Identification of the Virulence Gene of *Mycobacterium tuberculosis*

By

Halimah Adenike Rabiu
B.Sc. (Hons) Applied Biotechnology
University of the Western Cape

Thesis presented for the Degree of Masters of Science (Bioinformatics)
in the Department of Cell and Molecular Biology
University of Cape Town

June, 2007

Identification of the Virulence Gene of *Mycobacterium tuberculosis*

Halimah Adenike Rabiu

Supervisor: Dr Jane Nicola Mulder

Summary

The number of human deaths as a result of infections from microbial pathogens is enormous and is on the increase. Despite advance in drug design methods and effective drug and vaccine administration against major infectious diseases, infections due to microbes still constitute the major cause of death and disability worldwide. *Mycobacterium tuberculosis*, the causative agent of tuberculosis is the major cause of human death from a single infectious agent among adults in developing countries. It kills over 3 million people annually with the number on the increase due to the outbreak of multi-drug resistance strain in various places. In Sub-Sahara Africa as a consequence of the HIV-AIDS pandemic, the incidence of tuberculosis has worsened. In South Africa alone, roughly 1000 people die of the disease each day. This is even becoming more worrisome with the emerging threat of mutant strains of these pathogens that are resistant to available antibiotics and vaccines.

However, the emergence of complete genome sequences for various microbial pathogens means all potential targets are catalogued. One of the main goals of microbial comparative genomics using bioinformatics tools is to narrow down the search for genes that are essential for pathogen survival in their host and provide, opportunities for understanding and combating infectious diseases in human. Microbes have the uncanny ability to adapt to their environment, which results in ever changing targets for drug and vaccine research and making their control and eradication extremely difficult. We believe that genes that are unique to the pathogens should be a priority for future studies. The availability of the sequence data has enormous potential, but more relevant is the mechanisms involved in virulence, infection and disease progression.

The major thrust of this project is to identify and characterize potential virulence genes from *M. tuberculosis*. To this end, we have compiled and integrated information from various public databases to catalogue 246573 microbial genes from 84 organisms, including pathogens and non pathogenic microbes. We determined the phylogenetic distributions by grouping the proteins into families based on sequence similarity with the aid of BLASTP and the NCBI BLASTClust program. Three sets of experiments were generated with the two approaches and the results exported into binary and gene number matrices. Potential virulence genes were predicted to be those unique to pathogens or those unique to *M. tuberculosis* complex organisms. We identified 18022,

8459 and 7904 proteins clusters that are unique to pathogens from the three experiments BLASTClust1.0, BLASTP, and BLASTClust0.5, respectively, with the most constrained experiment producing the largest number of pathogen specific clusters. This experiment, BLASTClust1.0, produced 2362 pathogen-specific clusters that included a member from *M. tuberculosis*.

We then performed functional analyses of these potential virulence genes and predicted potential functions for some hypothetical proteins using InterPro and gene ontology annotations. Further evolutionary analysis was also performed to confirm that protein members from same cluster are evolutionary related to each other.

.....this work is dedicated to the efforts, ideals and visions of institutions providing financial supports for students to pursue and realize their dreams, in particular, the **Deutscher Akademischer Austrausch Dienst** [German Academic Exchange Service] and the SA **National Bioinformatics Network, NBN** for their generous scholarship and bursary.

Acknowledgements

How I wish I could find the all expressing words, sentences or phrase to express my gratitude to all that had contributed toward the successful completion of this work and more important to the further self-discovery.

My warmest gratitude to my supervisor and *teacher* **Dr Nicola Jane Mulder**. I will remain ever grateful for her unqualifiable sense of understanding, patience, academic insight and personal motivation. Her thorough grasp of the subject matter and painstaking research and writing efforts and concern, needles to stress let this work comes to fruition. You have reach out to a budding mind. The interest had been created and nurtured. Soon your rank will be swelled with one more enthusiastic and versatile bioinformaticist and academic. I consider myself lucky to learn from such an unassuming great mind and thinker. You have added a great value to my capability as a budding scientist and researcher. Thank you so much.

Professor Cathal Seoighe for his very helpful suggestions and for providing me the opportunity to experience, explore and learn bioinformatics and bioinformatic tools. Thank a lot for affording me with the privileged and cherished opportunity to come into contact with world-class resources, people and facilities.

I am also grateful to the wonderful people of Computational Biology Group, IIDMM for helping to enhance the clear focus; **Dr Gordon, Natasha, Venu, Bukky** you guys have made a notable contribution to this work; and **Graham** for helping with the program codes. **Vicky** who is always available to help at any time no matter the inconveniences. Humble, accessible and always enthusiastic, not for once did she failed me. Definitely great reward awaits her, who having reached higher on the ladder of knowledge and success eagerly pulls up others. It is great that you know that at least the top is wide enough to accommodate all who are willing to pay the price. **Nobubelo**, we just share too many things together, your word of encouragement and supports at most stressful time will be forever cherished. *Pity* you only choose to be my sister [uuh?].

To my mentor **Dr Zainu Arief**, the girl has grown and is growing. I always feel a great internal joy that I am living up to your expectations and hopes.

Sis. Kafaya sukran jazeelan for always being there even at the greatest inconviniences. You are more of a blood sister than a friend. There is just no way I could have forged ahead without your support as a mother for my child.

My heart goes out to the love of my life, the master motivator, my driver and the never-ending source of inspiration, **Abu Hamidah**. Your believe in me, engender the interest, and give me the courage and enthusiasm to forge ahead. You are my best friend and brother. To you I owe all my achievements. **Hamidah** and **Abdul-hamid** I do apprecaite your love and sacrifices, to lighten the burdens and responsibilities of motherhood.

To my father and mother thanks for your support and words of advise. I hope I will soonest be able to be there for you. Also to my in-laws for taking me as a daughter.

Finally, to God Almighty be the glory, for the good health, life and grace to continue to live rather than just existing.

Contents

List of Tables

List of Figures

1

Introduction

".....It may be expected that the elucidation of the aetiology of tuberculosis will provide new viewpoints for the study of other infectious diseases."

— Robert Koch, 1882

1.1 General Overview

Pathogens are microscopic organisms and include bacteria, viruses, fungi or parasites that infect other organisms using the host body to live and grow, and often affect normal cellular functions, leading to illness in the host. Human death annually as a result of infections from microbial pathogens is enormous and is on the increase. A significant advancement has been recorded in drug design methods and effective drug and vaccine administration against major infectious diseases. However, infections and diseases due to microbial pathogens are still the leading cause of death and disability worldwide.

The world is witnessing the emergence of new infectious diseases and re-emergence of old deadly ones [Smolinsk *et al.,* 2003; Cooksey, 1996; Pablos-Mendez, 1998]. Mutant strains of the virulent genes of the pathogens that are resistant to available antibiotics and vaccines are becoming more prevalent [Spratt, 1996; Böttger *et al.,* 1998]. The uncanny adaptability of these pathogenic mutant strains is greatly threatening public health and well-being. This necessitates a significant improvement in our ability to detect, control and eliminate these disease-causing microbes [Nair *et al.,* 1993; Riska *et al.,* 1999]. *Mycobacterium tuberculosis*, the causative agent for Tuberculosis (TB), for instance, kills over 3 million people annually [WHO, 1998; Murray and Nardell, 2002]. It is estimated by the World Health Organisation [2002] that one third of the world's population (around 2 billion people) is infected with TB. WHO [2004] reported an annual

incidence rate (number of new cases) of 356 per 100,000 in Africa. About 8.9 million new cases of active TB were diagnosed worldwide in 2006 [WHO, 2006] and these figures probably represent less than half the true number of new cases. In South Africa alone, roughly 1000 people die of the disease each day [WHO, 1998] probably due to HIV infection complications. Ironically, TB is a theoretically preventable as well as a curable disease, yet it is far from being a disease of the past. With the AIDS pandemic, the incidence of tuberculosis has increased further, leading to the WHO declaring tuberculosis a global emergency in 1993 [WHO, 1998].

Tuberculosis is caused by the invasion of the macrophage by an intracellular pathogen *Mycobacterium tuberculosis* [Dietrich *et al.*, 2006; Glickman and Jacobs, 2001; Raviglione *et al.*, 1995], an organism that has evolved a complicated and advanced survival and evasion mechanism [Dietrich *et al.*, 2006]. It is known that *M. tuberculosis* has survived inside even fully activated macrophages [Baumann *et al.*, 2006]. TB is one of the ancient infectious diseases that are still endemic in the human population today. The oldest evidence of TB was obtained in Egyptian mummies dated about 24000 BC [Heifets and Good, 1994]. Despite being a very old disease, little is still known about how to eradicate this and related diseases. TB still maintains the reputation of being the major cause of human death from a single infectious agent among adults in developing countries [Dietrich *et al.*, 2006; WHO, 1998; Dolin and Kochi, 1994].

Coupled with the HIV-AIDS pandemic in developing countries (sub-Saharan Africa in particular) the disease is fast becoming one of the most deadly of the major AIDS–related opportunistic infections [Edlin *et al.*, 1992]. While many of the common microbial pathogens can be prevented by vaccines or cured by antibiotics, slow diagnosis of the causative agents in patients and the emergence of multi drug-resistant strains of *M. tuberculosis* is a major impediment to eradication of TB [Dietrich *et al.,* 2006; Kim, 2005]. The understanding of the infection mechanisms of microbial pathogens requires the understanding not only of the biology of the pathogens concerned, but equally important is the whole range of the cellular and immune responses these microbes stimulate in the host organism [Lara-Tejero *et al.,* 2006].

Microbial pathogenesis investigates the mechanisms by which microbes cause diseases in their host by probing the molecular interactions between specific microbial products and the host cell [Glickman and Jacobs, 2001]. *Microbial virulence* includes strategies

that are used by microbes to evade the host defense mechanisms through targeting and modification of host cellular processes. This is achieved for instance by expression of genes that produce toxins, adhesions, capsules, and other molecules involved in the invasion of host cells and tissues, ensuring the microbe's survival but resulting in disease (and even death) in the host [Fruth and Young, 2004].

Microbes have the uncanny ability to adapt to their environment, which results in ever changing targets for drug and vaccine research, making their control and eradication extremely difficult [Sambandamurthy and Jacobs, 2005]. A detailed elaboration of the microbial pathogenicity pathways and strategies on a molecular level will enable the rapid development of effective drugs, antimicrobials, vaccines and diagnostic tools to combat the menace [Glickman and Jacobs, 2001; Weinstock *et al.,* 2000]. This is only recently feasible in the case of *M. tuberculosis* with the successful efforts to genetically manipulate the organism [Cole *et al.*, 1998; Bardarov, *et al.*, 1997; Pelicic *et al.,* 1997]. Towards this end, the emergence of complete genome sequences for microbial pathogens provides a database of the potential targets for customised drugs and vaccines [Weinstock *et al.,* 2000].

The availability of the sequence data has enormous potential, but for practical applications, a comprehensive elucidation of the mechanisms involved in virulence, infection and disease progression are essential too [Lonnroth *et al.,* 1999]. Recent advances in comparative and functional genomic studies have provided researchers with a valuable understanding of the biological functions of genes [Pelicic *et al.,* 1997] and hence made available vital information to rapidly zero in on potential drug and vaccine candidate genes [Lowrie, 2006]. Availability of genomic sequence has moved the focus of understanding the biology of an organism from the single gene to the whole genome, which provides the opportunity to look at genes within their context in a cell and achieve a global view [Scarselli *et al.,* 2005].

1.2 Potential of Bioinformatics Tools

Bioinformatics can literarily be summed up as the synergy of Information Science, Statistics (and to some degree Mathematics) and Applied Biology [Luscombe *et al.,* 2001]. It basically involves the storage, organization and indexing of sequence information on one hand, and the analysis of this information on the other [Becker, 2005]. The science of bioinformatics attempts to provide more insightful knowledge about the

fundamental biological functions of organisms. This information can be employed, among other things, in the development of drugs, antibiotics, biological systems and vaccines to combat the menace of microbes [Schultz *et al.*, 2002]. It can also find application in a plethora of other processes, *viz* environmental, agricultural and medical research.

The big challenge in bioinformatics is to provide the means and tools for efficient storage and management of the huge volume of genomic data being produced, and making it easily accessible to the public [Blanchard, 2004]. Analyzing a whole genome using bioinformatics tools provides insights into the potential functions of genes and hence greatly reduces the number of experiments needed to be carried out to confirm the function [Brosch *et al.*, 2000]. With the advent of a variety of bioinformatics tools, analysis of genomic sequence data has been made easier by reducing the number of genes of interest for follow-up studies with respect to a particular problem [van den Braak *et al.,* 2004].

The availability of genome sequences of diverse organisms provides a catalogue that can easily be accessed for any potential drug targets or vaccine candidates, since, for microbial genomes, the potential genes have been predicted [van den Braak *et al.,* 2004]. The difficulty lies in identifying which genes are the targets or candidates, a task that requires filtering the genomic data by some means and elucidation of the function of relevant gene products [Lowrie, 2006].

There is a tripod of fundamental theories upon which the study of bioinformatics revolves (http://www.ebi.ac.uk/2can/bioinformatics/bioin f_why_1.html):

- DNA sequence determines protein sequence
- Protein sequence determines protein structure
- Protein structure determines protein function.

A detailed elucidation of the biological pathways of each of these processes will enable a full understanding of the biology of organisms and the objectives set above.

1.3 Problem being addressed

The amount of genomic data from sequencing projects has grown exponentially over the last few years due to the introduction of new and faster sequencing technologies. Currently over 400 published complete genomes are available on the Genome On Line Database (GOLD) and more than a thousand are in the pipeline at about 10 major sequencing centres [Liolios *et al.*, 2006].

The effort to include genomes of different strains of the same species in the sequencing projects provides the opportunity for both intra-species genomic comparisons and comparisons between genomes of related species. Significant successes have been recorded in comparative and functional genomics studies performed on mycobacterial species. These efforts seek to understand the evolution of these pathogens [Brosch *et al.,* 2001; Cole *et al.,* 1998] and how they interact with their various hosts. This is achieved by techniques of proteomics, transcriptomics and microarray analysis, among others.

Of particular relevance to this work are the successful efforts of previous researchers to catalogue whole genome sequences of many members of the Mycobacterial species [Fleischmann *et al.,* 2002]. Researchers have compared the sequences with other bacterial species [Gordon *et al.,* 1999; TØnjum *et al., 1998*] and with the genome sequences of other organisms, including other microbes, pathogens, mice and men [Poux *et al.,* 2002]. This is to enhance specificity in identification of essential genes, by pinpointing potentially unique proteins. The comparative studies and other molecular techniques revealed the diversity and complexity within the genus. For instance, the analysis of isolates from environmental and clinical sources suggests that the prevailing theories of the origin of tuberculosis coinciding with the domestication of cattle are unlikely to be correct [Fleischmann *et al.,* 2002]. Significant deletions were observed when the *M. tuberculosis* genome was compared with that of *M. bovis.*

Many comparative genomics studies have been performed for identification of unique proteins between pathogens [Prentice, 2004; Gil *et al.*, 2004], non-pathogens, [Nascimento *et al.,* 2004] different species of the same genus [Nascimento *et al.,* 2004; Ferretti *et al.,* 2004; Moreira *et al.*, 2004] or between different strains of a species [Deng, 2003; Siew, 2004]. Identification of unique proteins at different taxonomic levels has provided knowledge of the metabolism, pathogenicity, physiology and behaviour of

different organisms. Some have been successful in identifying potential virulence factors, but few have proceeded to further characterise these factors. In addition, each study has, in general, used a single reference genome to determine the difference between organisms.

This project involves more detailed comparative genomics studies to produce a comprehensive microbial gene resource. The resource was used to identify genes that are unique to pathogens and genes unique to *M. tuberculosis*, and thus the genes potentially involved in virulence. However, it has also shed light on similarities and differences between other organisms, and not just the reference one. The work has also included a downstream analysis of candidate genes and proteins as opposed to the majority of comparative genomics studies that focussed more on the genome comparisons and only hinted at potential functions of interesting genes.

The current trends in bioinformatics studies often place major and disproportionate focus on statistics and computational techniques. Here the aim is to employ bioinformatics tools to elaborate on the biological functions and pathways to try to understand the process and evolution of virulence in microbial pathogens rather than the traditional large-scale computational biology approach. At the same time, a comprehensive catalogue of microbial genes and their phylogenetic distributions has been created.

1.4 General Objectives

This project focuses on using various Bioinformatics resources to compare genomes of pathogens and non-pathogenic bacteria using functional and comparative genomics. It is aimed at identifying and characterizing virulence genes from *Mycobacterium tuberculosis.*

Characterization of these specific genes will help to gain insight into their mode of pathogenicity and virulence, and subsequently aid in the identification of candidate vaccines and drug targets. A better understanding of the organism's mechanisms of infection and survival in the host will enable subsequent studies to investigate creating more selective and faster diagnostic reagents for tuberculosis and eventually eliminate the emergence of multiple drug resistant strains.

1.4.1 Specific Objectives

This study aimed to produce a catalogue of all microbial genes and determine their phylogenetic profiles. This was employed further for identifying homologous genes between 84 completely sequenced bacterial genomes and cataloguing the genes into pathogen and non-pathogen sets with the aid of a binary matrix. The genes that are common to pathogens were then characterized, paying specific attention to genes that are unique to *M. tuberculosis* strain CDC1551.

The approach involves comparative genomics and proteomics, data mining and storage of the results in a gene matrix. It has produced a comprehensive microbial gene catalogue that was used to answer the following questions:

1) What is the phylogenetic profile of all microbial genes?
2) What genes are common to all pathogens and absent from non-pathogens, and which genes are unique to *M. tuberculosis*?
3) What genes are potentially involved in virulence in *M. tuberculosis*?
4) What is the potential role of these genes in virulence?
5) How did these genes evolved?

1.5 General Approach

The project was carried out using whole proteome comparisons to produce sets of potential homologous groups using BLASTP [Altschul *et al.,* 1990], NCBI BLASTClust [http://biowulf.nih.gov/apps/blast/doc/blast clust.html] and the InterPro database [Mulder *et al.,* 2005]. Basic alignment search tool (BLAST) is a statistical algorithm used to find regions of similarity between two or more sequences. The aligned regions, referred to as segment pairs, usually consist of gapless alignments of any part of two sequences: the query sequence and the database. The sum of the scoring matrix values within the alignment regions is higher than the level that could be expected to occur by chance alone.

The NCBI BLASTClust program program clusters protein sequences into sets of related proteins based on sequence similarity, E-value, percentage sequence identity or score density and percentage overlap between sequences. InterPro is a database that integrates information and resources for protein families, domains and functional sites. It

assembles information from various protein signature databases that are generated through different protein signature methods. InterPro provides information on the families a protein belongs to and the functional domains it contains [Mulder *et al.,* 2005; Apweiler *et al.,* 2001].

Sequence similarity searches were carried out using BLASTP [Altschul *et al.,* 1990] followed by clustering proteins into related sets based on E-value and percentage overlap between two sequences with the aid of PYTHON scripts, and the NCBI BLASTClust program [http://biowulf.nih.gov/apps/blast/doc/blastclust.html]. Three sets of experiments were generated with the two approaches and exported into phylogenetic profiles by putting representative gene names/homologue set names down the first column and organism names across the first row of the matrices. The rest of the matrices were populated with the corresponding accession numbers.

Futher functional analyses were then carried out using various annotation data like UniProt description, GO slim terms and InterPro names for each experimental set. Evolutionary analyses of selected homologue sets (those that include candidate virulence genes from *Mycobacterium tuberculosis*) were also carried out. Figure 1 present a schematic view of the general approach used.

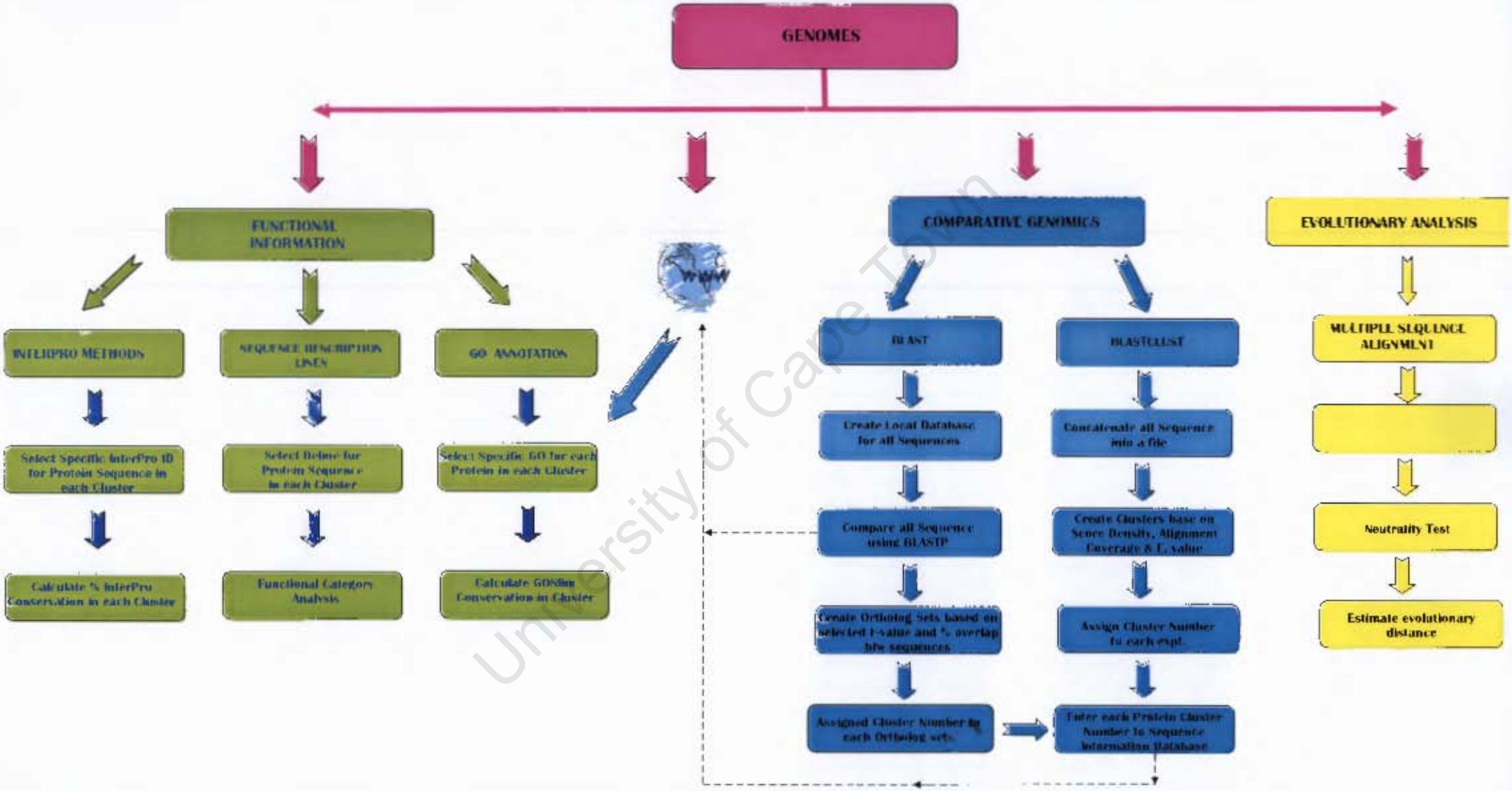

Figure 1: Schematic representation for the comparative genomics, evolutionary and functional analysis procedure

2

Survey of the Literature

"If the opportunity to end tuberculosis is not seized now, it may be lost indefinitely"

2.1 History of Tuberculosis

Human death as a result of TB infection has been long recorded [Bates and Stead, 1993]. Hard bones that were deformed by TB have been identified thousands of years later, the oldest of which were found in various fragments of Egyptian mummies from 2400 BC [Daniel, 2006; Gradmann, 2006]. Similar discoveries were found in Europe and the Middle East, suggesting that TB is an ancient, world wide disease [Haas and Haas, 1996; Smith, 2003]. There are various records of the disease in ancient works of Greek physicians (prominent among whom is Hippocrates) and Hebrews (*Schachepheth* in Modern Hebrew translates to tuberculosis). The disease, due to its widespread nature, was variously referred to as *white plague, white swelling*, *phthisis*, *Lupus vulgaris* (TB of the skin), *Scrofula* (lymph glands TB), *Mesenteric disease* (TB of the abdominal lymph glands), Pott's disease (TB of the spine), *consumption* and so on [Murray, 2004].

The infectious nature of TB was first described in the early 18th century by the English physician, Benjamin Marten [Sarrel, 2006]. He postulated that TB is caused by microbes- *wonderfully minute living creatures*- and is transmitted to a healthy person via frequent long time contact and interactions with a consumptive patient. He rightly opined then that short time contact with consumptive patients is seldom, or never, sufficient to transmit the disease. This is a landmark conceptual departure from the earlier belief that the disease just appears spontaneously in its victim and is incurable.

Marten's work was supported by the discovery of Jean-Antoine Villemin, a French doctor in 1865, that *consumption* could be passed from humans to cattle and from cattle to rabbits [Villemin, 1868 cited by Murray, 2004]. This further reinforces the postulation that

the disease is caused by a specific micro organism and did not arise spontaneously in its sufferers, and marks the beginning of the actual fight against the organism that causes TB. This first attempt was the creation of sanatoria to isolate the infected people and hence prevent the spread of the disease and at the same time provide a healthy rest house to aid the healing process [Trudeau, 1887 cited by Murray, 2004].

A new dawn towards a pathological understanding of the disease was witnessed with the isolation and description of the causative microbes for tuberculosis in March 1882 by a German bacteriologist named Robert Koch as *tuberculosis bacteria* or *tubercle bacilli* [Gradmann, 2001; Gradmann, 2006]. He developed a special staining technique that allowed him to view the organism [Krause, 1932 as cited by Murray, 2004]. The work earned him the Nobel Prize for physiology or medicine in 1905 [Murray, 2004]. He announced *tuberculin,* a glycerine extract of the *tubercle bacilli,* as a remedy for tuberculosis in 1890. This medication, though it proved ineffective, formed the basis of the efforts of Von Pirquet the search for a cure for pre-symptomatic tuberculosis [Sakula, 1982].

The discovery by Forlanini, an Italian physician that lung collapse had a positive effect on the outcome of the disease marked the beginning of active therapy for TB [Doetsch, 1978]. He developed and introduced surgical methods and artificial *pneumothorax* to reduce lung volume to combat the disease [Skeiky and Sadoff, 2006]. Wilhelm Konrad von Rontgen introduced radiation that could be used to view the progress and severity of the disease in a patient in 1895.

Although a number of other treatments were tried to combat the menace of this seemingly unstoppable disease, the first success in immunizing against tuberculosis was named after the two French biologists that discovered it, Albert Calmette and Camille Guerin, in 1906 [Gradmann, 2006]. The BCG vaccine, developed from an attenuated bovine-strain tuberculosis, was first used on humans on the 18th July 1921 in France [Daniel, 2000] and later adopted throughout the world. It is still in use today as prevention for TB.

Selman Waksman and his team at the University of California successfully isolated an effective antibiotic, *actinomycin.* This, however, was found to be too toxic for use in humans or animals [Murray, 2004]. The team eventually succeeded in 1943 to produce a compound called Streptomycin purified from *Streptomyces griseus.* This antibiotic turned

out to be effective against *Mycobacterium tuberculosis*, and non-toxic to humans. It was first administered to a human in November 1944, halting sensationally the progression of the disease [Daniel, 2000].

With time, more effective TB drugs and vaccines were developed - para-aminosalisylic acid, PAS (in 1946); Isoniazid, INH (1951); Pyrazinamide (in 1954); Cycloserine (1955); Ethambutol (1962); Rifampicin (1963) and so on, to combat TB [Global Tuberculosis Institute, http://www.umdnj.edu/ntbcweb/tbhistory.htm]. This was very important in light of the fact that antibiotic resistant mutants quickly began to appear. For instance, as early as 1947, resistance to streptomycin has been recorded [Ryan, 1992].

2.2 Modes of Infection, Transmission and Clinical Manifestations

Tuberculosis is a disease caused by a slow growing aerobic bacterium, *Mycobacterium tuberculosis* (MTB). The microbe is one of the most successful bacterial pathogens in the history of mankind in terms of its persistent uncanny adaptability to survive and thrive. A wide variety of anti-TB drugs have been developed yet the pathogen is still the leading cause of human death from infectious diseases [Burgos and Pym, 2002]. *M. tuberculosis* is a pathogen that can live for years in its host, in latent states, without manifesting any symptoms. It was found that the risk of latent TB progressing to active tuberculosis is highest during the first 2 years of infection [Ferebee, 1978 cited by Elad *et al.,* 2001]. While only 10% of TB infection progresses to TB disease, if untreated the diseases eventually causes chronic debilitation and death [Frieden and Munsiff, 2005].

The disease commonly manifests as pulmonary TB that affects the lungs, but infection can spread via blood from the lungs to all organs in the body, like bones and joints, central nervous system, and genital and urinary organs [Doetsch, 1978]. However, only the pulmonary form of the disease (or in the lung) is infectious, TB in other parts of the body, such as the kidney or spine, is usually not infectious.

TB is spread almost exclusively by airborne transmission [Blower *et al.,* 1995]. It is transmitted from person to person through the inhalation of the bacteria released into the air when a person with pulmonary or laryngeal TB coughs, sneezes or speaks. Infection occurs via inhalation of aerosols or air containing droplet nuclei of the bacilli [Frieden and Munsiff, 2005]. Depending on the environment, these tiny particles can remain in the air

for several hours. There is as yet no agreement about the safe exposure time to airborne *Mycobacterium tuberculosis* [Sudre *et al.,* 1992]. The bacteria usually settle in the lungs, begin to grow and then invade other parts of the body, such as the kidney, spine and brain. As mentioned previously, while TB of the lungs (pulmonary TB) or throat (Laryngeal TB) can be spread to other people, TB in other parts of the body, such as the kidney or spine, is usually not infectious [Blower *et al.,* 1996]. Ussery *et al.*, [1995] and Hutton *et al.*, [1990] have, however, reported cases of extra-pulmonary TB transmission.

The probability that TB will be transmitted depends on various factors like the number of organisms expelled into the air, duration of exposure, and the virulence of the bacterial strain [Sterling *et al.,* 2006]. The tuberculosis bacteria are killed when exposed to ultraviolet light, including sunlight. The organism's infection pathogenesis occurs in two stages: the first stage is regarded as latent TB (or TB infection) where the infected organism remains in a dormant (albeit alive) state and this state may persist for many years and sometimes for the entire life time of the host. During this state, the host show no sign of the disease and cannot transmit it.

The progression from latent TB to the active TB (TB disease) stage is triggered by a weakening in the immunological defence system of the host, allowing the bacteria to multiply. This may be as a result of many factors, including incidence of infections that compromise the immune system like HIV [Murray and Salomon, 1998; Shafer and Edlin, 1996; Selwyn *et al.*, 1989] and Diabetes [Rieder, 1989, CDCP, 1992]. Other factors that may cause the progression of TB are the infective burden of *Mycobacterium tuberculosis,* previous exposure to TB infection, virulence of the *Mycobacterium tuberculosis* strain and a host of other factors.

Selwyn *et al.* [1989] and Selwyn *et al.* [1992] reported that the risk of developing TB is 7% to 10% each year for persons who are infected with both *M. tuberculosis* and HIV. The risk is about 10% over a lifetime for persons infected only with *M. tuberculosis.* Also, while the global TB fatality rate is 23%, the value exceeded 50% in some African countries with high HIV rates [Dye *et al.,* 1999]. The major occurrence of active TB is in the form of Pulmonary TB -about 73% [CDCP, 1999]. This form is characterized by chronic or persistent cough and sputum production, fatigue, lack of appetite, weight loss, fever, and night sweats [CDCP, 1999]. If left untreated, this may be followed by coughing of blood. Extra pulmonary TB usually occurs in many forms like chronic, non tender

lymphadenopathy in Lymphadenitis TB, acute illness with cough, pleuritic chest pain, fever, or dyspnea in Pleural TB, and pain, joint swelling, and hampered mobility, draining sinuses and abscesses in chronic cases of Skeletal TB.

2.3 Epidemiology

The resurgence in the incidence of TB across the world has spurred research efforts to understand the epidemiology and pathogenesis of this disease [Burgos and Pym, 2002]. Human death as a result of infections and complications from TB is second only to that of HIV/AIDS, with about 1.7 million human deaths recorded in 2004 [Cole *et al.*, 1998; WHO, 1998]. About one-third of the world's population have latent TB infection; about 14.6 million people have active TB disease and a there is an annual incidence rate of about 9 million [Kurup and Chan, 2006]. Expectedly, the incidence rate reveals a direct relationship between the level of economic development and level of TB infection of the area [Blower *et al.*, 1995]. For instance, the figures range from 356 per 100,000 in Africa; 41 per 100,000 in the Americas; to about 90 per 100,000 in the UK [Bradford *et al.*, 1996].

Figure 2.1 shows the trend of notification of TB in WHO regions. While the figures for most of the regions are relatively stable over time, a drastic increase was recorded for Africa, leading to the World Health Organization declaring the region a TB epidemic area. This trend largely revealed that the current TB control measures are so far ineffective, which may be a direct consequence of the worsening HIV/AIDS epidemic in Africa [Nahid and Daley, 2006; Nunn *et al.,* 2005; Barnes *et al.*, 1991]. Figure 2.2 shows the percentage contribution of cases for WHO regions to TB notification in 2004. Africa alone contributed 24% of the total in 2004, and the majority of the top 12 counties are in the SADC region of Africa. Figure 2.3 shows data for the top 12 countries with the highest TB case notification rates per 100,000 in Africa for the year 2004.

The study of TB epidemiology is aimed towards enabling a more detailed understanding of the inter-relationships between *M. tuberculosis* and the host in each community in question and under natural conditions, that is, in the absence of any interventions [Styblo, 1985; Styblo 1984 cited by Murray and Nardell, 2002]. While a detailed elucidation of the transmission pathway of the diseases within populations is essential for evolving successful treatment and control measures, there are limited tools and resources for diagnosis and record-keeping [Murray and Nardell, 2002]. This would have

provided data on TB incidence cases, infectiousness and mortality rate, and hence provide a scientific basis for treatment and control measures to be formulated and implemented.

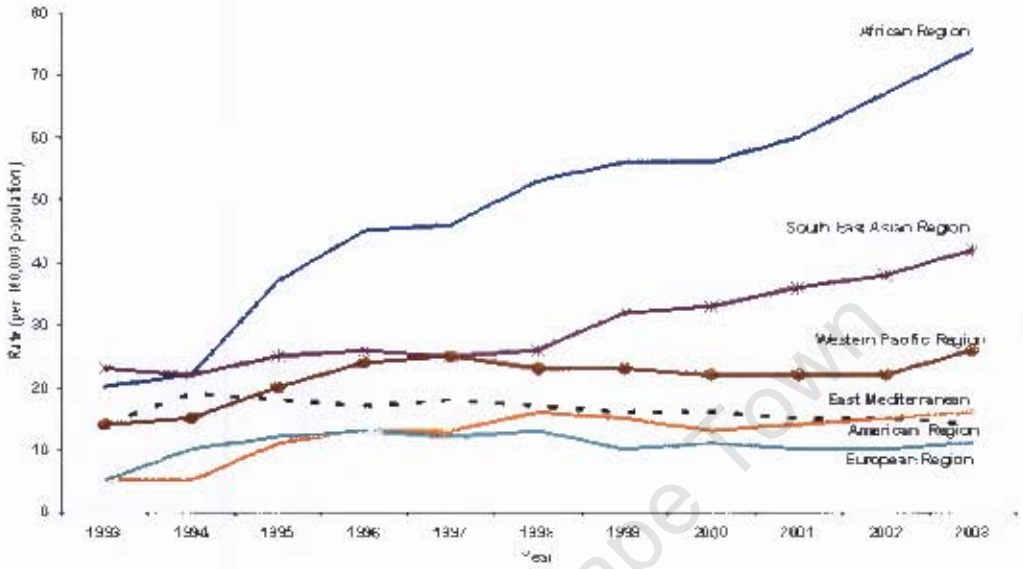

Figure 2.1: World TB notification rate per 100,000 population size
[Source: http://www.afro.who.int/tb/notificationtrend.html]

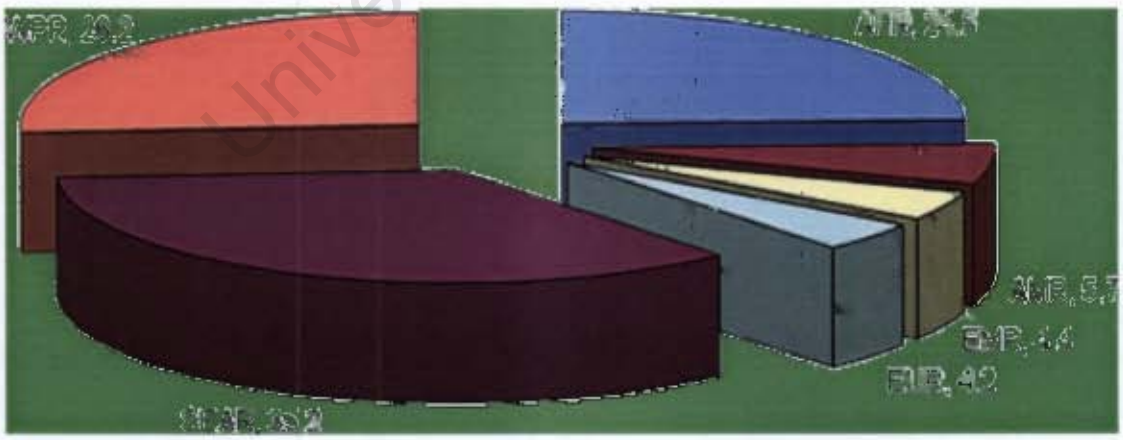

Figure 2.2: WHO Regions percentage contribution to TB notification in 2004. AFR-Africa; EUR-Europe; AMR –America; EMR- Eastern Mediterranean; SEAR – South East Asia and WPR – Western Pacific regions respectively.
[Source: http://www.afro.who.int/tb/notificationtrend.html]

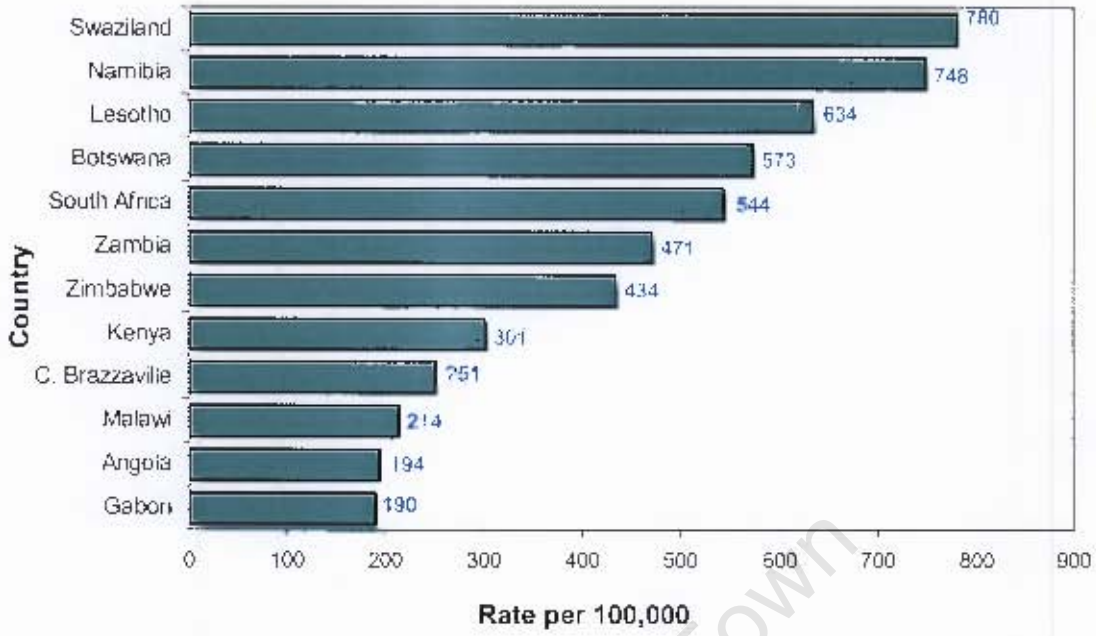

Figure 2.3: Top 12 countries with highest TB notification rates, African Region 2004. [Source: http://www.afro.who.int/tb/burdenreporting.html]

2.3.1 Drug and Multi- Drug Resistant Tuberculosis

The world wide increasing incidence of TB can be attributed to the emergence of drug resistant strains of *M. tuberculosis* (MTB). Drug resistant tuberculosis (DRT) is defined as TB cases with positive culture for an *M. tuberculosis* strain that was resistant to any of the drugs *isoniazid, rifampin, ethambutol* or *streptomycin* [Granich *et al.*, 2000; Park *et al.*, 1996]. The first case of *Mycobacterium tuberculosis* resistance to drugs occurred not long after the discovery of streptomycin in 1944 [Crofto and Mitchison, 1948 cited by El Sahly *et al.*, 2006].

There are two types of DRT: primary and secondary (or acquired) resistance. While primary DRT develops in persons who are initially infected with resistant strains of *M. tuberculosis*, secondary DRT, or acquired resistance, occurs during TB therapy. Acquired resistance results if, during the treatment period, the patient was treated with an inadequate regimen or if the patient failed to complete the prescribed regimen [Dooley *et al.*, 1992a; Edlin *et al.*, 1992; Fischl *et al.*, 1992].

Multi-drug resistant tuberculosis (MDRT) is defined as TB caused by strains that manifest a high degree of resistance to both *isoniazid* and *rifampin,* whether there is resistance to other drugs or not [Ormerod, 2005; Pearson *et al.*, 1992; Dooley *et al.*,1992b]. The emergence and worsening menace of drug (and multi-drug) resistant TB poses a serious threat to the effectiveness of the treatment protocol and the control measures for TB [Grannich *et al.,* 2000; Beck-Sague *et al.*, 1992].

In 1991, the Centre for Disease Control [CDC, 1992] conducted a survey across the United States. It found that 14.4% of new TB cases tested had organisms resistant to at least one anti-tuberculosis drug and 3.3% had MTB strains that are resistant to both isoniazid and rifampin. This is a far cry from a figure of about 0.5% recorded for the period 1982 through 1986. A similar alarming trend is noted for recurrent TB cases as well. Worthy of note is that the problem of DRT and MDRT is most prevalent in developing countries, accounting for about 90% of the reported cases [Ormerod, 2005].

2.3.2 Tuberculosis and the HIV epidemic

A critically important factor in the epidemiology of TB worldwide is HIV/AIDS [Comstock, 1999; Coronado *et al.*, 1993]. TB has been reported to be accountable for about 35% of deaths of AIDS victims. In Sub-Saharan Africa about 75% of individuals with tuberculosis are co-infected with HIV [Dlodlo *et al.,* 2005]. Figure 2.4 shows the estimated TB incidence in HIV positive adults per year for South Africa.

The immune mechanisms of an HIV-1 infected host that lead to the progression of latent to clinically active mycobacteria is still being actively debated [Shen *et al.,* 2004]. It is believed however that because of its adverse effect on the immune system, HIV infection facilitates acquisition of tuberculosis infection [Dlodlo, *et al.,* 2005; Shen *et al.,* 2004]. Co-infection with HIV is the most powerful risk factor associated with progression of latent TB infection to active tuberculosis. In effect, HIV serves to catalyze the acquisition and progression of TB and has been shown to be an important single-factor contributor to the spread of multi drug resistant TB strains (MDRT) [Small *et al.,* 1993].

It was found that latent TB infection in HIV positive patients is much more likely to progress to active tuberculosis compared to HIV negative ones. This phenomenon seems to be responsible for the high TB incidence and fatality rate in sub-Saharan Africa [WHO, 2003]. About 9 percent of new TB cases in 2000 were attributable to HIV. However, this figure varies greatly between regions and increases with an increase in

the HIV epidemic. In Sub-Saharan Africa, for example, some 31 percent of new TB cases are due to TB-HIV co-infection [Narain,1992].

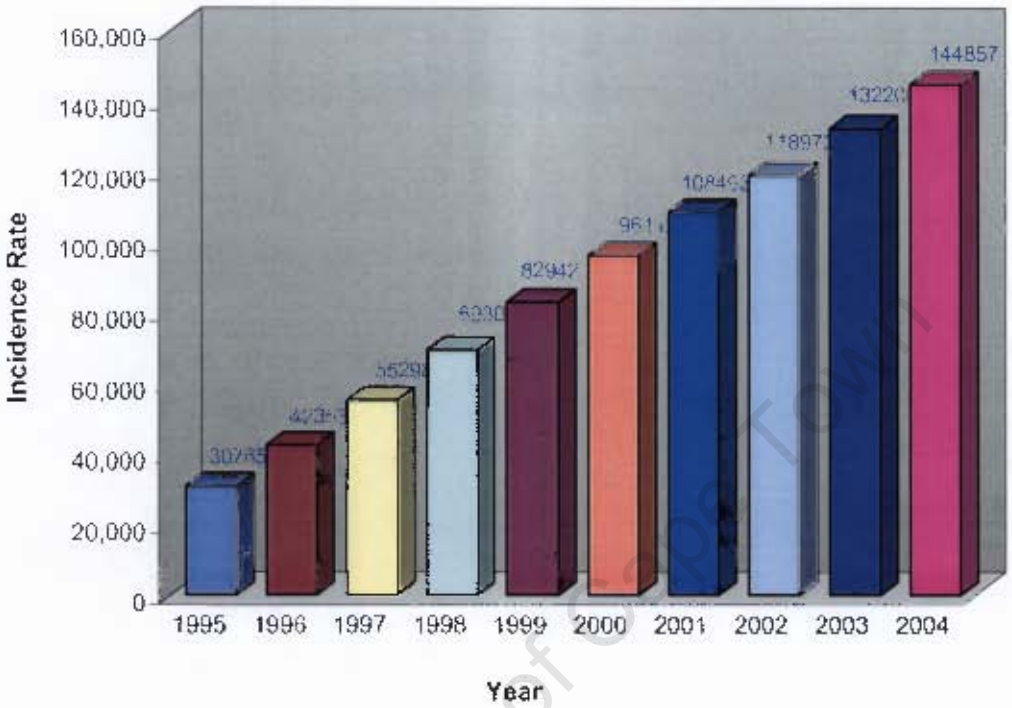

Figure 2.4: South Africa TB notification in HIV+ adults [WHO, 2006]

2.4 Bacterial Genomics

Bacteria are of immense importance because of their adaptability and capacity for rapid growth and reproduction [Wilson *et al*,. 2002]. While some are responsible for causing disease in other forms of life, other bacteria serve as sources of antibiotics, for instance streptomycin and nocardicin [Holt, 1994]. Some bacteria live symbiotically in the guts of human and manufacture vitamin K, an essential blood clotting factor [Davidson, 2006]. Others live in and/or on animals and on the roots of certain plants, converting nitrogen into a usable form [Adams, 2003; Holt, 1994]. Bacteria are also used for the production of desirable flavours in food production and to help break down dead organic matter [Griffiths *et al.*, 2006; Adams, 2003].

Genomics studies provide us with insightful details into how an organism functions, their genetic constitution, origin and evolution and their species diversity [Ward and Fraser, 2005]. Bacterial genomics is the study of bacteria using information derived from understanding the bacterial genome and its DNA sequence [Overbeek *et al.,* 2005]. A genome contains the complete hereditary information (both coding and non coding) of an organism that is encoded in deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) [Mora *et al.,* 2006]. Genomics studies provide a means of gathering information regarding the functional interactions between genes in an organism and genome relatedness, and as such, establish the commonality and differences in all forms of life [Ward and Fraser, 2005]. The ultimate goal of microbial genomics studies is to be able to make distinctions between genes that are essential for growth or those responsible for virulence. They can also facilitate the identification of the genetic differences between several genomes from the same species [Mora *et al.,* 2006; Overbeek *et al.,* 2005].

Over 300 bacterial and archaeal genomes that cut across wide species and strains of the same species have been successfully sequenced [Binnewies *et al.,* 2006]. The genome databases are expected to include about a thousand species soon [Field *et al.,* 2006; Overbeek *et al.,* 2005]. Genomic information has provided significant insights into the physiology and pathogenicity of many organisms. For instance, direct sequence analysis allows for genome-level analysis of pathogens, in particular those that are not amenable to genetic manipulation [Raskin *et al.,* 2006]. Such studies will also allow for examination of small differences such as single nucleotide polymorphisms (SNPs).

The development of gene cloning and sequencing techniques revolutionized the discipline of molecular biology [Efstratiadis *et al.,* 1977], providing researchers with a large and growing database of complete genomes of the smallest of prokaryotes to more complex ones [Edwards *et al.,* 2006]. Of significant note are the new sequencing technologies that allow for sequencing of random community DNA and single cells of bacteria, without the recourse to cloning or laboratory cultivation [Shendure *et al.,* 2005; Clarke, 2005 cited by Edwards *et al.,* 2006]. At present about 405 published complete genomes are available on the Genome On Line Database (GOLD) with a far larger number on stream as a direct consequence of improved new ultra-high-throughput sequencing technologies [Liolios *et al.,* 2006; Field *et al.,* 2006; Edwards *et al.,* 2006].

Cutting out the need for laborious laboratory experimentation, genomic analyses enable a rapid elimination of poorly conserved targets among various genomes and hence identification of well conserved genes and cis regulatory elements [Sturino and Klaenhammer, 2006]. The ongoing approach to focus research efforts on community whole genome sequencing rather than the present application of metagenomics to assay natural microbial communities will make bacterial comparative genomics more relevant to the understanding of microbial biology as a whole [Field *et al.*, 2006].

2.5 Bacteria Comparative Genomics and Phylogeny

Comparative genomics dwells upon understanding the function and evolution of genomes. For instance, a comparative genomics investigation into the interactions between phage and bacteria has been employed to engineer phage protection for industrially important bacteria used in bioprocessing activities [Sturino and Klaenhammer, 2006]. Many bacterial comparative genomic studies have being carried out to elucidate commonality and differences among various species of bacteria and to understand evolutionary relationships among these species. This is made possible by the employment of rapid methods for comparative sequence analysis of small subunit rRNAs [Woese, 1987].

Before the genomic era and the subsequent availability of comprehensive sequence databases, the phylogeny of bacteria was derived from the 16S rRNA. With the advent of comparative genomics and the availability of large sequence data for many organisms, this is now accomplished by comparing these sequence data to other potential phylogenetic marker molecules. Genomes are an excellent source of phylogenetic markers. These informative markers are based on genes and gene products that have sufficient sequence conservation and are universally distributed among various organisms and include, but are not limited to, large subunit rRNA, RNA polymerase, DNA gyrase, recA and tRNA synthetases [Edwards *et al.,* 2006]. For example, Figure 2.5 shows a maximum-likelihood tree produced from concatenated alignments of the universal subset of ribosomal proteins.

The comparative phylogenetic analysis of these bigger molecular markers is carried out as with the small subunit rRNA, save for some minor differences in local tree topologies, depending on the molecule analysed [Clarke, 2005]. The exception occurs only when

comparing a few distinct differences between phylogenetic trees derived from rRNA and protein genes [Field *et al.*, 2006]. Usually there are a large number of genes to work with. This may introduce some problems in the comparative phylogenetic analysis, particularly when there are genes that are paralogs from gene duplication events or genes that originate from horizontal gene transfer. The paralog markers can only be recognized as such if there are genomes containing all variants of the multiple genes.

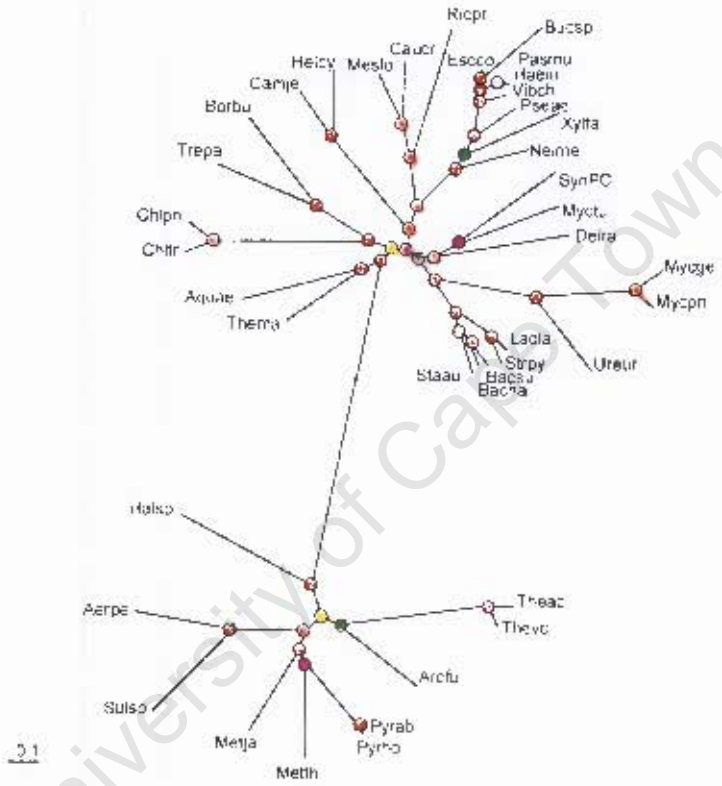

Figure 2.5: Maximum-likelihood tree produced from concatenated alignments of the universal subset of ribosomal proteins. The tree is unrooted. The circles indicate the level of bootstrap support, with the following colour coding: red: 90–100%, yellow: 80–90%, green: 70–80%, blue: 60–70%, magenta: 40–60%. The nodes with <40% support are unmarked [Wolf *et al.*, 2001].

However, with the availability of more genome sequences, it is possible to solve such problems practically with the use of genome-wide phylogenetic comparative analysis using large subsets or complete gene sets of all completely sequence genomes [Schleifer, 2004]. A conserved gene pair in prokaryotic genomes or distributions of percentage identity between probable ortholog sets to generate prokaryotic species trees

has revealed deep evolutionary relationships between prokaryotic lineages. Despite the problems with multiple genes and lateral gene transfer in organisms [Yuri *et al.*, 2001], however, distributions of identity with the phylogenetic analyses based upon alternative markers showed that small subunit rRNA derived trees globally reflect the phylogeny of the corresponding organism, and locally more their own history [Schleifer and Ludwig, 1999].

Figure 2.6 shows the corrected neighbour-joining evolutionary distance tree of the bacterial domain from approximately 8,000 bacterial 16S rRNA genes with 36 recognized divisions and putative candidate genes. The divisions which have cultivated representatives are shown in black; divisions represented only by environmental sequences are shown in outline [Pace *et al.*, 1998].

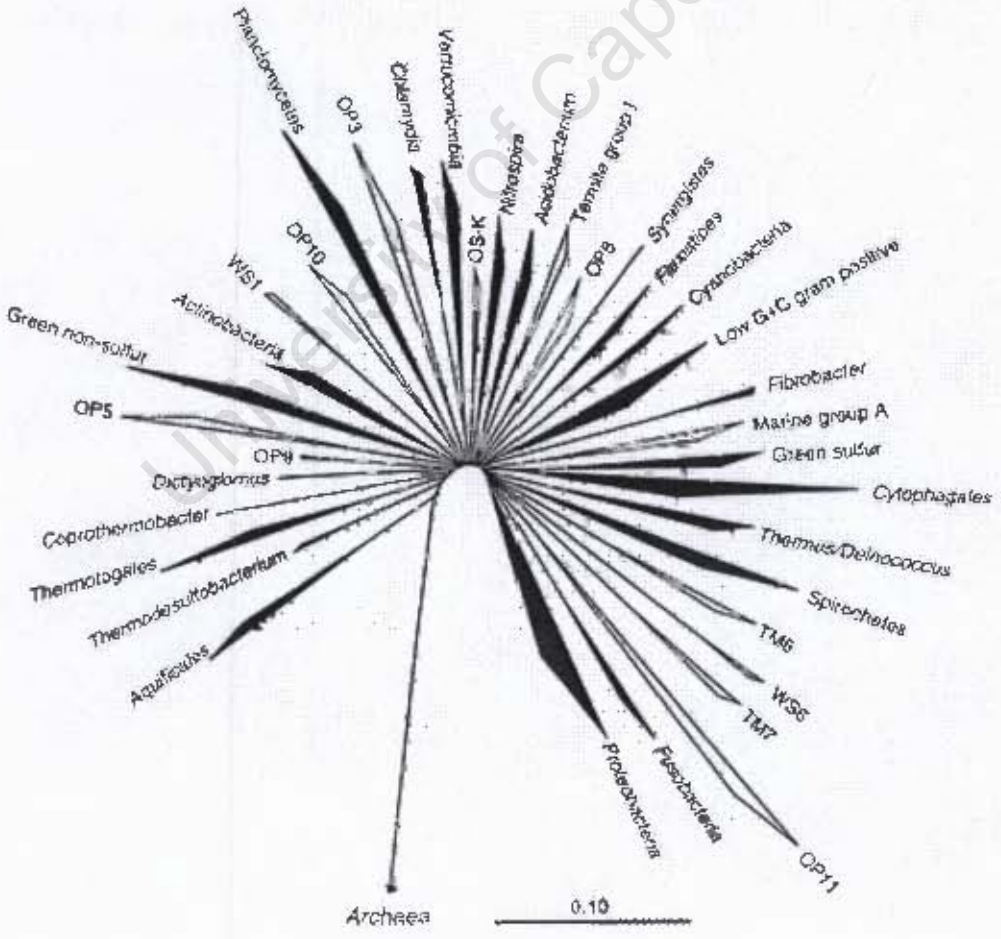

Figure 2.6: Corrected neighbour-joining evolutionary distance tree [Pace *et al.*, 1998]

Comparative genomics studies are not only used for phylogeny studies, but have also been performed for identification of unique proteins at different levels; between pathogens [Prentice, 2004; Gil and Moya, 2004], non-pathogens, [Nascimento *et al.,* 2004] different species of the same genus [Nascimento *et al.,* 2004; Ferretti *et al.,* 2004; Moreira *et al.,* 2004] or between different strains of a species [Mazumder *et al.,* 2005; Deng *et al.,* 2003]. Identification of unique proteins at different taxonomic levels has provided knowledge about the metabolism, pathogenicity, physiology and behaviour of different organisms [Siew *et al.,* 2004].

2.6 Mycobacterial Comparative Genomics

The whole complement of genes present within various mycobacterial species has been defined [Cole *et al.,* 1998; Camus *et al.,* 2002, Fleischmann *et al.,* 2002; Cole *et al.*, 2001; Gordon *et al.*, 2001] and their sequences compared to other mycobacterial species [Fleischmann *et al.,* 2002; Gordon *et al.,* 1999; TØnjum *et al.,* 1998]. Figure 2.7 shows a comparative phylogenetic tree of selected mycobactera, based on 16S rRNA sequences. The sequences of various mycobacterial species have also been compared to those sequences from other organisms, including other microbes and pathogens [Brosch *et al.*, 2001; Cole, 1999]. These studies have the potential to enhance specificity in identification of essential genes. Comparative genomics can also assist in pinpointing potentially antigenic proteins and narrow down potential targets for new and existing drugs and vaccines, as well as providing better diagnostic tools to detect mycobacterial infections [Cole, 2002].

The complete genome sequence of *Mycobacterium tuberculosis* (strain H37Rv) consists of 4,411,529 base pairs encoding approximately 3,986 proteins. Of these, 2058 proteins have a predicted biological function and 376 putative proteins share no homology with known proteins [Cole *et al.,* 1998; Camus *et al.,* 2002] The clinical strain, CDC1551, consists of 4,403,836 base pairs and approximately 4,187 open reading frames (ORFs), with an average G + C content of 65.6%. Predicted biological roles were assigned to 43% of the ORFs, while 15% have high sequence similarity with hypothetical proteins from other species and 42% are regarded as novel genes due to no match to any ORFs in the database at present [Fleischmann *et al.,* 2002].

The information revealed by the genome sequence of *Mycobacterium tuberculosis* has provided new and important insights into the biology of the tubercle bacillus and highlighted the significance of lipid metabolism to its way of life (about 8% of the genome is devoted to this activity) [Cole *et al.*, 1998]. While the cell envelope of *M. tuberculosis* was known to contain a notable numbers of lipids [Daffe and Draper, 1998], the genome sequence exposed many of the genes essential for their production. It was a surprise to find several genes and proteins that could code for enzymes that catalyze alternative sources of carbohydrates, that is a high number of enzymes involved in alternative lipid degradation from the host cell, which has not been reported in other bacteria.

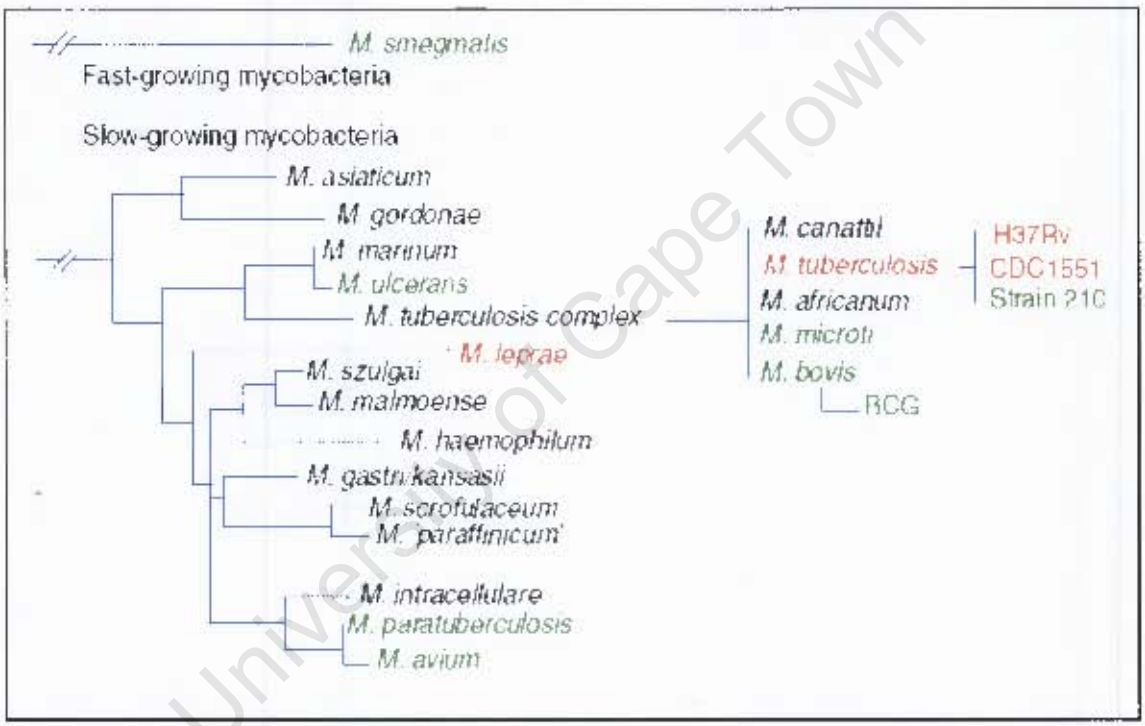

Figure 2.7: Phylogenetic tree of selected mycobacteria based on 16S rRNA sequences. Those shown in red and green are species whose genomes have been completely sequenced [Brosch *et al.*, 2001].

Despite *Mycobacterium tuberculosis* using lipolysis as its primary catabolic pathway, it also has a set of biosynthesis genes that are common to soil organisms involved in degradation of xenobiotics and modification of organic molecules to carbohydrates. This suggests the presence of a complete anabolic pathway, which supports the hypothesis that the tubercle bacillus has recently emerged as a human pathogen or that it has limited access to metabolic precursors within the phagosome [Cole, 2002]. Support for

the latter justification is provided by evidence that genes for anabolic functions have been greatly preserved in the genome of *Mycobacterium leprae*, a related obligate intracellular pathogen, despite massive reductive in the *M. leprae* genome compared to *M. tuberculosis* [Cole *et al.*, 2001; Eiglmeier *et al.*, 2001].

The genome sequence also enabled identification of some novel gene families which were either unknown before or poorly understood. Among these, the major ones are the Pro-Glu (PE) and Pro-Pro-Glu (PPE) families, including about 100 and 67 members, respectively [Cole *et al.*, 1998], which correspond to about 8% of the genome. Members of each of these families share a conserved N-terminal domain of about 110 and 180 amino acid residues, with the characteristic motifs 'Pro-Glu' (PE) or 'Pro-Pro-Glu' (PPE) at positions 8 to 9, or 8 to 10, respectively. The families are further subdivided into smaller groups, based on their C-terminal domains. Among the PE family are the polymorphic GC-rich sequence (PGRS) and major polymorphic tandem repeat (MPTR) subfamilies. The PE and PPE families are discussed in more detail below.

Various comparative and functional genomics approaches such as proteomics, bioinformatics, structural biology, transcriptomics, and microarray analysis have been performed on mycobacterial species to understand the evolution of these pathogens [Brosch *et al*, 2001; Cole *et al.*, 1998; Supply *et al.*, 2000] and their interaction with their various hosts [Camacho *et al.*, 1999]. *In silico* comparative sequence analysis of *M. leprae* and *M. tuberculosis* provides evidence for reductive evolution in the former, that is genes becoming pseudogenes or inactive as functions are no longer required in highly specific niches. In *M. leprae* the total number of genes encoded by the genome is about half (1602 genes) that of *M. tuberculosis,* with 27% of genome coding for pseudogenes with functional counterparts in *M. tuberculosis* [Cole *et al.*, 2001].

About 1,400 genes where found to be common between *M. leprae* and *M. tuberculosis,* of which, 333 were common among *Actinomycetes* and 219 were specific to the mycobacterium complex organisms (*Mycobacterium tuberculosis* strains CDC1551 and H37Rv, *Mycobacterium bovis*, *Mycobacterium leprae* and *Mycobacterium paratuberculosis avium*) [Cole *et al.*, 2001; Tekaia *et al.*, 1999]. Since *M. tuberculosis* and *M. leprae* are both intracellular pathogens, some of these genes might play an important role in the ability of *M. tuberculosis* to survive in the host [Cole, 2002].

Moreover, amino acid sequence comparative analysis on a region corresponding to the PE and PPE protein families resulted in the identification of the variability and possible roles of these multiple gene families when the PE genes of *M. tuberculosis* strains H37Rv and CDC1551 were compared. It was discovered that the genes encoding a PE domain alone, or a PE domain followed by a unique protein sequence, were identical in both genomes [Banu *et al.*, 2002; Betts *et al.*, 2000].

In contrast, Banu *et al.* [2002] found 39 of the 62 regular PE-PGRS proteins displayed variability as a result of in-frame insertion or deletion of diverse Ala, Gly-rich coding sequences in the PGRS component of the gene, or harbored frame shift mutations. This seems to occur with no loss of enzymatic activity due to available evidence that the antibodies cross reacted with more than one member of PE-PGRS proteins [Cole, 2002]. Potential structural functions were also suggested for this family, as comparative analysis observed similarity between the PE-PGRS protein and some structural proteins of insects [Cole, 2002].

Comparative analysis between these families of proteins and amino acid sequences of tandem repeats that have been found to be conserved in some other bacteria and *Archaea* may be useful for detecting related proteins from other genomes, and to assist in designing suitable experiments to test their potential functions. Equally important is an increased awareness of the diversity and complexity within the genus of mycobacteria. This was obtained from the application of comparative studies and molecular techniques to the analysis of isolates from environmental and clinical sources. Various evidences have shown that gene loss occurred at a high rate within species of the *M. tuberculosis* complex as a result of homologous recombination events [Brosch et *al.,* 2000; Fleischmann *et al.,* 2002].

Fleischmann and co-workers [2002] compared multiple genome sequences of Mycobacteria, including *M. bovis* and *M. tuberculosis* clinical and laboratory stains. Figure 2.8, taken from their paper, shows a circular representation of the *M. tuberculosis* genome showing the locations of each predicted protein coding region. They observed a region with different numbers of genes, which revealed contradictory evolutionary relationships among CDC1551, H37Rv and *M. bovis*.

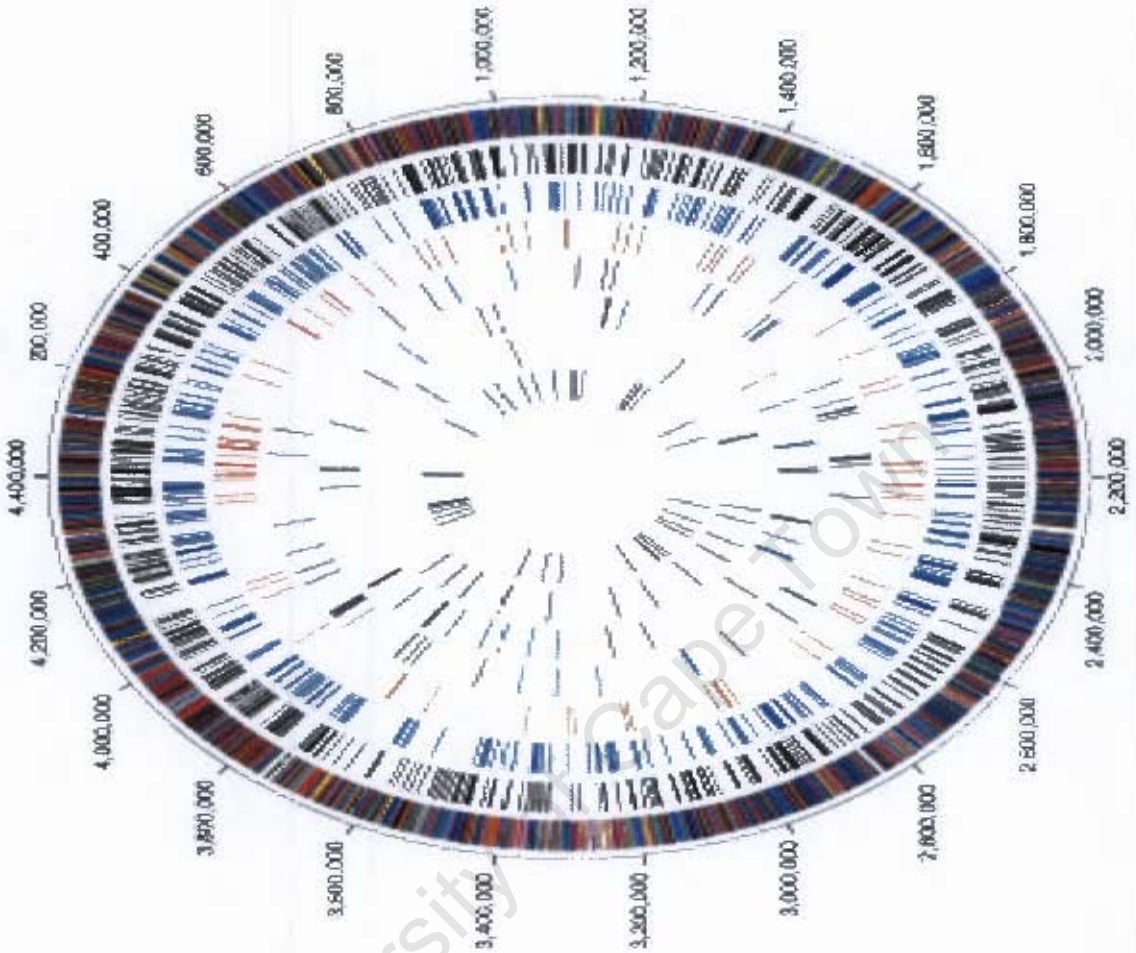

Figure 2.8: The circular representation of the *M. tuberculosis* genome. [Fleischmann *et al.*, 2002]

Brosch *et al.* [2002] also found the presence of intact *mmpS6* and *mmpL6* genes in *M. bovis*, and a truncated region corresponding to these genes in most *M. tuberculosis* strains. These show that genetic variability in *M. tuberculosis* arises through complex evolutionary processes that involve recombination of multiple insertion-deletion events occurring independently at the same locus, suggesting that popular theories of the origin of tuberculosis coinciding with the domestication of cattle are unlikely to be correct [Fleischmann *et al.*, 2002].

There has also been contradictory evidence in sequence polymorphisms among various *M. tuberculosis* complex organisms. For instance, some work recorded the ratio of non-synonymous substitutions to synonymous substitutions and single nucleotide polymorphisms (SNPs) to be silent and much lower than 1.0 [Wade *et al,* 2004; Scorpio and Zhang, 1996] compared to the 1.0 in 2000 to 4000 base pairs that is generally observed among bacteria [Sreevatsan *et al.*, 1997].

However, Fleischmann and his colleagues discovered a high degree of sequence divergence between multiple genome sequences of *M. tuberculosis* and *M. bovis*. *M. tuberculosis* clinical and laboratory stains had an unexpectedly high ratio of non-synonymous to synonymous mutations across all coding sequences compared to what was previously observed in comparative analysis among *M. tuberculosis*. They further observed most of these differences in other clinical *M. tuberculosis* isolates tested, suggesting a possible evolutionary hypothesis that the divergence of *M. bovis* and *M. tuberculosis* was so recent that there has been insufficient time for purifying selection to operate against non-synonymous mutations [Fleischmann *et al.,* 2002]. Another important comparative phylogenetic study of *Mycobacterium tuberculosis* species is the resent study by Filliol and his colleagues [Fillol *et al.,* 2006] who analyzed various clinical and major *M. tuberculosis* and *M.bovis* species, single nucleotide polymorphism (SNP) markers to classify clinical strains from various geographical sites into SNP cluster groups.

Based on their analysis they observed that *M. tuberculosis* has a stable association with human host populations. For instance, strains predominant in an environment depend on the origin and diversity of people in the environment, for example they discovered that the United States *M. tuberculosis* isolates showed more diverse cluster groups due to large immigrant populations compared to isolates obtained from the same indigenous community or closed populations.

3

Comparative Genomics Analysis

The rapidly emerging field of comparative genomics has already yielded dramatic results. ..... study comparing the fruit fly genome with the human genome discovered that about 60 percent of genes are conserved between fly and human. ..., the two organisms appear to share a core set of genes.

- NHGRI, 2005

The first major objective of this project is to generate a phylogenetic profile to catalogue microbial genes and answer the following questions:

- What is the phylogenetic profile of all microbial genes?
- What genes are common to all pathogens and absent from non-pathogens, and which genes are unique to *M. tuberculosis*?

3.1 An Overview

In this chapter we provided a phylogenetic profile of all 84 bacterial genomes involved in this project and used the output to:

- catalogue all 246573 proteins into a phylogenetic profile
- identify genes that are conserved in all bacteria studied
- identify genes that are unique to pathogens
- identify genes that are unique to *Mycobacterium tuberculosis* complex organisms
- Identify genes that are potentially involved in virulence in the reference genome (*M. tuberculosis* strain CDC1551).

To achieve these goals, it was necessary to identify sets of homologous clusters of proteins between the 84 complete genomes involved and examine the protein clusters created for conservation of protein description lines, available functional information and protein family and domain composition using different strategies. Various homology

approaches were tested and two sequence similarity-based methods at different levels of stringency were decided on to create three sets of phylogenetic profiles of all species involved in the experiment.

We identified sets of homologous clusters of proteins between 84 complete bacterial genomes by performing pair-wise comparisons between them using the BLASTP algorithm [Altschul, 1990]. This was followed by clustering proteins into related sets based on E-value and percentage overlap between two sequences or using the NCBI BLASTClust program. This program clusters protein sequences into sets of related proteins based on sequence similarity, E-value, percentage sequence identity or score density and percentage overlap between sequences [http://biowulf.nih.gov/apps/blast/doc /blastclust.html]. A phylogenetic profile was generated for each protein cluster in the different experiments. The profiles were re-ordered to select the genes of interest.

3.2 Background

Comparative genomics analysis is a powerful tool for investigating evolutionary changes among (and sometimes within) organisms. It enables the identification of the genes (of interest) that are conserved among species as well as the genes that give each organism its own unique characteristics.

Experimentally determined functions are only known for a very small fraction of the proteins in sequenced genomes. There is a need for automated methods of transferring knowledge from well studied organisms to less known organisms, given the increased amount of sequence data from completely sequenced genomes. Since *in vivo* experiments of all genes are not feasible, the ability to infer the function of genes from the function of corresponding genes in model organisms is highly desirable. So far, the genomes from more than 150 genomes of bacteria have been fully sequenced. Of particular interest for human medical research are the full genome sequences of human and human pathogens, such as *Mycobacterium, Bacillus*, *Salmonella*, *Haemophilus*, *Clostridium* species, as well as various non human pathogens.

Many approaches have been used to infer functional linkages between genes within the genome of the same and different organisms [Groenen *et al*, 2006]. Using phylogenetic profiles is a common approach to infer function, comparable to the study of comparative genomics using the evolutionary concept of homology. To make comparative genomics meaningful, there is an inevitable need to introduce a natural way to classify sets of

genes in different genomes that might be functionally equivalent, that is, orthologs or co-orthologs. This will help in reducing individual experimental analysis to be performed to the number of identifiable clusters of orthologs in genomes being compared [Koonin, 2005].

The term **homolog** refers to genes in different species that are derived from a common ancestor Homologous genes are usually observed on the basis of statistically significant sequence similarity to a gene or protein of interest. **Orthologs** are homologous genes or protein sequences that evolved by vertical descent from a single ancestral gene, in different species that are derived from a common ancestor. They usually occur as a result of a speciation event. Orthologous genes may or may not have the same function, but are usually assumed to have maintained equivalent functions. History of the gene reflects the history of the species; here phylogeny of genes represents the true phylogeny of species. **Paralog** is used to describe homologous genes within a single species that result from a duplication event within a genome. It usually leads to functional specialization in which one of the genes evolves a new function through random mutation, or both genes share the function of the original gene [Fitch, 1970].

Ortholog and paralog formation might be considered to be the primary events of genome evolution and have been well recognized in the pre-genomic era [Koonin, 2001]. The concept of orthology arises as a result of describing evolutionary relationships with accuracy [Fitch, 2000]. It is important in inferring gene functional conservation as a consequence of orthology based on sequence similarity. Functional equivalence of orthologous genes, which is one of the important properties of orthologs, is theoretically plausible and has been experimentally supported in various studies [Amos *et al.*, 2004; Koonin, 2005]. These genes typically perform similar functions in the respective organisms.

Orthology identification is further complicated by other primary elements of gene evolution such as *in-paralogs*, paralogs that occur after speciation; and *out-paralogs*, paralogs that precede speciation [Sonnhammer *et al*, 2001]; horizontal gene transfer; gene loss; gene fission and fusion and gene rearrangement [Doolittle, 2000; Koonin, 2005]. Since orthology and paralogy are evolutionary events, the best method of identifying orthologs is through phylogenetic analysis, or tree reconciliation [Page and Charleston, 1997; Eulenstein *et al.*, 1998; Mirkin *et al.*, 1995], where a gene tree is

compared with a chosen species tree. This is carried out using the parsimony reconciliation principle through selecting the number of allowed minimum duplication and gene loss events. This approach is expected to show orthologous relationships, however there are major shortcomings with this method of orthology inference, principally among which is the prevalence of gene acquisition through horizontal gene transfer, *xenology*.

Xenolog describes the relationship between two genes in which one has been derived by horizontal gene transfer. This phenomenon is very common in prokaryotes and makes tree reconciliation a difficult task [Ragan, 2001; Garcia-Vallve *et al.*, 2000; Koonin *et al.*, 2001; Doolittle, 1999], leading to the use of consensus trees from various genes. This is done, for example, by construction of several individual gene trees, and comparing the trees to produce a consensus tree [Doolittle, 1999; Koonin, 2005; Doolittle, 2000; Wolf *et al.*, 2002; Daubin *et al.*, 2002; Bininda-Emonds *et al.*, 2002].

Eukaryotic phylogeny also faces the major issue of uncertainty of artefacts, a problem relating to the evolutionary clock which is the dating of speciation events and gene fission and fusion (where a protein might be a part of multi domain protein and vice versa). While phylogenetic analysis can help to resolve some of these issues, the major problem of phylogenetic comparative analysis is the inability to automate the process, and the high cost of the computational analysis.

3.2.1 Principles and Techniques for Identification of Orthologs and Paralogs

In solving the above problems, most computational comparative analyses of genomes resort to a more straight forward sequence similarity method to infer orthologs between the genomes. The assumptions are that orthologs (genes) sequenced from different species are more similar to each other in sequence than paralogs from the same or different species [Tatusov *et al.*, 1997; Koonin, 2005]. The level of functional conservation between orthologous proteins makes orthology widely used in genome analysis for protein function prediction and annotation, where the information about a statistically significantly similar protein sequence from a well studied species is used for annotation of the orthologous protein in another species [Groenen *et al.*, 2006].

The concept of identification of orthologous relationships between genes is probably the fastest way of transferring functional information from numerous model organisms that

are well studied and can be used to elaborate on gene functions for most completely sequenced genomes. Inferring orthology, coupled with further annotations like structural similarity and expression data, can be used to ascertain functional similarity between genes. For example, the level of protein-protein interactions, allows networks of orthologous sequences to be investigated to detect conservation of processes and pathways [Groenen *et al.*, 2006].

The concept of the *Reciprocal Best Hit*, (RBH) technique is the most frequently applied method to infer ortholog pairs. In this method, orthologous genes form reciprocal best hits (RBH) when the genomes are compared. For example, gene X used as query sequence from genome A has the highest sequence similarity score with gene Y in genome B and gene Y as query has gene X as its best score. This assumption is based on the evolution of orthologs and the possibility of them occupying the same functional niche in their individual genomes [Koonin, 2005]. Though the sequence similarity approach faces various problems of false positive and negative results due to the inherent evolutionary problems of out-paralogs, in-paralogs [Koonin, 2001; Sonnhammer, 2001] and xenologs, genes forming RBH in terms of functional similarity are still considered to be orthologs due to their species to species origin of transfer.

While out-paralogs might not be regarded as orthologs, in-paralogs are considered to be orthologs since duplication occurs after speciation. There is practical evidence of genes from closely related species forming RBH [Tatusov *et al.*, 1996]. For example, Koonin [2005] shows that RBH is common among prokaryotic genomes and it decreases as the evolutionary distance between organisms increases. As mentioned previously, however, the conventional RBH method faces the issue of in-paralogs when comparing two genomes and it also becomes more difficult to rely on when comparing multiple genomes due to the complex mix of in- and out-paralogs.

Various other methods for identification of orthologs, based on specially designed sequence clustering procedures, explicit phylogenetic analysis, or a combination of both have been developed to solve these problems and better unravel orthologs and paralogs [Tatusov *et al.*, 1997; Sonnhammer *et al,* 2001; Li *et al,* 2003; Zmasek and Eddy, 2002]. A Cluster of Orthologous Groups (COG) is the first representative of such systems [Tatusov *et al.*, 1997]. The COG (and KOG) database is a collection of BLAST-based ortholog groups from multiple species with further manual annotations.

It is based on the assumption that any set of at least three proteins from fairly distant genomes that are more similar to each other than they are to any other proteins from the same genomes are the most probable orthologs. Here the notion of a genome-specific best hit was extended to multiple genomes such that the algorithm sought to distinguish clusters of triangles of mutually consistent, genome-specific best hits. Highly similar proteins from the same genome that are more similar to each other than they are to any proteins from other species are treated as one [Tatusov *et al.*, 2001].

Another popular approach developed by the *Inparanoid* group formed by Sonnhammer and co-workers, identified orthologs and in-paralogs between two genomes A and B, by determining all possible pair-wise similarity scores between genome A-A, B-B, A-B and B-A that score higher than a preset cut-off of BLAST bit score and overlap. Then, the RBH, are marked as potential orthologs. The in-paralogs that score higher than these orthologs are then regarded and marked as additional orthologs with a confidence value chosen upon the developed statistical criteria. The quality of in-paralogs can also be assessed by an out-group proteome score [Remm, 2001]. OrthoMCL [Li *et al.*, 2003] also takes the result of all-against-all BLASTP and group orthologs and paralogs with a Markov Cluster algorithm based on probability and graph flow theory. It allows consecutive classification of global relationships in a similarity space.

The steps involve all-against-all BLASTP, after which, the reciprocal best hits between species are marked as putative orthologs and the reciprocal better similarity hits within species that are better than RBH are marked as recent paralogs. A similarity matrix is calculated, followed by a Markov clustering which determines the orthologous groups [Li *et al.*, 2003]. Another effort by Sonnhammer and co-workers uses a phylogenomic procedure for inference of orthologs by comparing gene trees with species trees, and selecting the subset of the gene tree that has the same topology as the species tree, as orthologs [Storm and Sonnhammer, 2002].

Zmasek and Eddy [2002] used the output generated via multiple alignments and subsequent tree calculation to automate the ortholog detection procedure. Other groups use pair-wise sequence similarity or phylogenetic trees coupled with genome positional information to cluster genes into ortholog sets [Overbeek *et al.*, 1999; Cannon and Young, 2003; Tekaia and Yeramian, 2005; Chen, 2006], with the assumptions that sets

of genes from different organisms that occupy the same regions and are similar in organization are orthologs.

In this work, we decided not to use reciprocal best hit (RBH) methods or other ortholog searching algorithms for a number of reasons. Firstly, as mentioned above, ortholog and paralog prediction is not straight forward, particularly in prokaryotes, where this is complicated by horizontal transfer and other genetic rearrangements, as well as in- and out-paralogs. Secondly, our interest is on proteins that are in any way functionally related, not just strict orthologs. These include proteins that might be important for pathogenicity and virulence of *Mycobacterium tuberculosis,* which might include both orthologs and paralogs from the same and different organisms. Whereas ortholog prediction is used primarily for functional annotation and thus needs to be strict, in this study we wanted to identify protein sets that have some potential to now or previously share a similar function, without missing any potential functional links.

Finally, since the evolution of *M. tuberculosis* is thought to be relatively recent, it is possible that paralogs would not yet have had sufficient time to evolve completely new functions. We therefore decided to use BLAST and clustering algorithms and confirmation by InterPro matches (protein signatures are able to detect functionally related proteins, and those that are more distantly related), and not try to separate closely related orthologs and paralogs. In this way, when we attempt identify proteins unique to pathogens, they should truly be unique in sequence and thus function.

3.2.2 Possible problem with using BLAST to infer Orthologs

Orthologs based on BLAST are usually selected based on E values, which depends on (and is affected by) the length of the match sequences. Since BLAST generates several *high scoring segment pairs* (HSPs) for each pair of genes and the HSPs may overlap with each other, a simple addition of their scores may overestimate the similarity between the two genes.

The best hit for a particular protein may be a local domain hit of a multiple domain protein, which may lead to predicted genes being gene fragments and not the true length of the gene. These problems were solved by calculating the percentage alignment coverage and selecting orthologs with ≥50% coverage for both sequences.

3.3 Tools Selected for Generating Clusters

3.3.1 Sequence Alignments

Biological sequence alignment is the method of comparing two (pair-wise alignment) or more (multiple sequence alignment) biological sequences by searching for a series of individual characters or character patterns that are in the same order in the sequences [Mount, 2001]. Alignment methods are further divided into global and local alignments. In global alignments an attempt is made to match the entire sequence onto each order using as many characters as possible, extending to both ends of each sequence. Sequences that are similar and more or less the same length are suitable candidates for global sequence alignment.

In local alignments, stretches of sequences with the highest density of matches are aligned, thus generating one or more islands of matches or sub-alignments in the aligned sequences. Local sequence alignments are suitable for aligning sequences that are similar along some of their lengths, but dissimilar in other regions and sequences that differ in length or sequences that share conserved domains [Mount, 2001]. The primary sequence alignment task is to ask if sequences are evolutionarily, structurally or functionally related to each other.

3.3.2 Alignment Algorithms used

BLAST - The Basic Local Alignment Search Tool is a tool that uses a heuristic algorithm to find regions of high local similarity between a protein or DNA sequence and a database of sequences [Altschul *et al.*,1990], and calculates the statistical significance of the matches. It makes a list of all fixed length words (3 for protein and 11 for nucleotides) that align with the query sequence with at least some pre-set threshold, and then searches through the database to produce high scoring pairs HSP, that have a score of at least the pre-set threshold to produce an un-gapped alignment in both directions [Durbin *et al.*, 1998]. The statistical significance of the recorded HSP is evaluated to determine whether the match score recorded is higher than what is expected to occur by random chance.

BLASTP - BLASTP uses the BLAST algorithm to compare a query protein sequence against a database of protein sequences. Like other BLAST programs, the BLASTP algorithm is optimized to find local regions of similarity, but to report a global alignment, when sequence similarity cuts across the whole sequence [Altschul *et al.*, 1990].

BLASTClust - BLASTClust is a single-linkage clustering method to group proteins or nucleotide sequences based on pair-wise similarity found using the BLAST algorithm. The program accepts as input a file of protein sequences in FASTA format, each with a unique sequence identifier. It returns a file of sequence identifiers arranged in clusters. It uses the BLASTP algorithm for proteins and MegaBLAST algorithm for DNA sequences [http://www.ncbi.nlm.nih.gov/Web/Newsltr/Spring04/ blastlab.html].

For each pair of sequences the top-scoring alignment is evaluated based on several pre-set parameters, like thresholds, T for score density, S; percent identity, *p* and alignment length, *b* to control the stringency of clustering. Two sequences are considered to be neighbours if the coverage and the score density are over or equal to the pre-set threshold. It then assigns a sequence to a cluster if the sequence is a neighbour to at least one sequence in the cluster [http://www.ncbi.nlm.nih.gov/Web/Newsltr/Spring04 /blastlab. htm].

3.3.3 Data Sets - Sequence Data

The data collection steps involved selection and downloading, in FASTA format, of the whole genomes of 56 pathogens, including *Mycobacterium tuberculosis* strain CDC 1551 (reference genome) and H37Rv. Also included were 3 other species of mycobacteria and 28 non-pathogenic bacteria including gram positives and alpha, beta, delta, gamma and epsilon proteobacteria from the TIGR website [http://www.tigr.org/tigr-scripts/CMR2/CMR Home_Page.spl] (Tables 3.1a and 3.1b). Non redundant protein sets for the 84 selected bacterial genomes were also downloaded from Integr8 project of the European Bioinformatics Institute (EBI) [htttp://www.ebi.ac.uk/integr8].

3.4 Methods

More than one approach was used to generate sets of homologous genes from the 84 genomes for the phylogenetic profiles. We used a combination of publicly available software and custom scripts to generate three different datasets. In the first method, pair-wise sequence similarity searches of all non redundant protein sets were carried out by creating a local database using formatDB according to NCBI. For every protein in each genome, we ran a BLASTP search [Altschul *et al*, 1990] against all the remaining 84 genomes to identify possible homologs with the aid of Python scripts. Homolog sets were selected according to the following criteria:

- Blast thresholds set with E-value of less than or equal to 1e-6 (0.000001)
- Percentage [%] alignment coverage: an overlap across the query and hit proteins length of greater than or equal to 50%, 70% and 90% were tested.

A group of homolog sets that satisfied the above two criteria were created from the first genome, generated from the BLAST output. For each experiment, each cluster was then labelled with the aid of a specific identifier. These steps were repeated until all the proteins from all genomes had been assigned to a group using each protein as a query protein. Redundancy was filtered out by making sure that no protein belonged to two or more groups. This was achieved by assigning a protein to group in which it had highest significant match. For these homolog sets we then checked whether they contained more than one protein from the same organism, that is, potential paralogs, and decided to treat the set of paralogs as one gene. The above method is referred to as Myblastclust.

3.4.1 Identification of Clusters with BLASTClust

The second and third datasets were generated using the Basic Local Alignment Search Tool protein clustering program (BLASTClust) from NCBI [http://www.ncbi.nlm.nih.gov/BLAST/]. BLASTClust was used to cluster the 246573 protein sequences from all 84 genomes into clusters of similar homolog sets, with the default E-value of 1e-6 and parameters –p, T, -b, F and –S. We selected Score Density option (S) ranging from 0 to 2 over an area covering 50%, that is -b value of 0.5 for both sequence lengths. For each pair of sequences the Score Density,

$$S = -\frac{N}{AL'}$$

where, L' is length of sequence in the alignment, L; N is the number of identical residues and A is the total alignment length which is equal to L plus number of gaps in the alignment. The cluster outputs were examined and clusters with 0.5S and 1.0S were selected for further processing.

Table 3.1a: Organism phylogeny-pathogens

TAX ID	Organism code	*Common name*	Strain	Gram	Description
83332	MYCLEH	*Mycobacterium tuberculosis*	H37Rv	positive	High GC, G+
83331	MYCTU	*Mycobacterium tuberculosis*	CDC1551	positive	High GC, G+
233413	MYCBV	*Mycobacterium bovis*	AF2122/97	positive	High GC, G+
262316	MYCPA	*Mycobacterium paratuberculosis*	k10	positive	High GC, G+
272631	MYCLE	*Mycobacterium leprae*	TN	positive	High GC, G+
257309	CORDI	*Corynebacterium diphtheriae*	NCTC13129	positive	High GC, G+
267747	PROAC	*Propionibacterium acnes*	KPA171202	positive	High GC, G+
203267	TROWT	*Tropheryma whipplei*	Twist	positive	High GC, G+
169963	TROW8	*Troplerma whipplie*	two/27	positive	High GC, G+
195102	CLOPE	*Clostridium perfringens*	13	positive	Low GC, G+
212717	CLOTE	*Clostridium tetani*	E88	positive	Low GC, G+
261594	BACAN	*Bacillus anthracis*	Ames	positive	Low GC, G+
222523	BACC1	*Bacillus cereus*	ATCC10987	positive	Low GC, G+
158878	STAAM	*Staphylococcus aureus*	Mu50	positive	Low GC, G+
12228	STAES	*Staphylococcus epidermidis*	ATCC	positive	Low GC, G+
272626	LISIN	*Listeria innocua*	CLIP 11262	positive	Low GC, G+
169963	LISMO	*Listeria monocytogenes*	GSC1	positive	Low GC, G+
243273	MYCGE	*Mycoplasma genitalium*	G-37	positive	Low GC, G+
272634	MYCPN	*Mycoplasma pneumoniae*	M129	positive	Low GC, G+
272633	MYCPE	*Mycoplasma penetrans*	HF-2	positive	Low GC, G+
171101	STRR6	*Streptococcus pneumoniae*	R 6	positive	Low GC, G+
208435	STRA5	*Streptococcus agalactiae*	2603V/R	positive	Low GC, G+
210007	STRMU	*Streptococcus mutans*	UA159	positive	Low GC, G+
700294	STRP1	*Streptococcus pyogenes*	SF370 /ATCC	positive	Low GC, G+
226185	ENTFA	*Enterococcus faecalis*	V583	positive	Low GC, G+
243274	THEMA	*Thermotoga maritima*	MSB8	negative	Thermatogs
272561	CHLTR	*Chlamydia trachomatis*	serovar D	negative	Chlamydia
243161	CHLMU	*Chlamydia muridarum*	Nigg	negative	Chlamydia
115713	CHLPN	*Chlamydia pneumoniae*	CWL029	negative	Chlamydia
227941	CHLCV	*Chlamydophila caviae*	GPIC	negative	Chlamydia
198214	SHIFL	*Shigella flexneri*	301	negative	Gamma Proteo
83334	ECO57	*Escherichia coli*	O157:H7 VT2-Sakai	negative	Gamma Proteo
187410	YERPE	*Yersinia pestis*	KIM5	negative	Gamma Proteo
273123	YERPS	*Yersinia pseudotuberculosis*	IP32953	negative	Gamma proteo
99287	SALTY	*Salmonella typhimurium*	LT2 SGSC1412	negative	Gamma Proteo
601	SALTI	*Salmonella typhi*	CT18	negative	Gamma Proteo
32741	SALCH	*Salmonella choleraesuis*	SC-B67	negative	Gamma Proteo
233412	HAEDU	*Haemophilus ducreyi*	35000HP	negative	Gamma Proteo
71421	HAEIN	*Haemophilus influenzae*	KW20 Rd	negative	Gamma Proteo
272843	PASMU	*Pasteurella multocida*	PM70	negative	Gamma Proteo
243277	VIBCH	*Vibrio cholerae*	El Tor N16961	negative	Gamma Proteo
223926	VIBPA	*Vibrio parahaemolyticus*	RIMD 2210633	negative	Gamma Proteo
196600	VIBVY	*Vibrio vulnificus*	YJ016	negative	Gamma Proteo
243233	METCA	*Methylococcus capsulatus*	Bath	negative	Gamma Proteo
242231	NEIGO	*Neisseria gonorrhoaeae*	FA1090	negative	Beta proteo
122586	NEIME	*Neisseria meningitidis*	MC58	negative	Beta proteo
520	BORPE	*Bordetella pertussis*	Tohama I	negative	Beta proteo
257311	BORPA	*Bordetella parapertussis*	12822	negative	Beta proteo
243160	BURMA	*Burkholderia mallei*	ATCC:23344	negative	Beta proteo
12472	CHRVO	*Chromobacterium violaceum*	ATCC:12472	negative	proteo
782	RICPR	*Rickettsia prowazeki*	Madrid E	negative	proteo
781	RICCN	*Rickettsia conorii*	Malish 7	negative	proteo
29461	BRUSU	*Bucella suis*	1330	negative	proteo
29469	BRUME	*Brucella melitensis*	16M	negative	proteo
11168	CAMJE	*Campylobac jejuni*	NCTC11168	negative	proteo
85963	HELPJ	*Helicobacter polari*	J99		

Table 3.1b: Organism phylogeny-non pathogens

TAX ID	Organism code	Common Name	Strain	Gram	Description
63363	AQUAE	*Aquifex aeolicus*	VF5	negative	Thermophile
1299	DEIRA	*Deinococcus radiodurans*	R 1	positive	Deinoc/therm
1718	CORGL	*Corynebacterium glutamicum*	ATCC 13032	positive	High GC, G+
152794	COREF	*Corynebacterium efficiens*	YS-314	positive	High GC, G+
1902	STRCO	*Streptomyces.coelicolcr*	A32	positive	High GC, G+
33903	STRAW	*Streptomyces.avermitilis*	nil	positive	High GC, G+
216816	BIFLO	*Bifidobactarium.longum*	NCC2705	positive	High GC, G+
1488	CLOAB	*Clostridium acetobutylicum*	ATCC824	positive	Low GC, G+
1423	BACSU	*Bacillus subtilis*	168	positive	Low GC, G+
86665	BACHD	*Bacillus halodurans*	C-125	positive	Low GC, G+
182710	OCEIH	*Oceanobacillus iheyensis*	HTE831	positive	Low GC, G+
1590	LACPL	*Lactobacillus plantarum*	WCFS1	positive	Low GC, G+
33959	LACJO	*Lactobacillus johnsonii*	NCC 533	positive	Low GC, G+
1360	LACLA	*Lactococcus lactis*	lactis IL1403	positive	Low GC, G+
97948	PSEAE	*Pseudomonas aeruginosa*	nil	negative	Gamma Proteo
76868	PSEPK	*Pseudomonas putida*	KT2440	negative	Gamma Proteo
98794	BUCAP	*Buchera aphidicola*	sg	negative	Gamma Proteo
118099	BUCAI	*Buchnera sp.*	APS	negative	Gamma Proteo
562	ECOLI	*Escherichia coli*	K12	negative	Gamma Proteo
36870	WIGBR	*Wigglesworthia glossinidia*	brevipdpis	negative	Gamma Proteo
915	NITEU	*Nitrosomonas europaea*	ATCC 19718	negative	Beta proteo
155892	CAUCR	*Caulobacte crescentus*	CB15	negative	Alpha proteo
375	BRAJA	*Bradyrhizobium japonicum*	USDA 110	negative	Alpha proteo
882	DESVH	*Desulfovibrio vulgaris*	Hildenborough	negative	Delta proteo
35554	GEOSL	*Geobacter sulfurreducens*	PCA	negative	Delta proteo
844	WOLSU	*Wolinella succinogenes*	DSMZ 1740	negative	Epsilon proteo
21948	SILPO	*Silicibacter pomeroyi*	DSS-3	negative	Epsilon proteo

3.4.2 Creating Matrix and Ordering of Result

The results from the clustering experiments were entered into a binary matrix of zeros, 0 and ones, 1 as well as additional matrices with the actual number of proteins, with the aid of PYTHON scripts. In each matrix the cluster identifiers are down the first column while the organisms' codes are across the first row. In the binary matrix, for each organism, 1 is entered into the column of that organism if it has a protein present in the corresponding cluster and 0 if it does not, and the case of the other matrices, the number of proteins present in the cluster was entered.

The total number of proteins from all organisms, and the total number from pathogens and non-pathogens present in each cluster were evaluated by summing up across corresponding rows. A summary matrix was then created to include additional information like average number of proteins per cluster and presence or absence of the reference organism in each cluster. The matrices were ordered to create static summary tables of the total number of clusters, number of single protein clusters, number of pathogen only clusters, number of non-pathogen only clusters, number of MTB complex only clusters and number of shared clusters by all genomes in the project.

3.4.3 Selection of genes common to Pathogens and *M. tuberculosis*

The columns of the matrices were clustered into pathogens and non pathogens, with each set having its total derived from adding together the numbers across each row. The rows were then ordered in ascending numerical order based on the total column for non-pathogens, followed by ordering the total column for the pathogen set in descending numerical order. For the binary matrix, the rows of interest were those containing a total of 0 for the non-pathogen set and n, $n-1$, $n-2$, $n-3$ *etc*, for the pathogen set, where n, is the number of pathogenic organisms used for the experiments. In this instance, n is selected as those proteins common to pathogens and absent from non-pathogens and is chosen to be 56. For the numerical matrices, rows with 0 in the total column for non-pathogens and >1 for the total proteins from pathogens were selected. Those pathogen clusters unique to the MTB complex organisms were also selected for further analysis.

3.4.4 Generation of Virulence Gene Test Sets

Information on known microbial virulence genes was retrieved using the Sequence Retrieval System (SRS) searching the UniProtKB database [http://srs.ebi.ac.uk/srsbin/],

and Pubmed abstracts from NCBI [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=Pager&DB=Pubmed/]. Specific attention was on *Mycobacterium tuberculosis* virulence genes using the following test queries:

- *Mycobacterium tuberculosis* and virulence
- *Mycobacterium tuberculosis* and pathogen
- *Mycobacterium tuberculosis* and host immunity
- Bacterial and pathogen
- Bacterial virulence gene.

3.5 Results and Discussion

3.5.1 Phylogenetic patterns of proteins in the comparative proteome clustering experiments

Three sets of experiments were conducted - NCBI BLASTClust with Score Density of 0.5 (0.5S) and Score Density of 1.0 (1.0S), and Myblastclust with E-value of 0.000001 and 90% alignment coverage (as explained earlier). Results from the three experiments are presented. The results of phylogenetic profiles of the proteins for each experiment were represented in the form of protein number and binary matrices. Figures 3.1 and 3.2 are examples of the results represented as heat maps, which show the number of proteins in each organism for individual clusters by the colour intensity. These were obtained by converting the numeric number of proteins to a colour code gradient. The colour code ranges from green for 0.0, red for 1.0 to yellow for 2 and above; the darker the yellow colour the higher the number of proteins in the cluster for that genome. The complete matrices for all the 84 genomes will be uploaded to the CBIO webpage [http://www.cbio.uct.ac.za].

Figure 3.3 shows a heat map of hierarchical clustering of the organisms for the first 100 clusters using the complete linkage algorithm of the TIGR Multiple Experiment Viewer (MEV) [Saeed *et al.*, 2003]. Here the colour code ranges from green for 0.0, black for 1.0 to red representing 2 and above proteins.

Table 3.2 shows the general statistics resulting from the clustering experiments and Figure 3.4 shows the percentage distribution of cluster types (in terms of organisms represented) in each of the experiments. NCBI BLASTclust with 1.0 score density and 50% alignment coverage created the largest number of protein families with 121905

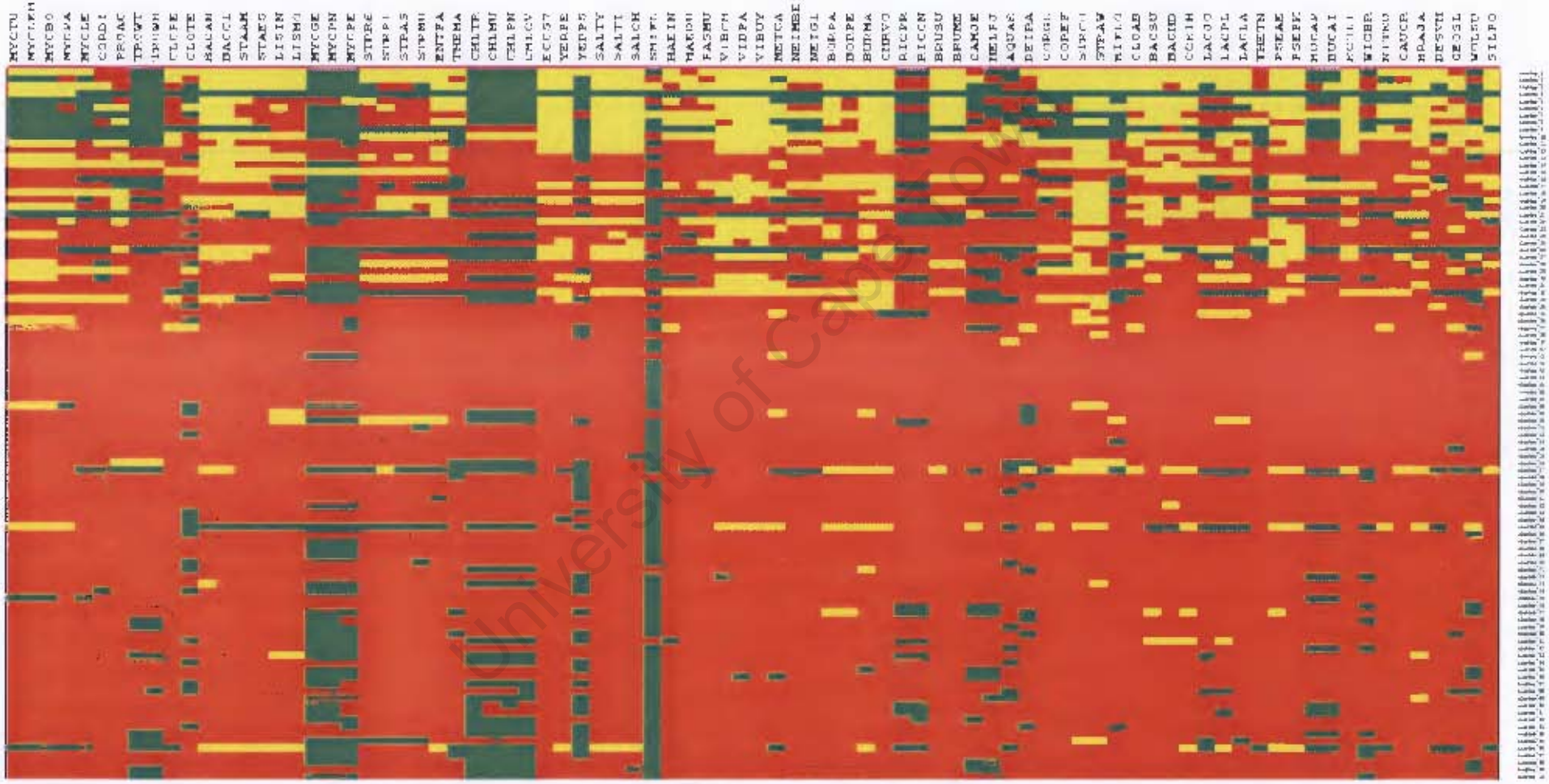

Figure 3.1: Heat map representation of phylogenetic matrix for 1.0S: The colour code ranges from green for 0.0, red for 1.0 to yellow for 2 and above; the darker the yellow colour the higher the number of proteins in the cluster for that genome.

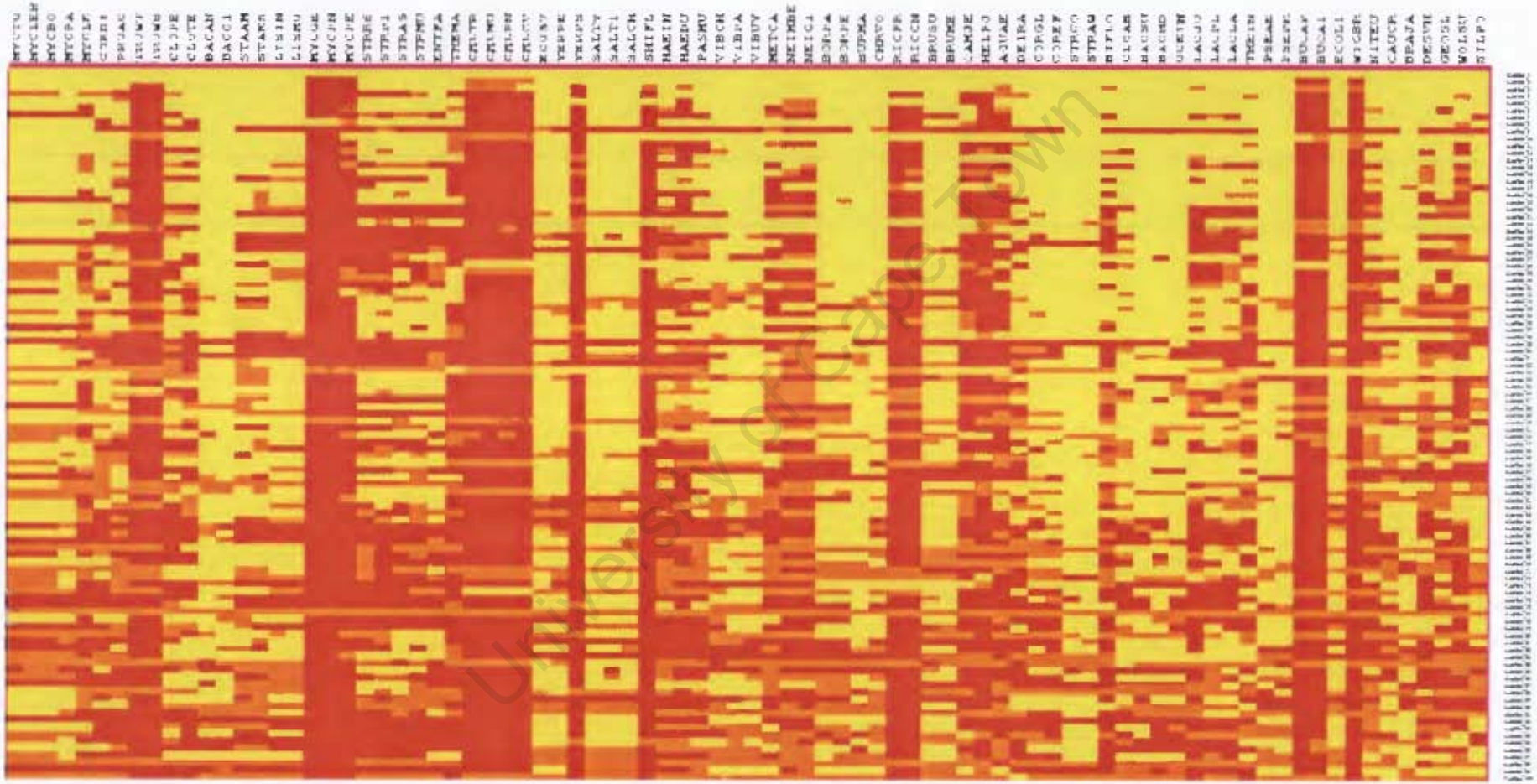

Figure 3.2: Heat map representation of phylogenetic matrix 0.5S: The colour code ranges from green for 0.0, red for 1.0 to yellow for 2 and above; the darker the yellow colour the higher the number of proteins in the cluster for that genome.

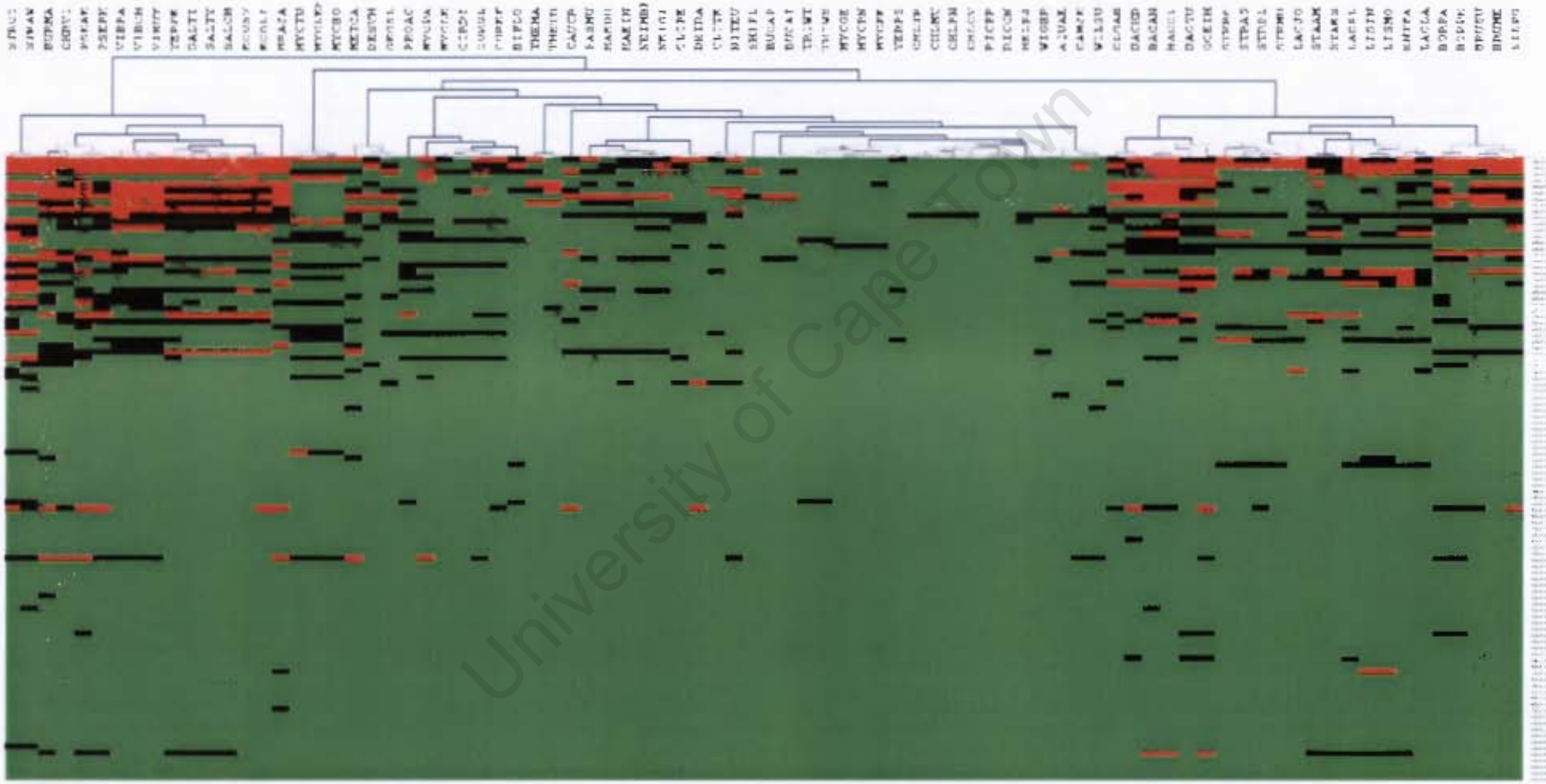

Figure 3.3: Heat map representation of complete lineage hierarchical clustering of organisms

Table 3.2: General statistical distribution of the experiments

Experiments	0.5 score density	1.0 score density	My BLASTClust 90%
Total clusters created	63819	121905	74223
Total Singletons (one protein families)	42770	85735	53959
Total more than one protein clusters	20849	36170	20264
Total Pathogen only clusters	7904	18022	8459
Total non Pathogen only clusters	3824	7165	4232
Total MTB Complex only clusters	1060	2362	1249
All organisms (84 genomes)	10	2	7
All organisms (at least 80 genomes)	135	39	85
All organism (at least 60 genomes)	448	130	397
Number with at least two genomes per cluster (no paralogs)	19056	24110	17045
Number of clusters with at least 2 proteins (paralogs) in one genome	1794	2061	3210

of 246573 proteins from 84 genomes, followed by Myblastclust clustering with expected value (E-Value) 10^{-6} and 90% alignment coverage and the NCBI BLASTclust with 0.5 score density and 50% alignment coverage with the smallest number of clusters of protein families.

The Myblastclust clustering gave the highest percentage of orphan families of proteins with about 73% of all clusters generated being single protein clusters, followed by 1.0S with 70%. The 0.5S results, with the value of 67% of all protein families created being singletons, is the lowest. The reverse is the case with the percentage of clusters with more than one protein from any genome, 0.5S created the highest percentage, and Myblastclust the lowest.

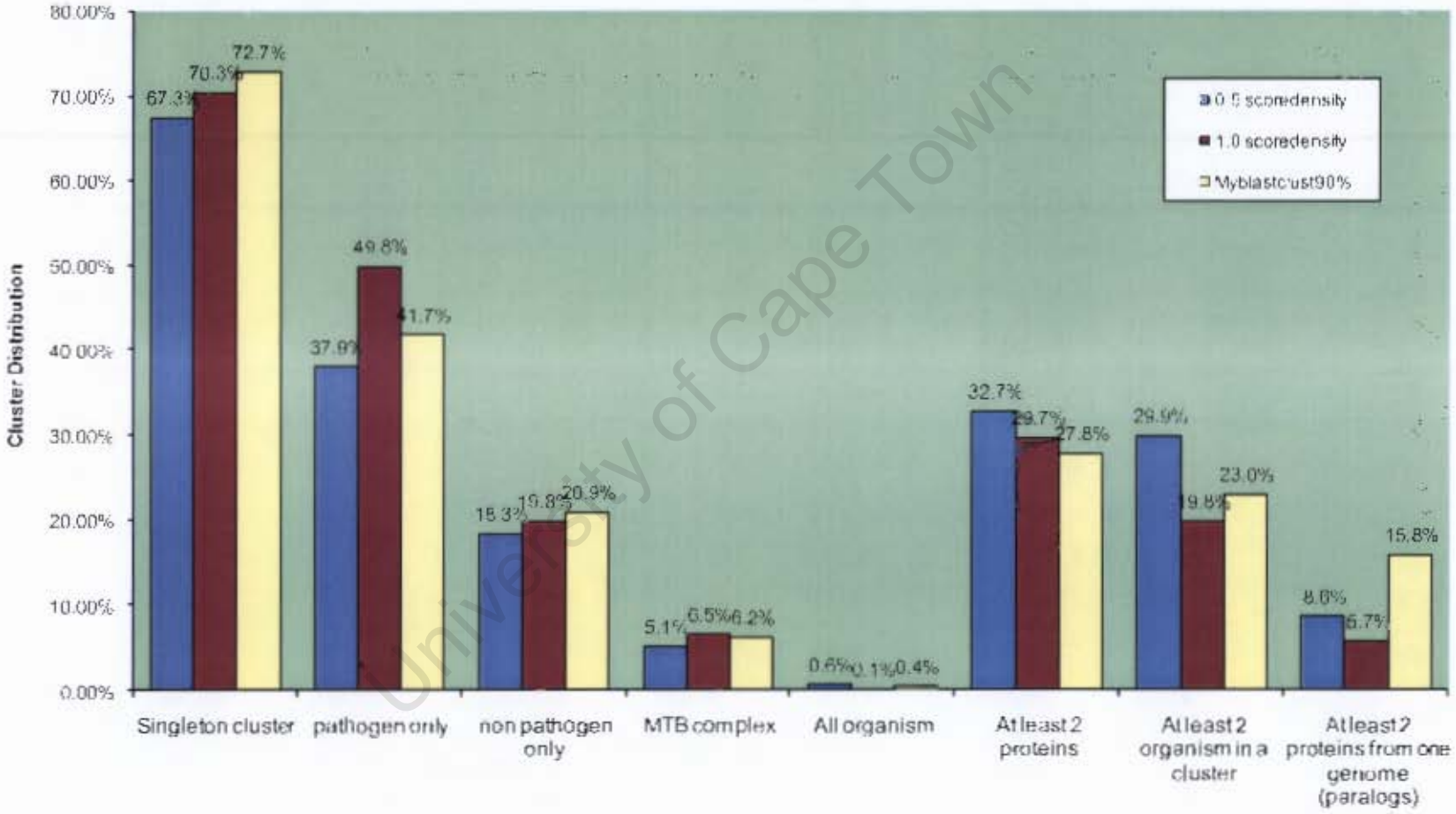

Figure 3.4: Percentage statistical distribution of the experiments

Table 3.3: The result of known virulence bacterial genes from BLASTClust 1.0S

Cluster No	Total Protein in pathogens	Virulence organism code	Total Protein in non pathogens	Average no of proteins	No of non pathogens	Common DE line	Common InterPro
14	1117	SALTY	959	34.25	26	ATP-dependent_Clp	IPR001907
29	617	SALTY	506	18.071	25	RNA polymerase sigma_factor rpoD	IPR011991, IPR007624 IPR007627
131	271	VIBCH	209	7.4643	22	Chaperone_clpB	IPR003959, IPR004176 IPR001270
341	128	LISMO	67	2.3929	25	Aldehyde_dehydrogenase	IPR002086, IPR012303 P23240
366	76	LISMO	46	1.6429	28	Potassium-transporting_ ATPase	IPR008250, IPR005834 IPR001757
522	72	SHIFL	39	1.3929	27	Potassium_binding and translocating	IPR000131
1052	73	STRR6	29	1.0357	19	Zinc-binding_lipoprotein_ adcA_ precursor	IPR006127, IPR006128 IPR006129
2470	51	PASMU	27	0.9643	27	Phosphomannomutase/ phosphoglucomuta	IPR005845, IPR005841 IPR005843
4241	34	BACAN	27	0.9643	15	Virulence_factors_ transcription_ regulator	IPR011991, IPR001789 IPR000792
5166	34	PSEAE	16	0.5714	14	Transcriptional_regulatory_	IPR001789, IPR001867 IPR011006
6264	24	ENTFA	11	0.3929	11	Sensor_protein_ basS/pmrB	IPR003661, IPR003660 IPR003594
8270	271	VBCH	209	7.4643	22	Capsule_biosynthesis_capB	IPR008337
10321	31	YERPS	0	0.0357	0	Capsule_biosynthesis_ protein_capC	IPR008338
10631	4	BORPE	7	0.25	6	Cytotoxic_protein_ccdB	IPR002712, IPR011067
10744	4	ECO57	2	0.0714	2	_ccdA_(Protein_letA)_ (Protein_H)_LynA	IPR009956
16892	2	BACAN	2	0.0714	2	Virulence_sensor_protein_b vgS	IPR001789, IPR001638 IPR003661
30401	2	BACAN	2	0.0714	2	Attachment_invasion _locus precursor	IPR000758
36375	3	ECO57	0	0.0357	0	Putative_surface-exposed_ virulence_bigA	IPR005546

Table 3.4: The result of known virulence bacterial genes from BLASTClust 0.5 S

Cluster No	Total Protein in pathogens	Number of pathogens	Virulence organism code	Total Protein in non pathogens	No of non pathogens	Common DE line	Common InterPro
2	1117	54	SALTY	959	26	Virulence_factors_putative_ positive_transcription_regulato r_bvgA	IPR011991, IPR001789 IPR001867
5	617	47	SALTY	506	25	Sensor_protein_basS/pmrB_ (EC_2.7.3.-)	IPR003661, IPR003660 IPR003594
14	271	42	VIBCH	209	22	Aldehyde_dehydrogenase_ (EC_1.2.1.3)	IPR002086, IPR012303
60	128	55	LISMO	67	25	Chaperone_clpB	IPR004176, IPR003959 IPR003593
144	76	53	LISMO	46	28	ATP-dependent_Clp_ protease_ proteolytic_ subunit_(EC_3.4.21.92)	IPR001907
176	72	51	SHIFL	39	27	Acyl_carrier_protein_(ACP)	IPR009081, IPR006163 IPR003231
201	73	39	STRR6	39	19	Zinc-binding_lipoprotein_ adcA_precursor	IPR006127, IPR006128 IPR006129
435	51	51	PASMU	27	27	ATP_synthase_gamma_chain _(EC_3.6.3.14)_(ATP_synthas	IPR000131
675	34	21	BACAN	27	15	Capsule_biosynthesis_ protein_capD	IPR000101
814	34	29	PSEAE	16	14	Phosphomannomutase/phosp hoglucomutase_ (PMM_/_PGM)	IPR005845, IPR005841 IPR005843
1116	24	22	ENTFA	11	11	Potassium-transporting_ ATPase_B_chain _(EC_3.6.3.12)	IPR008250, IPR005834 IPR001757
1236	14	8	VBCH	18	7	(ATP_phosphohydrolase_ potassium-transporting_B_chain)	
1237	31	6	YERPS	0	0	Attachment_invasion_locus_ protein_ precursor	IPR000758
2985	4	3	BORPE	7	6	Probable_tonB-dependent_ receptor_bfrD_precursor_ (Virulence)	IPR000531, IPR012910 IPR010105
5485	4	3	ECO57	2	2	Cytotoxic_protein_ccdB_ (Protein_letB)_(Protein_G) _(LynB)	IPR008337
7143	2	2	BACAN	2	2	Capsule_biosynthesis_ protein_capB	IPR008338
8282	2	2	BACAN	2	2	Capsule_biosynthesis_ protein_capC	
8629	3	3	ECO57	1	1	Protein_ccdA_(Protein_letA)_ (Protein_H)_(LynA)	IPR005546
12119	1	1	SALTY	1	1	Putative_surface-exposed_ virulenc_protein_bigA_ precursor	IPR011991 IPR011608

Table 3.5: The result of known virulence *Mycobacterium tuberculosis* genes in 1.0 S

Cluster No	Total Protein in pathogens	Number of pathogens	Virulence organism code	Total Protein in non pathogens	No of non pathogens	Common DE line	Common Inter\pro
29	57	49	ENTIFA	33	26	RNA_polymerase_sigma_ factor_rpoD_ (Sigma-A)	IPR011991, IPR007624 IPR007627
341	22	21	VIBCH	15	12	Aldehyde_dehydrogenase_ family_protein/Hypothetical_ protein	IPR002086, IPR012303
366	24	22	ENTIFA	11	11	Potassium-transporting_ATPase _B_ chain_ (EC_3.6.3.12)	IPR008250, IPR005834 IPR001757
652	16	8	MTB	8	5	DNA-binding_ response_regulator	IPR001789, IPR001867 IPR011991
1583	7	7	MTB	7	4	Mycobacterial persistence regulator MRPA	IPR001789, IPR001867 IPR005829
2441	10	5	MTB	0	0	Virulence_factor/mce-family_protein	IPR003399, IPR005693
2613	6	6	MTB	4	4	Cytotoxin_/haemolysin_ homologue	IPR002942, IPR002877 IPR004538
2818	9	5	MTB	0	0	Virulence_factor_mce_ family_protein	IPR003399, IPR005693
2820	9	5	MTB	0	0	Mce-family_protein_mce2d	IPR003399, IPR005693
2822	9	3	MTB	0	0	Phospholipase_C_1_precurso r_ EC_3.1.4.3)_(MTP40 _antigen)	IPR007312, IPR006311
2823	9	5	MTB	0	0	Virulence_factor_mce_ family_protein	IPR003399, IPR008360 IPR005693
2931	9	5	MTB	0	0	Mce-family_protein_mce1b_	IPR003399, IPR005693
3332	8	5	MTB	0	0	Virulence_factor_mce_ family_protein	IPR003399, IPR005693
7120	5	5	MTB	0	0	Heparin-binding_ hemagglutinin_(Adhesin)	IPR000897
8232	4	4	MTB	0	0	Sulfatase family protein	IPR000917
8234	4	4	MTB	0	0	Virulence_factor_mce_ family_protein	IPR003399, IPR005693
8935	4	4	MTB	0	0	Exported_repetitive_protein_p recursor_(Cell_surface_ protein_pirG)	IPR008165
11724	3	3	MTB	0	0	Virulence_factor_mce_ family_protein	IPR003399, IPR005693
12944	3	3	MTB	0	0	Virulence-regulating_ (arac/xyls_family)	IPR012287, IPR000005 IPR009057

Table 3.6: The result of known virulence *Mycobacterium tuberculosis* genes in 0.5 S

Cluster No	Total Protein in pathogens	Number of pathogens	Virulence organism code	Total protein in non pathogen	No of non pathogens	Common DE line	Common InterPro
2	1117	54	SALTY	959	26	Mycobacterial_persistence_ regulator_mrpa	IPR001789, IPR001867 IPR005829
49	124	54	MTB	95	27	RNA_polymerase_sigma_fact or_rpod_(Sigma-A)	IPR011991, IPR009042 IPR007624
887	27	27	MTB	19	19	Cytotoxin_/haemolysin_homol ogue_ (Cytotoxin/hemolysin)	IPR000943, IPR012760
1679	20	5	MTB	2	2	Mce-family_protein_mce2d	IPR002942, IPR002877 IPR004538
1681	20	5	MTB	2	2	Virulence_factor_mce_ family_protein	IPR003399, IPR005693
1699	20	5	MTB	2	2	MCE-family_protein_mce1b	IPR003399, IPR008360 IPR005693
1748	19	5	MTB	2	2	Virulence_factor_mce_family _protein_MCE4A)	IPR003399, IPR005693
1823	18	5	MTB	2	2	Virulence_factor_mce_ family_protein	IPR003399, IPR005693
1883	14	7	MTB	5	4	Phospholipase_C_1_precurso r_(EC_3.1.4.3)_ (MTP40_antigen)/	IPR003399, IPR005693
1905	17	5	MTB	2	2	Possible_MCE-family_lipoprotein_lprm	IPR007312, IPR006311
5060	6	3	MTB	0	0	Possible_virulence-regulating_38_kDa_protein	IPR003399, IPR008995 IPR005693
5995	5	5	MTB	0	0	Exported_repetitive_protein _precursor_(Cell_surface protein_pirG)	IPR012287, IPR000005 IPR009057
6392	5	5	MTB	0	0	Heparin-binding_ hemagglutinin_(Adhesin)	IPR008165

Regarding the extent of genome inclusion in the experiment, 0.5S produced close to 30% of its clusters with at least two proteins per cluster and with more than one organism. 1.0S is the least sensitive with only 20% of the clusters containing at least two organisms. In other words, this experiment produces the most exclusive, smallest clusters. There were very few clusters covering all organisms, probably because of the diversity of taxonomies chosen.

3.5.2 Detection of paralogs families

Myblastclust produced the highest number of families of proteins with more than one member from the same organism, accounting for 16% of all non singleton clusters. It is followed by 0.5S and then 1.0S with only 5.7% paralog clusters in the latter.

3.5.3 Selection of genes common to pathogens only and those unique to MTB.

We observed that close to 50% of the clusters with more than one protein created by BLASTclust with 1.0 are pathogen-specific. This is followed by Myblastclust clustering with 41% of its clusters, while 0.5 created 37% of its clusters as pathogen only clusters. The percentage ranges of *M. tuberculosis* complex only clusters created by the experiments were very close but with 1.0S still having the highest percentage (Table 3.2).

3.5.4 Production of virulence gene test set

Known virulence genes were extracted from the literature to form a test set of 38 genes from *M. tuberculosis* complex organisms and 47 genes from other pathogens. The phylogenetic distribution pattern of 47 known bacterial virulence genes from pathogens other than *M. tuberculosis* in the test set obtained from PubMed abstracts and Swiss-Prot annotation were determined. These bacterial virulence genes were distributed between 19 clusters in BLASTclust 1.0S and 21 clusters in BLASTclust 0.5S experiments (Tables 3.3 and 3.4 respectively). Of these, 5 clusters from the 1.0S experiment and 11 from 0.5S include *M. tuberculosis* complex proteins.

Three and one clusters of each experiment, respectively, contain proteins that are also annotated to be virulence genes in *M. tuberculosis* complex organisms (Tables 3.5 and 3.6). The tables show the cluster number, organisms from which the gene was characterised as being involved in virulence, and number of other pathogen and non-pathogen proteins in the cluster.

Tables 3.5 and 3.6 show that 38 of the known *M. tuberculosis* complex virulence genes tested were distributed between 19 and 13 clusters in the 1.0S and 0.5S experiments respectively. While the virulence genes from other pathogens appear to have familyi members in non-pathogens, the *M. tuberculosis* virulence genes appear to be a bit more restricted to pathogens.

3.6 Discussion

We have tested two methods of sequence clustering algorithms to detect sets of homologs from our 84 selected bacterial genomes. Each was used to create a matrix of phylogenetic profiles for all the bacterial proteins and organisms used. The output was employed to answer questions related to Mycobacterium complex virulence genes.

The results show that most of the clusters of more than one protein observed are relatively large protein families, with more than one protein from each organism and few clusters in each experiment generally contain one protein representative from each included species (Figure 3.4). For example, Figure 3.2, shows no green in colour on the heat map for 0.5S, which indicates that all the 84 organisms have at least one protein present in the fist 100 clusters of this experiment (the hetamaps show only a subset of clusters and these are different in both figures).

In general, the mean number of proteins per cluster of course decreases as the number of paralogous proteins from the different genomes decreases and as the number of clusters increases (Figure 3.5). A plausible explanation for this relationship in the experiments might be that the highly conserved paralogs have retained their original functions, while the functions of the less conserved paralogs have changed over the course of evolution.

One of the observed large paralog families of proteins is the PE and PGRS protein family of *M. tuberculosis* complex organisms, in which each of the *M. tuberculosis* complex organisms have at least 60 copies of the gene, the exception being *Mycobacterium leprae*. This is consistent with the findings of Gordon *et al.* [1998] and Cole *et al.* [1998]. They reported the distribution of some novel *M. tuberculosis* complex gene families which were either unknown before or poorly understood. They found that each *M. tuberculosis* complex organism has between 67 and 100 genes, which are annotated to be of the Pro-Glu (PE) and Pro-Pro-Glu (PPE) families.

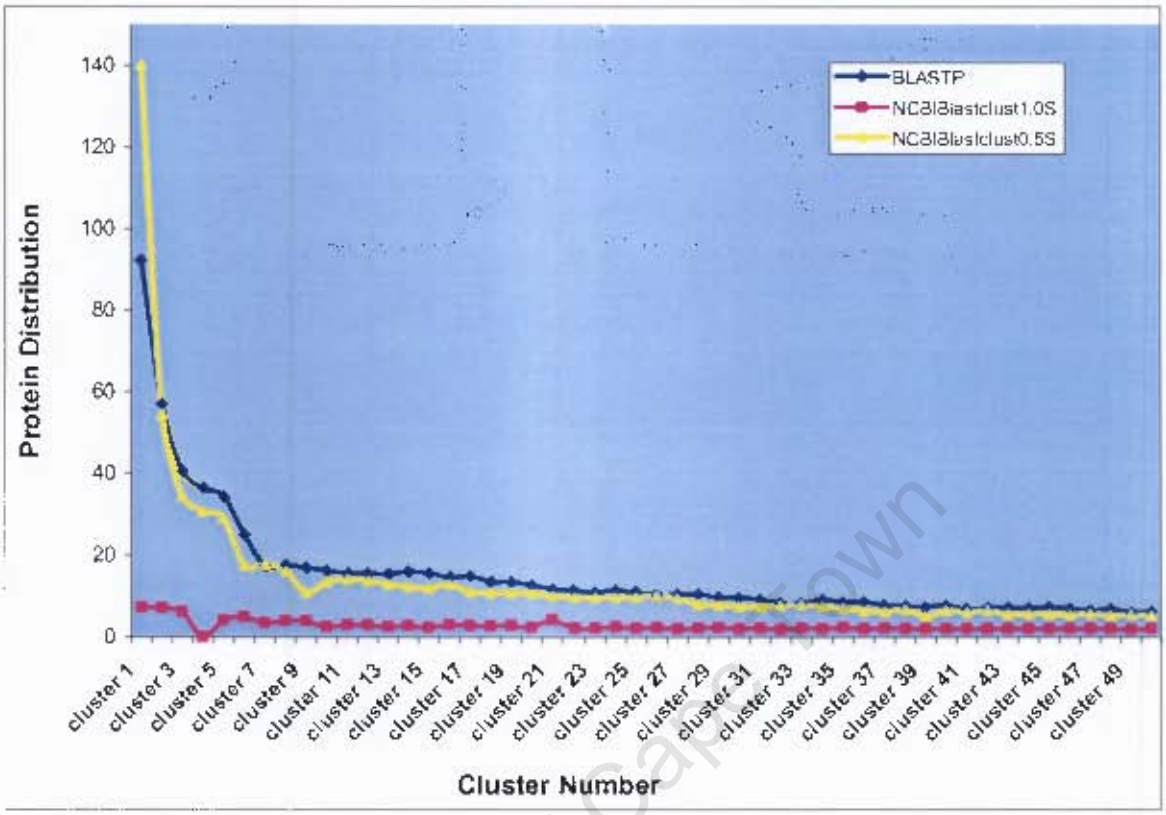

Figure 3.5: Mean Protein Distribution

Furthermore, we also noted that there are several large clusters in the NCBI BLASTclust 0.5S experiment with complex relationships between members (Figure 3.2). For instance, the adenosine triphosphatase (ATP) components of ABC transporters; histidine kinases (cluster1) and the PE and PGRS families (cluster 9) contain over 5600, 2076 and 577 proteins respectively.

The 1.0S experiment splits these families into several distinct clusters of a few hundred members each, for example, for the PE and PGRS family mentioned above, the 1.0S experiment seems to further divide them into subfamilies on the basis of their C-terminal domains. This was also observed by Poulet and Cole [1995]. However, it was found that the PE and PE/PGRS family seems to include proteins from other pathogenic and non pathogenic *actinomycetes* in the 0.5S experiment. A recent study showed that the non-mycobatcterial proteins do not contain the PE domain, and thus may not be true family members [Gey van Pittius *et al.*, 2006].

This observed pattern in the PE/PGRS and other clusters in the case of the 0.5S experiment tends to support the postulations that *Mycobacterium tuberculosis* has a set of biosynthesis genes that are common to soil organisms involved in degradation of xenobiotics and modification of organic molecule to carbohydrate. This has led to the suggestion that the Tubercle bacillus has only recently emerged as a human pathogen and still shares some of its genes with non pathogenic organisms from same lineage [Cole, 2002].

Our results also show that the occurrence of clusters of proteins that are specific to a number of pathogens are generally very few in all the clustering experiments tested compared to the numbers of pathogens included in the experiments. This is likely to be as a result of the diversity of pathogens involved in the study, and the fact that phylogenetic profiles usually follow species phylogeny rather than gene functions or organism phenotype. Allowing for some non pathogenic organisms genes to be included in the study set increases the numbers of clusters which results in a deliberate overrepresentation of the pathogens. For most of these clusters, the range of organisms included correlated with their phylogenetic lineages (Figure 3.3).

Pathogen-specific clusters that also contain *Mycobacterium tuberculosis* genes are observed to be very few and include representatives from a very small number of the total pathogenic organisms included in the study. The largest *Mycobacterium tuberculosis* containing pathogen-specific cluster observed for the 1.0S experiment contains 8 organisms viz: 5 *Mycobacterium tuberculosis* complex and 3 beta proteobacteria pathogens. The 0.5S experiment on the other hand, produced a 9 organism pathogen-specific cluster as its largest, consisting of 5 mycobacterial, 2 beta and 2 gamma proteobacterial genomes.

A close examination of the clusters that have *Mycobacterium tuberculosis* genes shows that *Mycobacterium tuberculosis* complex organisms form clusters with other *actinomycetes* (both pathogen and non pathogen) and some other phylogenies, such as high and low GC gram positive and beta and gamma proteobacteria (Figure 3.3).

Interestingly, about 90% of the clusters that contain *M. tuberculosis* genes are MTB complex-specific clusters with only the 5 MTB complex organisms included. The presence of known bacterial virulence and *M. tuberculosis* virulence genes from the

literature allowed for validation of the experimental results and enabled us to examine which of the clustering experiments identified known families of proteins correctly. To achieve this, we examined the protein description line and InterPro annotations in the known bacterial virulence gene clusters. Generally a very good correlation was observed, with majority of the member proteins having similar description lines and the same InterPro identity. The InterPro matches are more conserved in 1.0S with over 90% of the members mapping to the same InterPro families.

Surprisingly, most of these clusters containing known bacterial virulence genes are not pathogen-specific in all the experiments. The distributions of known bacterial genes in 1.0S shows that out of 19 clusters designated for these proteins only 2 are pathogen-specific while 0.5S detected only 1 pathogen-specific cluster. It should be note however that most of these test set clusters include proteins from organisms with close phylogenetic distances in the 1.0S experiment, for example cluster 8270 in Table 3.3 contains 4 non pathogenic organisms and they are all *Bacillus* species.

Examination of known virulence *M. tuberculosis* genes in all the members also revealed a similar description line and they mapped to same InterPro entries with 0.5S seeming to combine two or more clusters from 1.0s into a single cluster. Out of the 19 clusters that are designated to these known virulence *M. tuberculosis* genes in 1.0S, 11 of them were pathogen-specific and also *M. tuberculosis* complex-specific, while 0.5S had only 3 of the clusters as *M. tuberculosis* complex specific.

3.7 Conclusions

In general, the results of the experiments show that the 1.0S experiment is more sensitive and selective, generally including less paralogs from closely related genomes while 0.5S includes more paralogous proteins and also includes proteins from more distantly related organisms. The Myblastclust clustering experiment gave an intermediate result.

Because of the high level of functional similarity between orthologous proteins, the quality of orthology prediction is a central aspect in the transfer of functional annotation. To generate phylogenetic profiles of all bacterial genomes involved in this project we tested two sequence similarity-based methods and different levels of stringency to create

phylogenetic profiles of all species. We measured the functional similarity of proteins within clusters using InterPro and Gene Ontology functional data.

We as well, detected a sensitivity/selectivity trade-off: the functional similarity within a cluster of homologs increases when the number of proteins included in the groups decreases. The method or amount of stringency to be used is dependent on the research question that needs to be answered. It also depends on, for example, the evolutionary distance between the studied species and the desired size of clusters, that is many-to-many or one-to-one orthologous relationships between species.

From the results we have two sets of genes, from now on referred to as predicted virulence genes (pathogen only clusters), one from 0.5S and one from 1.0S for further study.

4

Functional Analysis

"Biology is a discipline rooted in comparisons..... Genomics is the most recent branch of biology to employ comparison-based strategies..."

-Nobrega and Pennacchio, 2004.

The second major objective of the project is to select and characterise potential virulence genes in *M. tuberculosis* by answering the following questions:

- What genes are potentially involved in virulence in *M. tuberculosis*?
- What is the potential role of these genes in virulence?

4.1 Overview

In this chapter, in analyzing the clusters generated, we use the hypothesis that functionally equivalent homologs should perform similarly in functional characterization of the protein set in each cluster. This aspect of conservation of function can be measured in various ways, e.g. having similar expression profiles and involvement in the same biological and molecular processes (GO annotation). Identical domain annotation (InterPro annotations) and conservation of protein interaction [Groenen *et al.,* 2006], could also be used to measure function conservation.

Once the clusters were generated, we were able to determine the best clustering method by calculating, for each experiment, the percentage of proteins in each cluster that had the same UniProtKB description line, same InterPro accession numbers and those involved the same GO slim biological process and molecular function. To achieve this, we examined the protein description lines, InterPro annotations and GO annotations and calculated the percentage conservation in these functional labels.

We went on to analyze the functions of predicted clusters by comparing functional categories between the predicted and non predicted sets for the 0.5S and 1.0S experiments. We determined which functions were over-represented in the predicted

sets. Since many of the predicted proteins are hypothetical proteins we tried to predict functions for some of the unknown clusters.

4.2 Background

Homologs are proteins that have a common ancestor or are evolutionarily related. It is common that very close homologs, particularly orthologs, frequently have a similar function. Homology-based transfer of functional annotations is a raw prediction technique that assigns proteins that have not been annotated with the function of their annotated homologs. These methods are based on sequence similarity between proteins.

However, functional inference through these approaches has some short comings as well as functional limitations. One major predicament is that it is difficult to be sure of the level of sequence similarity to ascertain that two proteins have the same function [Shah *et al*, 1997]. Other problems of homology-based functional transfer includes common errors when transferring annotation based on identifying one domain of multiple domain proteins as homologs [Liu *et al.* 2004]; errors in the original database annotation of the homolog; and short comings caused by evolutionary divergence, for example, when the closest homologue has lost the function or acquired another function through mutations [Yanay *et al.*, 2005].

In the absence of high sequence similarity, motifs and patterns can also be used to analyze function, by performing multiple sequence alignment of a functionally annotated protein family. This can lead to the identification of similar motifs or patterns in a target protein. It also allows for annotation transfer from experimentally characterized proteins to an unknown protein target even in the absence of a significant level of overall sequence similarity.

Protein signature databases like **Pfam, ProDom** and **PROSITE** are databases that provide essential tools for identifying distant relationships in new sequences and thus are used for the classification of protein sequences and for inference of functional similarity between protein sequences. For instance, PROSITE [Sigrist *et al.,* 2002] contains manually chosen biologically important motifs and consists generally of three types of signatures: patterns, rules and profiles, with each signature using a different automated method for searching motifs. Although the two most local signatures, patterns and rules, usually extend over few residues, profiles expand the similarity to the level of complete domains.

Another important motive-based library is Pfam [Bateman et *al.* 2004], in which motifs usually span over complete domains of at least 100 residues [Liu *et al.,* 2004]. It is based on a combination of specialist manual curation and automated analysis. The annotation in Pfam includes a description of each family and links to other resources and literature references.

INTERPRO, Integrated Resource for Protein families, domains and functional sites, is a database that catalogues information derived from protein signatures. It assembles information from various protein signatures databases that each generate protein families and domains through different protein signatures methods [Mulder *et al.*, 2003]. The InterPro database can be used to assess the conservation of molecular function within our homolog sets of proteins because each InterPro accession number represents a protein family or domain, containing a cross-species set of homologous proteins with its own functional annotation. Proteins within a homolog cluster should belong to the same InterPro family and have the same domain compositions. The higher the percentage of proteins with the same InterPro accession numbers within a homologous set, the better the conservation of function.

GENE ONTOLOGY - The Gene Ontology (GO) project arose as a result of the need to have a common annotation system for describing gene products. It provides structured, standard terms for describing the function of gene products. The GO vocabulary is divided into three ontologies viz, molecular function, biological process and cellular component. Each ontology is represented as directed acyclic graphs (DAG) where each node can have one or more parents and zero or more children. There are two types of associations, *is a*, which means the child is a subclass of the parent; and *part of* indicating the child is a component of the parent [The Gene Ontology Consortium, 2001; Gene Ontology Consortium, 2006].

Gene ontology annotations can also be used to determine which protein homologous sets are involved in the same biological process and have the same molecular functions. According to Groenen *et al.* [2006] sets of proteins were said to be active in the same process if they shared a 4th level element of the GO biological process tree, in which the root is the first level element and every succeeding branch is one level higher. GO annotations provide the available functional information of a gene product and can thus

be used as a source of validating functional similarity between gene products generated from sequence similarity methods [Calamita *et al.,* 2005]

GO SLIM – GO Slims (GS) are summarized versions of the GO vocabulary. A GS contains a division of the terms in the whole GO that give a general idea of the ontology content without the detail of the specific terms. GS mainly functions to generate a summary of the results of GO annotation of a gene when wide classification of gene products function is required [http://www.geneontology.org/GO.slims.shtml]. Since each protein from different organisms could be annotated to different levels of the GO hierarchy, we decided to use GS to more easily compare the functional annotations within a cluster. We utilized the Integr8 GOSLIM file for each genome which is a set of high-level terms selected to cover major aspects of each of the three GO ontologies without overlapping in paths of the GO hierarchy [Biswas *et al.*, 2002].

4.3 Methods

4.3.1 Functional Information

We downloaded InterPro matches for all 84 proteomes from the Integr8 Proteome Analysis Database. We also obtained complete GO Slim data for each of the selected genomes in this study [htttp://www.ebi.ac.uk/integr8].

4.3.2 Creating Database for Gene Sequences and Functional Information

A MySQL database was created to store all the results from this project. The data stored in the database is outlined below.

Protein Sequence information - 246573 protein sequences from the 84 complete genomes together with their, standardized names/description (DE) lines, accession numbers and organism codes (OS codes) were extracted, and loaded into the MySQL database.

InterPro methods table - the InterPro signature methods, matches for each domain (start and stop positions), and corresponding protein accession numbers were extracted and added to the database.

.

Gene ontology and GO slim - GO identifiers (GOid); GO ontology (one of the three - biological process (P), molecular function (F) and cellular component (C)); evidence supporting the annotations (all evidence codes were included); GO slim identifier and GO

slim name were retrieved from Integr8 GO files for each proteome and maintained in the database.

BLAST output table- the BLASTP output was parsed to generate a BLAST table in the database representing each query sequence accession number and all the resulting hit protein accession numbers, organism codes (OS code), expected values, bit scores, percentage length coverage, and percentage sequence identities across all 84 genomes.

Cluster output table – the clustering outputs from the BLASTP clustering method, and BLASTClust 0.5 and 1.0 were added to the protein sequence information table to create linkages between each protein accession number and all other data in the database to each clustering experiment.

4.3.3 Analysis of functional conservation within clustering experiments

Conservation of functions were examined by using various SQL queries to count unique InterPro ids and GO Slim ids (by ontology) in each cluster and python scripts were written to generate the percentage of these functional labels present in each cluster for each clustering experiment. The protein description lines (DE line) were also examined for conservation of terms. Since counting unique DE lines was difficult due to major differences in naming of protein sequences, this was not used further for the analysis.

4.3.4 Selection of pathogen and MTB specific clusters

For each BLASTClust experiment, all the clusters that were pathogen-specific and contained a representative from MTB, were selected as potential virulence gene clusters for further functional investigation. These are referred to as the 'predicted' clusters. Those clusters that contain MTB but were not pathogen-specific were regarded as not predicted MTB clusters.

4.3.5 Functional analysis of predicted virulence gene clusters

The protein DE lines were examined for the two sets above to investigate those that had been assigned a function and those annotated as hypothetical proteins. A set of functional categories (13 high-level function terms) was generated and a single category applied to each *M. tuberculosis* CDC1551 protein based on the DE line. The categories included: unknown, cell cycle, DNA/RNA/protein metabolism, enzyme, ESAT, foreign (transposons, phages and so on), PE/PPE/PGRS, regulator, transporter, stress

response, surface (membrane and secreted proteins) and virulence. The functional composition of both predicted and not predicted MTB clusters was determined using these functional categories. The results were summarized in charts.

In addition to the functional analysis above, a statistical analysis was done to compare the predicted and non-predicted protein sets and determine whether any GO terms were significantly over-represented in the predicted set. For this we used the Ontologizer tool [Robinson *et al.*, .2006] and the GO annotations for each MTB (CDC1551) protein, where available. Ontologizer takes as input a set of selected protein accessions (study) and the reference set (population), as well as gene association files for the organism. It performs a modification of the Fischer's Exact test and uses the Bonferroni correction to correct for multiple testing. The program takes into account parent-child relationships in GO terms, and thus considers the structure of the hierarchy, which is supposed to produce fewer false positive results.

4.3.6 Analysis of hypothetical proteins

For the hypothetical proteins in the predicted virulence set the InterPro matches were downloaded and examined to see whether we could assign a potential function based on the protein family and domain hits. We also investigated some clusters that contain hypothetical proteins to check if the predicted functions are conserved among the members of hypothetical proteins in same cluster based on their InterPro matches.

4.4 Results and Discussions

4.4.1 Functional conservation

InterPro and GO annotation data capture the available functional information of a gene product and can be used as a source for assessing functional similarity between gene products in each cluster. Figures 4.1 and 4.2 show the existence and conservation of InterPro and GO annotations of proteins that exist in each cluster in the two experiments. The percentage of proteins with InterPro matches, ranges from 75 to 100% in 0.5S and 75 to 97% in 1.0S. In total, InterPro annotations cover above 70% of all available bacteria predicted gene sequences.

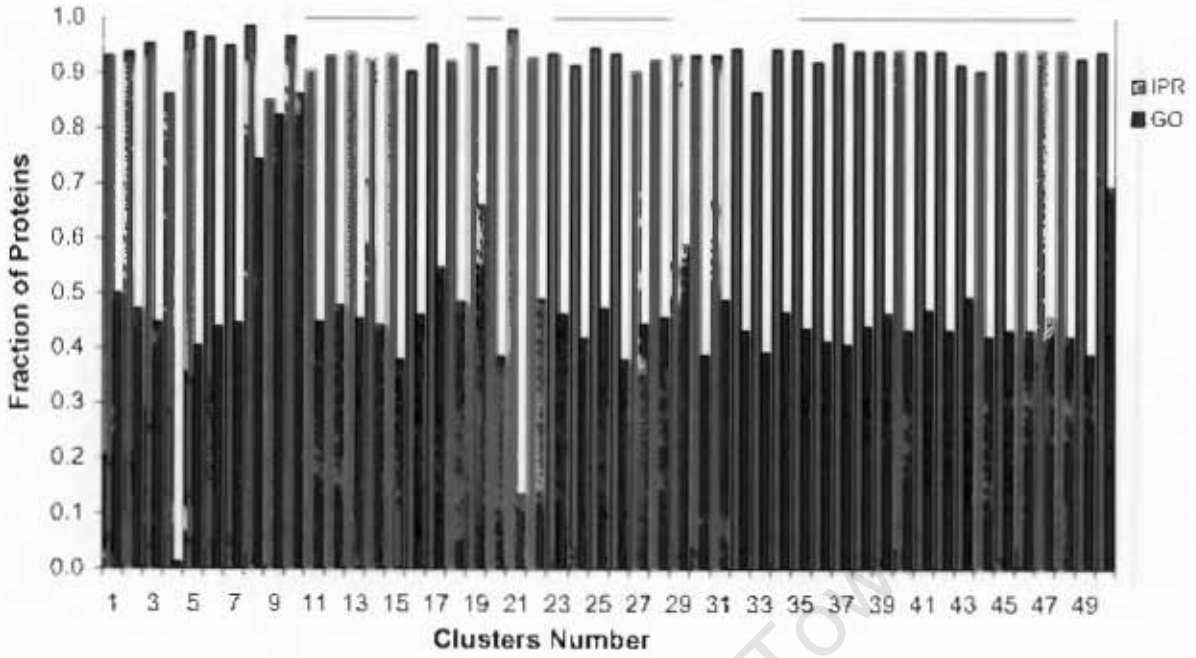

Figure 4. 1: Existence of Proteins with InterPro and GO in BlastClust 1.0

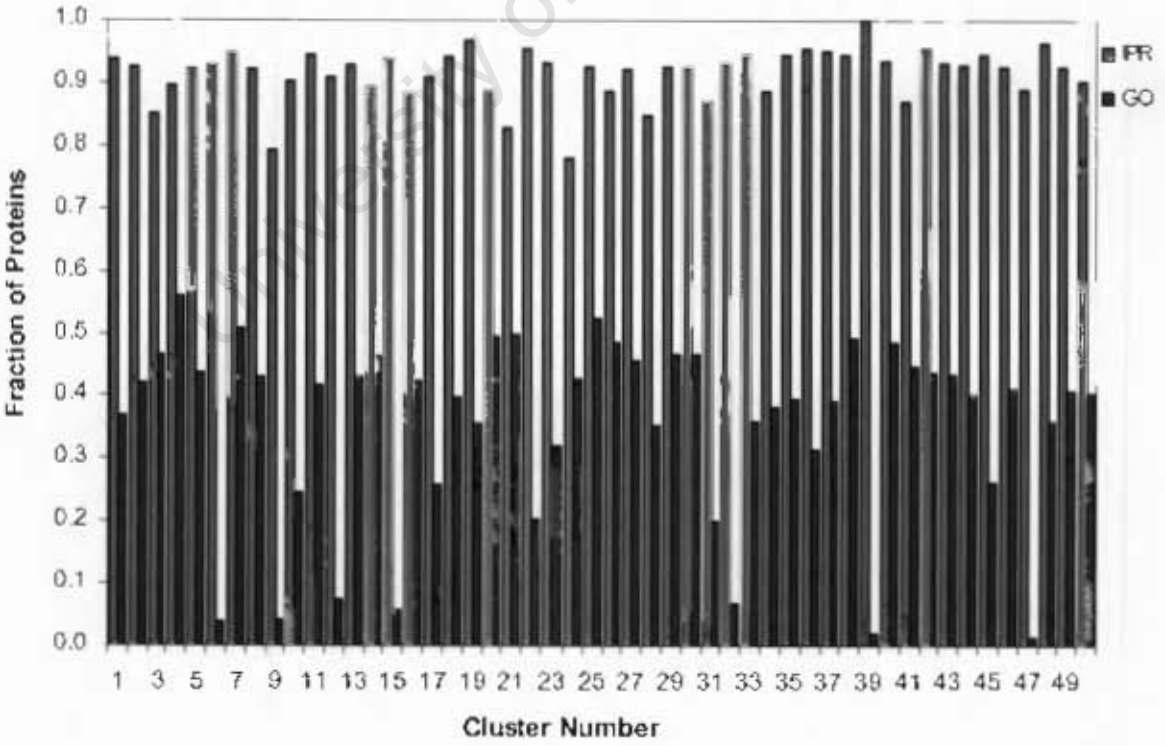

Figure 4.2: Conservation of Proteins with InterPro and GO in BlastClust 0.5

It is observed in our result that the number of proteins in each cluster that have GO term annotation is very low compared to InterPro annotations. This ranges from as low as 3 proteins in a cluster of over 200 proteins (0.013%) to a high of 50% in 0.5S and 0.011% to 80% in 1.0S experiments. This is due to the fact that a large number of gene products are not yet annotated with GO terms as was observed by Andreas *et al.* [2006].

Regarding conservation, a very good correlation was observed, with the majority of the member proteins having similar description lines and the same InterPro identity. At least 60% and 92% of protein members of every cluster in 0.5S and 1.0S respectively are annotated to same InterPro and GO slim term, where available. However the annotations are more conserved in 1.0S with over 90% of the members mapping to the same InterPro families.

4.4.2 Selection of pathogen and MTB specific clusters

Experimentally we predicted about 8000 and 18000 pathogen specific clusters for 0.5 and 1.0S respectively. Out of these figures, 1015 and 2355 clusters, respectively, contain potential virulence MTB genes due to their being predicted as pathogen-specific clusters with MTB gene members. These numbers are portions of the pathogen specific clusters that have MTB complex organisms generally

The whole *M. tuberculosis* CDC1551 proteome (the reference virulence MTB complex organism) contains over 4000 proteins. Of the 4172 proteins (that is, the number of proteins of the reference organism predicted as member of the above number clusters), 1167 in 0.5S and 2614 in 1.0S were members of predicted pathogen-specific clusters and the rest are members of non pathogen-specific clusters. From our result it can be noted that 1.0S, which is the most stringent experiment, predicted more potential virulence MTB genes with close to 65% of the whole *M. tuberculosis* CDC1551 proteome being a member of predicted virulence gene clusters (that is, either unique to MTB complex organisms or MTB and other pathogens).

The 0.5S experiment, on the other hand, only predicted about 29% of the whole CDC1551 proteome to be members of potential virulence gene clusters. This is due to the fact that some of these families of genes that are predicted in the 1.0S experiment as virulence gene clusters are included in non-predicted families of genes in 0.5S, due to these clusters containing some proteins from non pathogenic phylogenetic lineages in

the latter data set. For example, the PE/PGRS family includes 2 non pathogenic *actinomycetes* in the 0.5S experiment.

When looking more closely at some of the clusters from the two experiments, the 1.0S results better reflect what has previously been reported. For example, the PE/PPE family has been reported to be unique to MTB organisms [Gey van Pittius *et al.*, 2006], and 0.5S identifies some cluster members in other organisms and is probably therefore not strict enough. Gey van Pittius *et al.* report that some similarity has been found to proteins in other closely related organisms, but this is due to non-specific alignment of repeated regions, and these other proteins do not contain the conserved PE/PPE domains and motifs. The 1.0S results are still, however, quite surprising in the high number of proteins not found in non-pathogenic organisms. There are relatively few proteins conserved across all pathogens and not found in non-pathogens, and the results also indicate a high proportion of MTB complex-specific proteins.

4.4.3 Functional Analysis of the *M. tuberculosis* proteome in predicted and not predicted pathogen specific clusters

Figures 4.3 and 4.4 compare the functional composition of both predicted and not predicted potential virulence MTB gene sets for each experiment, BLASTClust0.5 and 1.0.

In the 0.5S results, the cases where the number of proteins in the predicted set exceeds that in the not-predicted set are the "Antigens,"ESAT" and "surface" categories, the latter of which includes membrane and secreted proteins. This is probably a result of the fact that the not-predicted set is bigger for all other categories. For the 1.0S results, the bigger set now includes all the PE/PPE proteins, and a significant proportion of the unknown, surface, antigen and virulence proteins, which may all play a role in virulence or the intracellular lifestyle of the organism as predicted compared to non predicted sets.

Of particular interest are the proteins that are characterized as unknown or hypothetical proteins, which occupy the highest percentage in both predicted and not predicted gene sets in both of the experiments. For 0.5S, 63% of the total predicted virulence genes are unknown (Figure 4.5a) and 35% of the total not predicted genes are unknown (Figure 4.5b). The percentage of unknown proteins in the 1.0S predicted and not predicted sets are close to 40%, but the predicted set still produced highest percentage (Figures 4.6a and 4.6b).

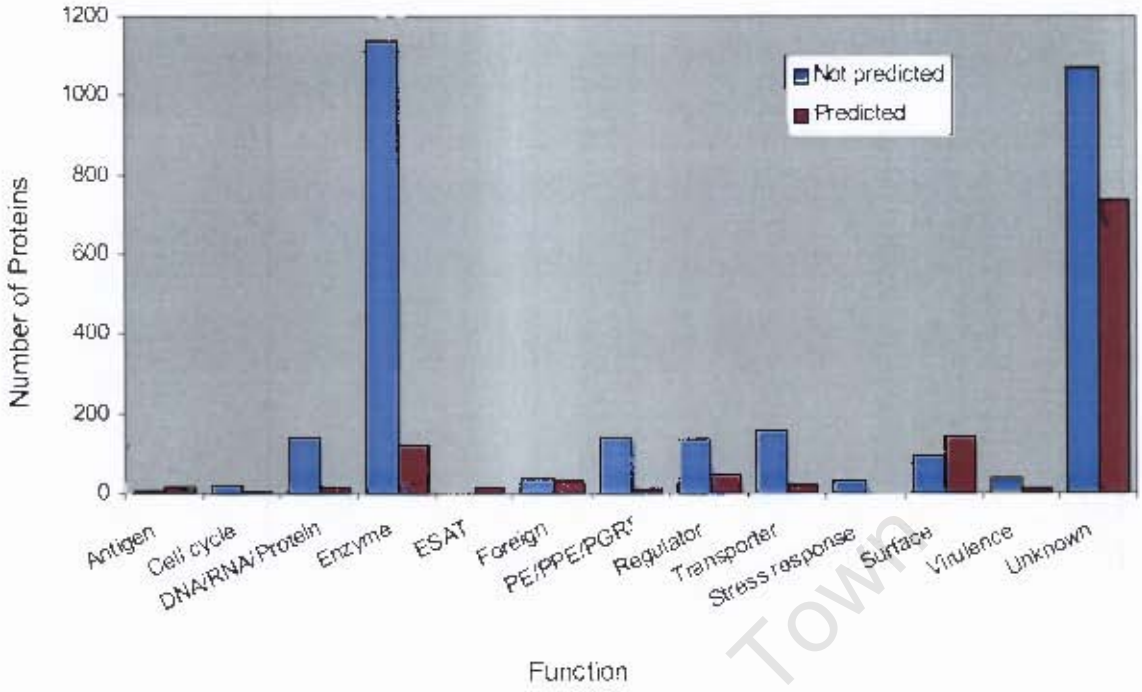

Figure 4.3: Summary of functions in the predicted and not-predicted sets from the 0.5S experiment.

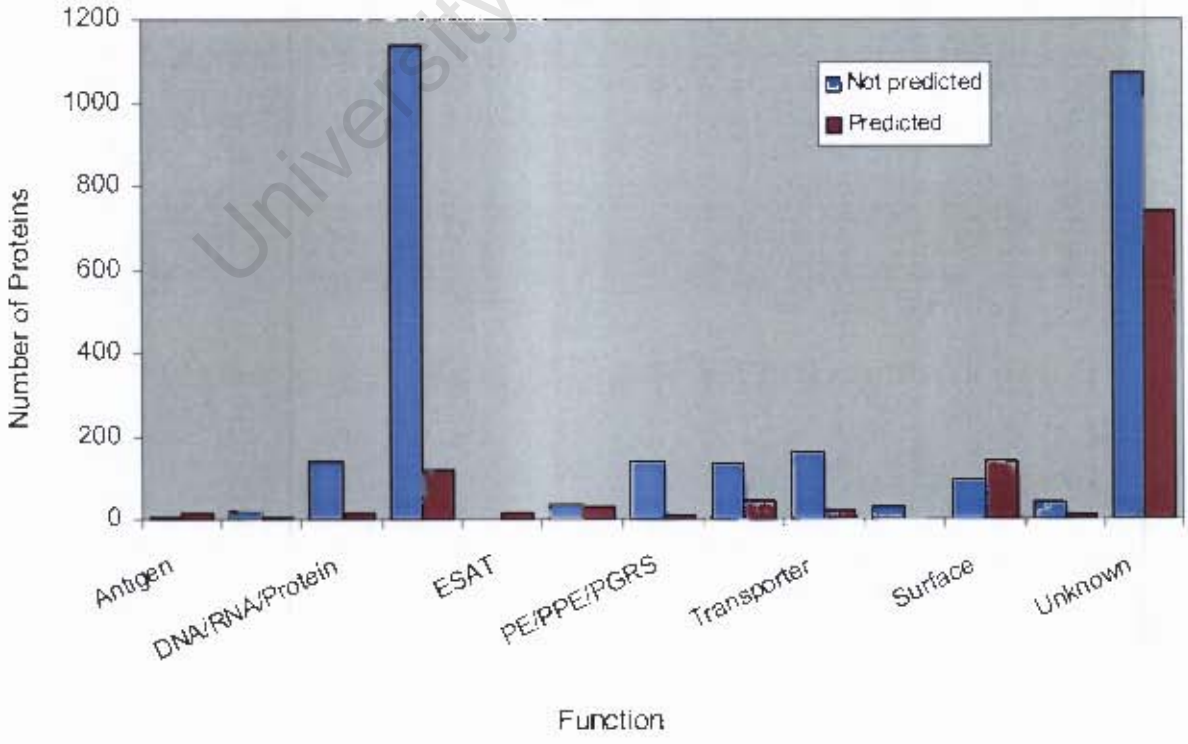

Figure 4.4: Summary of functions in the predicted and not-predicted sets for 1.0S

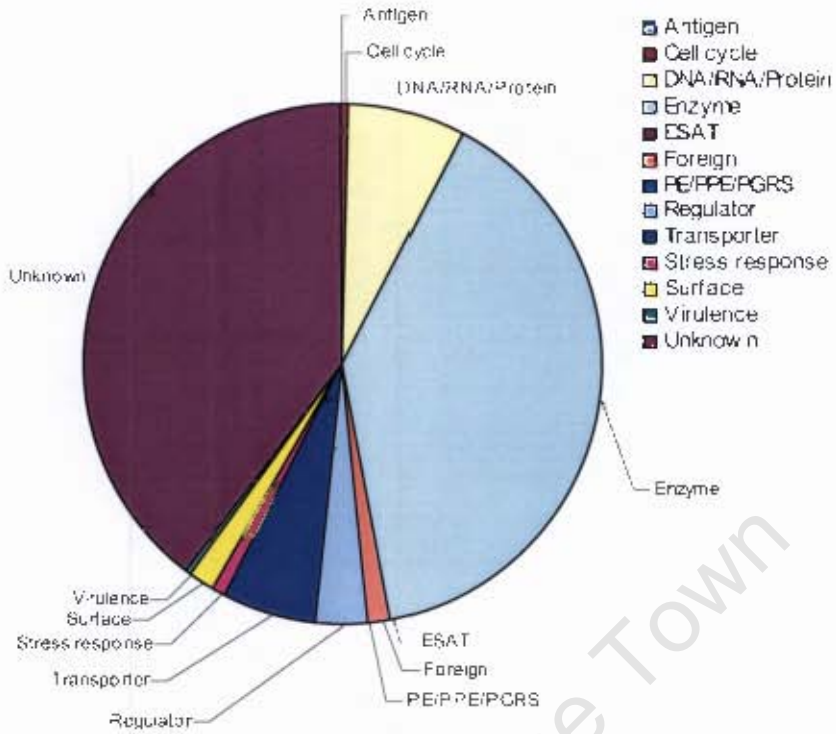

Figure 4.5a: Summary of Functions in the Not Predicted Proteins in the 0.5S experiment

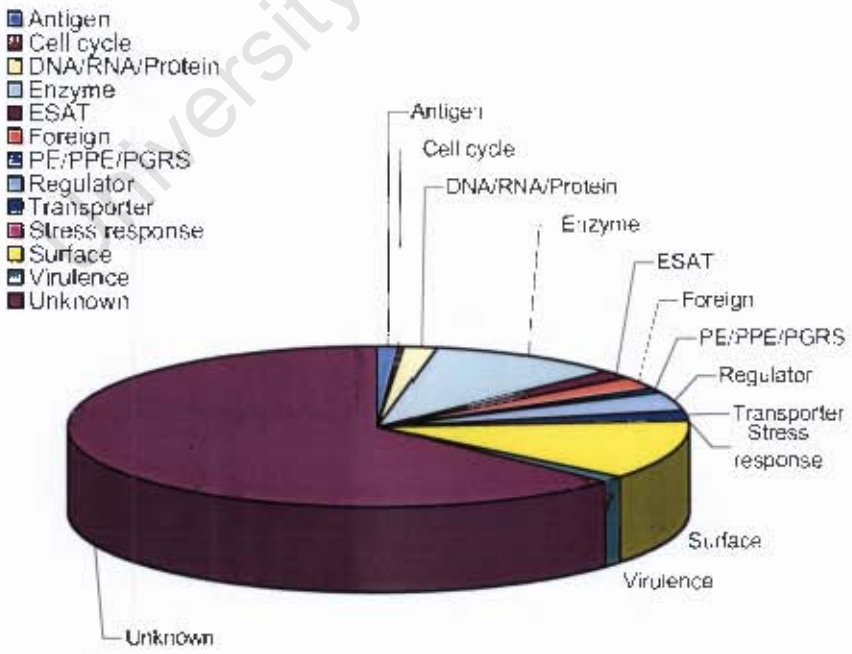

Figure 4.5b: Summary of Functions in the Predicted Proteins in 0.5S

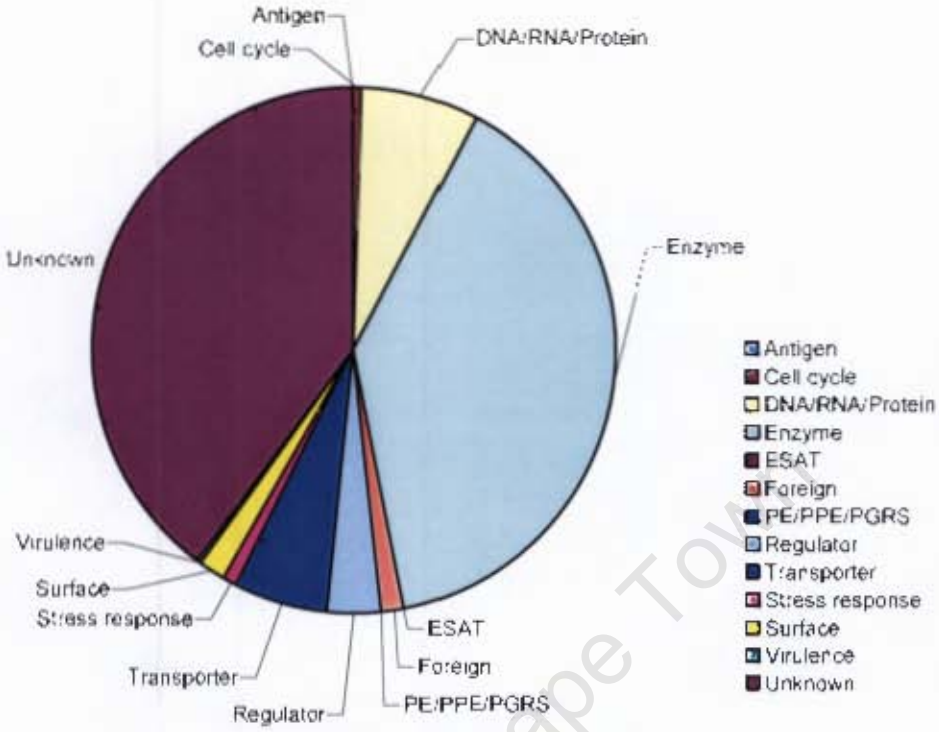

Figure 4.6a: Summary of Functions in the Predicted Proteins in 1.0S

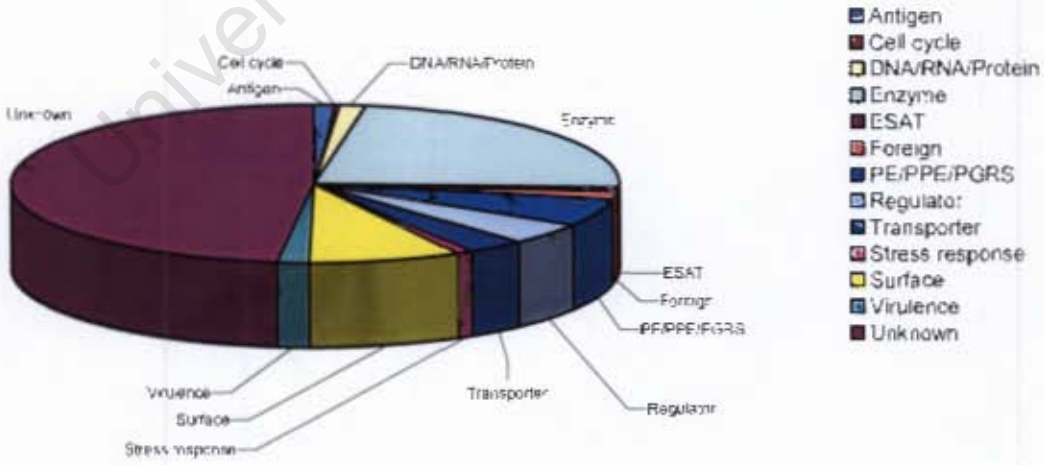

Figure 4.6b: Summary of Functions in the Not Predicted Proteins in 1.0S

Also apparent is the distribution of enzymes and the PE/PPE families of protein in the experiments. For the predicted virulence gene set, the enzyme set only occupied about 10% of all proteins in 0.5S and 23% in 1.0S, while the proportion of proteins that are annotated as enzymes are similar in both experiments for the not predicted set. Unlike the other functional categories, the percentage of PE/PPE proteins in the predicted virulence gene set is almost 6 times bigger in 1.0S experiment compared to 0.5S.

On the other hand, the 0.5S experiment produced a larger proportion of PE/PPE proteins in the not predicted sets. This is a result of the 0.5S experiment being more lenient than 1.0S for inclusion of members. It was previously observed that most members of these families form a cluster with some related non pathogenic *actenomyces* organisms as discussed above.

In contrast to our expectations, the fraction of the proteome that is functionally annotated as being involved in virulence is very low, only 1 and 1.4% of the predicted virulence set in 0.5 and 1.0s respectively. However, this reflects the current low level of knowledge about which genes are involved in virulence in this organism. The number of characterized virulence genes in the predicted set is significantly higher than the not predicted set for the 1.0S experiment.

We also examined the functions that are over-represented in MTB proteins that have been predicted as potentially virulent using GO annotations and the Ontologizer software [Robinson *et al.*, 2006]. Using Ontologizer, we performed a statistical analysis to determine which GO terms were overrepresented in the predicted sets to see whether certain functions were unique to this organism and pathogens, and further confirm whether those gene sets are involved in pathogenesis and virulence of the organism. However it must be noted that this analysis was carried out for only part of the data set that has GO annotation.

Figure 4.7 shows a summary of GO terms that are significantly overrepresented in the predicted virulence set compared to all proteins in the genome for the 1.0S experiment. The terms that are over-represented are shaded in three different colours based on the three GO ontologies. The terms shown are those with a p-value of less than E-5. Other terms that are positioned between the significant terms and the root term are shown with no shading. These GO terms, as well as additional GO terms with a p value of less than E-4 are shown in Table 4.1.

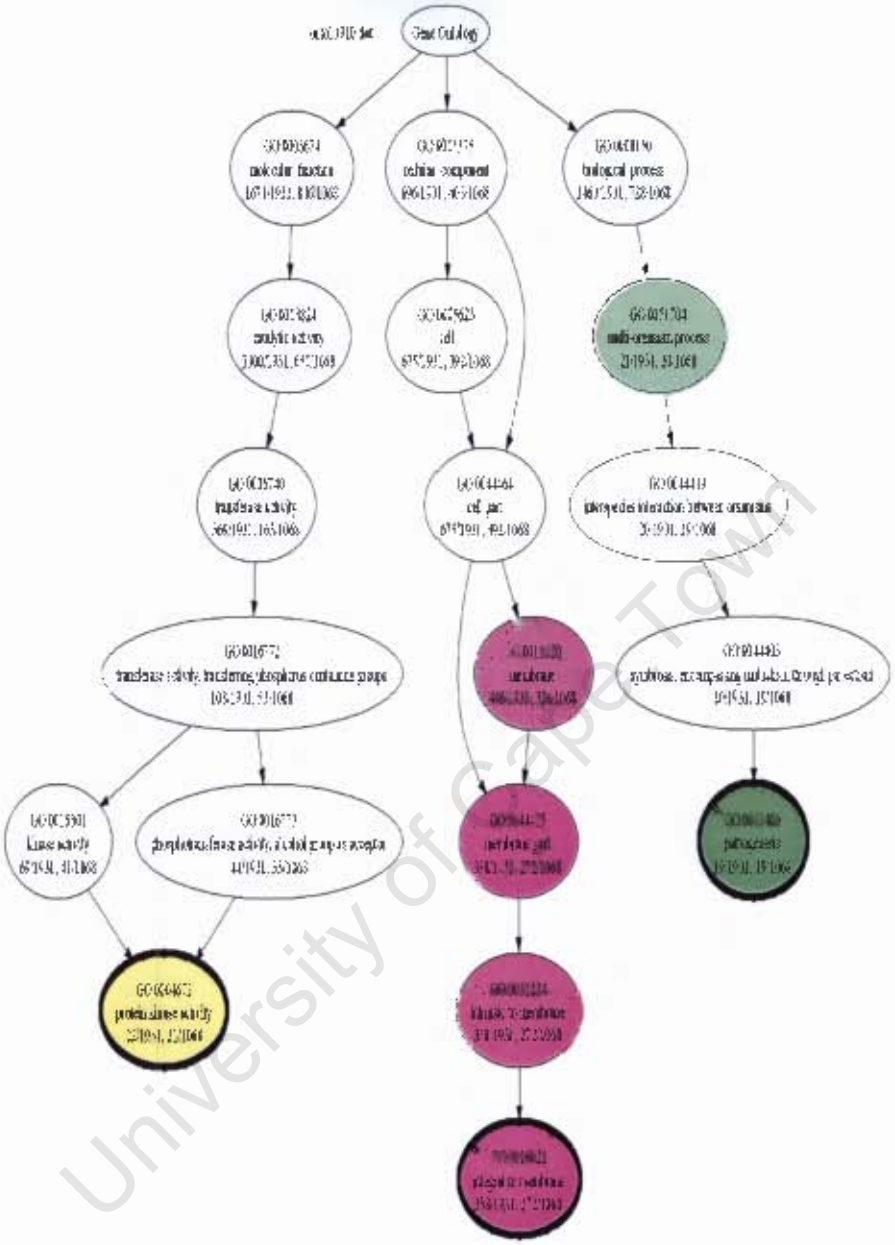

Figure 4.7: Summary of the significantly overrepresented GO terms in the predicted set for 1.0S

The result of our GO analysis revealed that a number of terms related to the membrane and membrane components were significantly overrepresented in the cellular component ontology, and *protein kinase activity* was prominent among the molecular function annotations of the predicted set. Interestingly, *pathogenesis* and *multi-organism process* were also the most significant biological processes shown to be overrepresented in the predicted virulence gene set.

Table 4.1: Significantly (<E -4) over-represented GO terms in predicted set from the 1.0S experiment.

GO ID	GO name	P value
GO:0006468	protein amino acid phosphorylation	1.246 E-4
GO:0043687	post-translational protein modification	8.124 E-4
GO:0009187	cyclic nucleotide metabolic process	7.969 E-4
GO:0051704	multi-organism process	6.642 E-5
GO:0007242	intracellular signaling cascade	4.385 E-4
GO:0044419	interspecies interaction between organisms	1.157 E-4
GO:0009405	pathogenesis	1.207 E-5
GO:0044403	symbiosis, encompassing mutualism through parasitism	1.156 E-4
GO:0004497	monooxygenase activity	4.710 E-4
GO:0004672	protein kinase activity	3.8083 E-5
GO:0003700	transcription factor activity	5.837 E-4
GO:0044425	membrane part	3.459 E-19
GO:0016021	integral to membrane	3.459 E-19
GO:0016020	Membrane	1.022 E-17
GO:0031224	intrinsic to membrane	3.459 E-19

The 0.5S results had fewer overrepresented GO terms in the predicted set, and these included terms such as *membrane, integral to membrane, regulation of transcription* and *DNA binding.*

It is difficult to draw too many conclusions from the data because of all the proteins that do not have GO annotations. However, the appearance of the above GO terms in the summary of significantly overrepresented GO annotations results suggested that many of the proteins in the predicted set are membrane proteins that are involved in pathogenesis activities of the organism and its interaction with other organisms that is, host-pathogen interactions.

4.4.4 Analysis of hypothetical proteins

Approximately 40% of the MTB genome is made up of hypothetical proteins or proteins of unknown function. Since many of the predicted protein sets are hypotheticals, we tried to predict functions for some of them. Table 4.2 shows some of the proteins in the 1.0S experiment predicted set that have been classified as unknown, together with their InterPro matches. We observed that some of these groups can be assigned potential functions based on InterPro matches, which could serve as the basis for further investigation of some interesting clusters. Most of these proteins appear to be enzymes, but some are involved in transcription regulation, and some *in prevent-host-death* and *abortive infection*, which could be involved in the pathogenesis processes of the organism. To compare InterPro matches of some of the interesting clusters of hypothetical proteins we selected 3 clusters from the 1.0S experiment,,based on the fact that all the hypothetical members of these clusters hit the same InterPro domains and some have predicted structures that can be modelled. The above characteristics make them good examples of hypothetical gene clusters that can be further investigated and removed from the list of hypothetical proteins.

Figure 4.8a shows the protein matches for the first cluster tested, that contained 11 hypothetical proteins. In this cluster all the members hit an O-methyltransferase domain, and we also observed that members of this cluster have a predicted structure that can be modeled base on a known protein databank (PDB) structure, 1rjd. The PDB structure is for a carboxy methyl transferase protein that is involved in regulating protein phosphatase 2a activity.

It is observed that the next cluster seems to contain a nuclease domain that may be involved in transposition as shown in Figure 4.8b, while the last cluster include proteins that are *guanylyl cyclases* with a *histidine kinase* domain. This domain is predicted to be involved in signaling and signal transduction (Figure 4.8c).

Table 4.2: Proteins in 1.0S predicted set that are previously classified as unknown and their InterPro matches.

Protein Accession	InterPro ID	InterPro Name
O05294	IPR008262	Lipase, active site
O05305	IPR003593	AAA+ ATPase, core
O05305	IPR011990	Tetratricopeptide-like helical
O05309	IPR001173	Glycosyl transferase, family 2
O05460	IPR000641	CbxX/CfqX
O05573	IPR013974	SAF domain
O05592	IPR013781	Glycoside hydrolase, catalytic core
O05770	IPR002641	Patatin
O05815	IPR009078	Ferritin/ribonucleotide reductase-like
O05854	IPR011047	Quinonprotein alcohol dehydrogenase-like
O05856	IPR002197	Helix-turn-helix, Fis-type
O05866	IPR009187	Predicted Ku, prokaryotic type
O05882	IPR006037	TrkA-C
O05918	IPR000846	Dihydrodipicolinate reductase
O06178	IPR003736	Phenylacetic acid degradation-related protein
O06178	IPR006683	Thioesterase superfamily
O06218	IPR003779	Carboxymuconolactone decarboxylase
O06218	IPR004675	Alkylhydroperoxidase AhpD core
O06232	IPR000836	Phosphoribosyltransferase
O06233	IPR007712	Plasmid stabilization system
O06242	IPR000403	Phosphatidylinositol 3- and 4-kinase, catalytic
O06242	IPR000601	PKD
O06250	IPR000631	Carbohydrate kinase
O06288	IPR002502	N-acetylmuramoyl-L-alanine amidase, family 2
O06328	IPR011251	Bacterial luciferase-like
O06351	IPR005149	Transcriptional regulator PadR-like
O06351	IPR011991	Winged helix repressor DNA-binding
O06412	IPR004360	Glyoxalase/bleomycin resistance /dioxygenase
O06415	IPR002716	PilT protein, N-terminal
O06547	IPR013217	Methyltransferase type 12
O06572	IPR001054	Adenylyl cyclase class-3/4/guanylyl cyclase
O06580	IPR003615	HNH nuclease
O06580	IPR013324	Sigma factor, regions 3 and 4
O06592	IPR004378	*Mycobacterium tuberculosis* paralogous family 11
O06619	IPR009097	Appr>p cyclic nucleotide phosphodiesterase
O06630	IPR003675	Abortive infection protein
O06630	IPR000015	Fimbrial biogenesis outer membrane usher protein
O06632	IPR000182	GCN5-related N-acetyltransferase
O06780	IPR003477	PemK-like protein
O06780	IPR011067	Plasmid maintenance toxin/Cell growth inhibitor
O06800	IPR003779	Carboxymuconolactone decarboxylase
O06800	IPR004675	Alkylhydroperoxidase AhpD core
O07187	IPR000802	Arsenical pump membrane protein
O07187	IPR004680	Citrate transporter
O07205	IPR013813	YjgF/chorismate mutase-like
O07205	IPR006175	Endoribonuclease L-PSP
O07238	IPR001854	Ribosomal protein L29
O07239	IPR001023	Heat shock protein Hsp70
O07251	IPR013216	Methyltransferase type 11
O07429	IPR006025	Peptidase M, neutral zinc metallopeptidases
O07733	IPR002589	Appr-1-p processing
O07738	IPR004136	2-nitropropane dioxygenase, NPD
O07742	IPR007372	YceI
O07743	IPR003455	O-methyltransferase, N-terminal
O07751	IPR006992	Amidohydrolase 2
O07754	IPR012349	Pyridoxamine 5-phosphate oxidase, FMN-binding
O07764	IPR004027	SEC-C motif
O07764	IPR011990	Tetratricopeptide-like helical
O07770	IPR011660	Rv0623-like transcription factor

Table 4.2 contd.

Protein Accession	InterPro ID	InterPro Name
O07782	IPR006442	Prevent-host-death protein
O07810	IPR000150	Cof protein
O07810	IPR006379	HAD-superfamily hydrolase, subfamily IIB
O07810	IPR005834	Haloacid dehalogenase-like hydrolase
O33195	IPR000577	Carbohydrate kinase, FGGY
O33239	IPR011101	Phage Gp37Gp68
O33266	IPR002711	HNH endonuclease
O33302	IPR011660	Rv0623-like transcription factor
O50380	IPR012951	Berberine/berberine-like
O50389	IPR000253	Forkhead-associated
O50389	IPR008984	SMAD/FHA
Q8VKS1	IPR000477	RNA-directed DNA polymerase (Reverse transcriptase)
Q8VKQ8	IPR005031	Streptomyces cyclase/dehydrase
Q7D9C8	IPR010093	Excisionase/Xis, DNA-binding
Q7D9D6	IPR003455	O-methyltransferase, N-terminal
Q7D9H0	IPR003675	Abortive infection protein
Q7D964	IPR005467	Histidine kinase
Q7D967	IPR011701	Major facilitator superfamily MFS_1
Q7D980	IPR001647	Bacterial regulatory protein, TetR
Q7D980	IPR009057	Homeodomain-like
Q7D9H0	IPR003675	Abortive infection protein
Q7D9X9	IPR010419	Carbon monoxide dehydrogenase subunit G
Q7D9T2	IPR001104	3-oxo-5-alpha-steroid 4-dehydrogenase, C-terminal
Q7D9L2	IPR000318	Nitrogenase component 1 alpha and beta subunits
Q7D9M2	IPR002925	Dienelactone hydrolase
Q7D8W2	IPR001087	Lipolytic enzyme, G-D-S-L
Q7D8W2	IPR013831	Esterase, SGNH hydrolase-type, subgroup
Q7D8P0	IPR013228	PE-PPE, C-terminal
Q7D7Q8	IPR011067	Plasmid maintenance toxin/Cell growth inhibitor
Q6ARF7	IPR008975	Viral coat and capsid protein
Q50593	IPR000644	Cystathionine-beta-synthase
Q11037	IPR012338	Penicillin-binding protein, transpeptidase fold
Q11034	IPR013656	PAS fold-4
Q11034	IPR010822	Sporulation stage II, protein E C-terminal
Q10880	IPR001750	NADH/Ubiquinone/plastoquinone (complex I)
P96936	IPR004147	ABC-1
Q10384	IPR006055	Exonuclease
P96916	IPR006442	Prevent-host-death protein
P95041	IPR003785	Creatininase
P64751	IPR002218	Glucose-inhibited division protein A

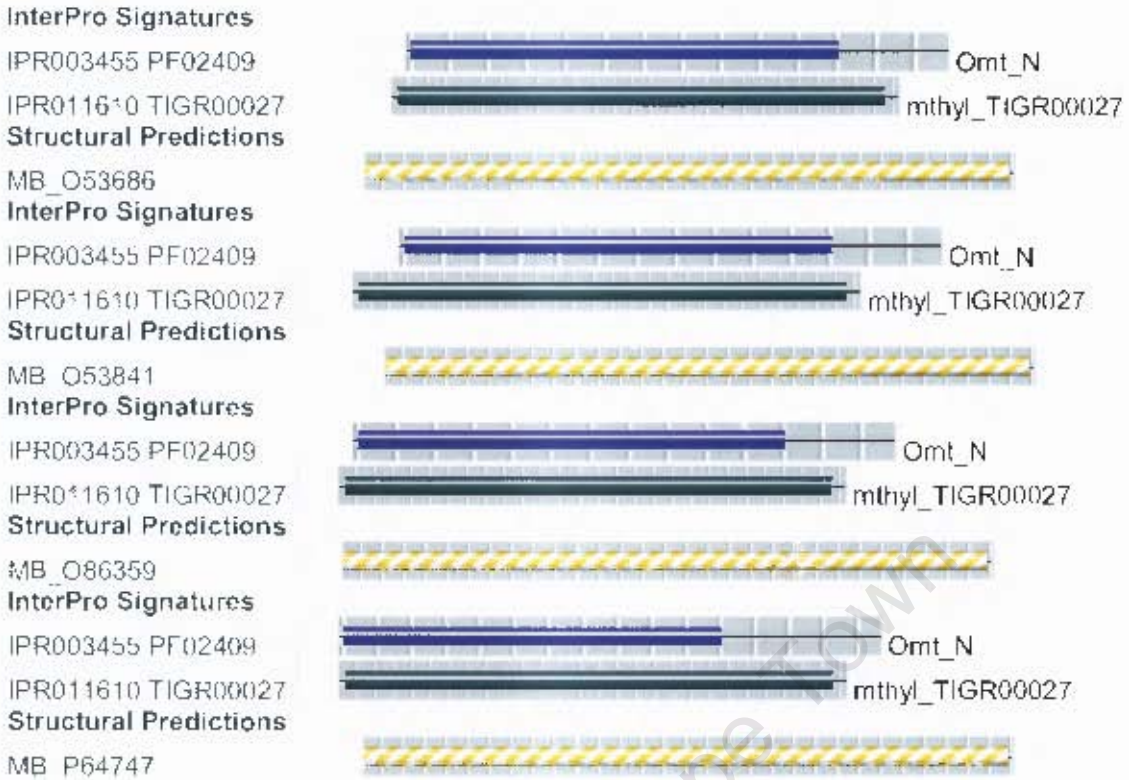

Figure 4.8a: InterPro matches for a hypothetical protein cluster showing the presence of an O-methyltransferase domain and a modeled structure.

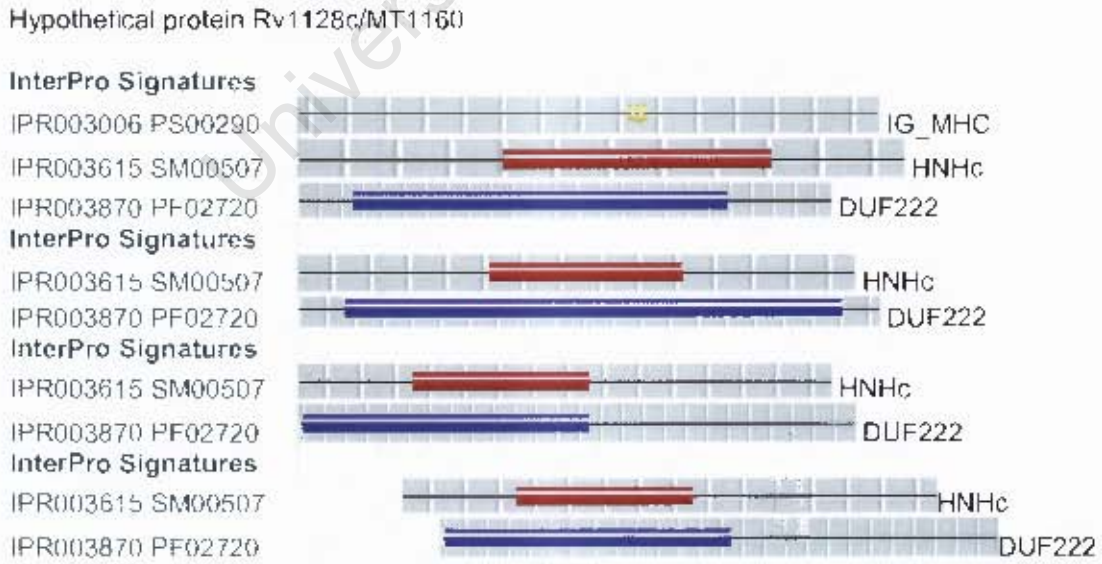

Figure 4.8b: InterPro matches for a hypothetical protein cluster showing the presence of a nuclease domain.

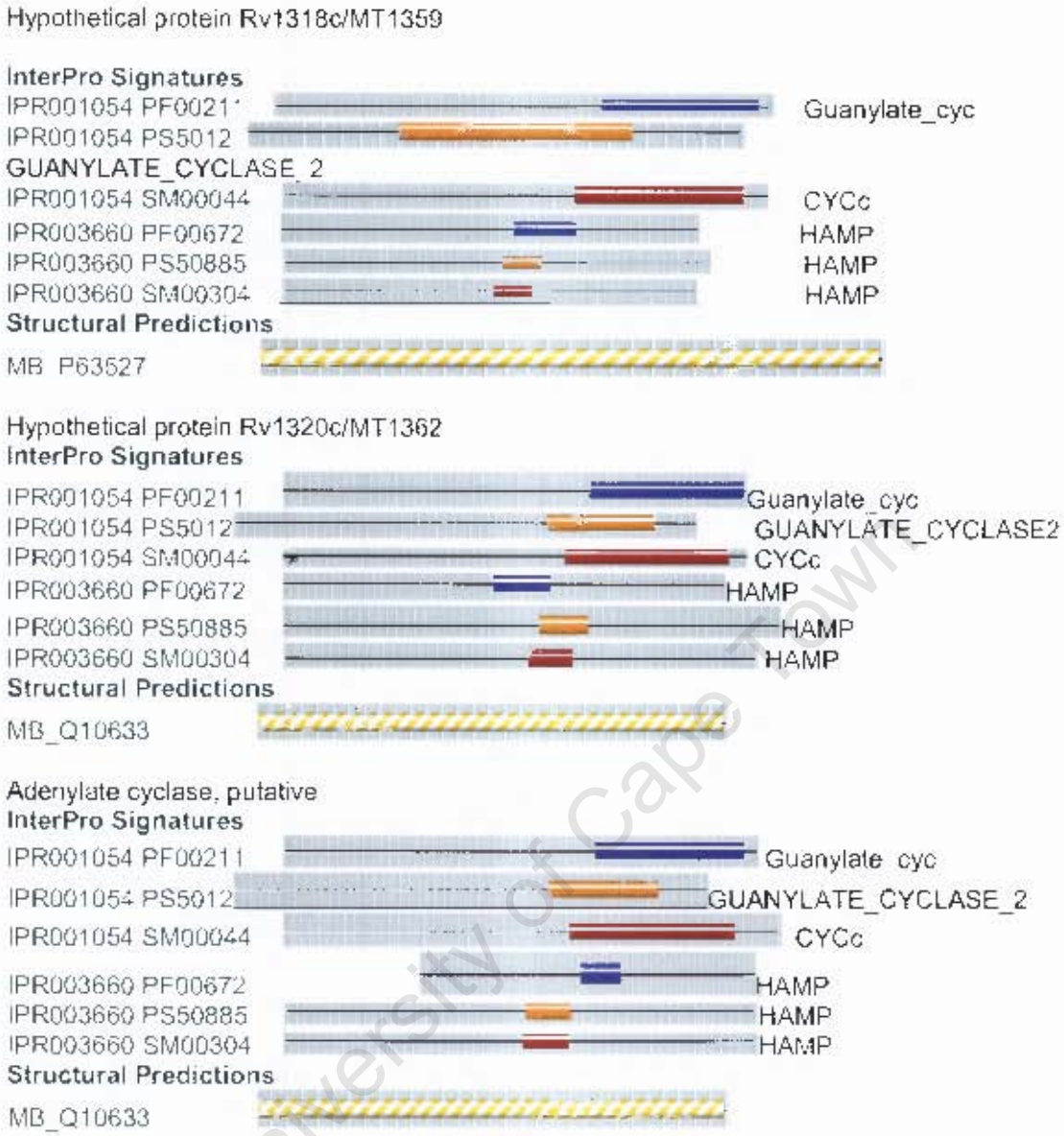

Figure 4.8c: InterPro matches for a hypothetical protein cluster showing the presence of guanylyl cyclase family signatures and modeled structures.

4.5 Conclusions

In summary, from the results of this chapter we observed a very good correlation of functional information within member proteins of each cluster, with the majority of the same cluster proteins having similar description lines and identical InterPro and GO annotations, where available. However, it should be noted that conclusions based on GO annotations should be considered with caution, due to the nature of the annotations, especially the use of electronic (IEA) evidence codes which are generally based on

sequence similarity. The InterPro and GO annotations were more conserved in 1.0S with over 90% of the members mapping to the same InterPro families and GO annotations. The functional analysis and statistical results of GO term annotations for the predicted set suggested that many of these proteins are involved in pathogenesis of *Mycobacterium tuberculosis* (the reference genome), with some being part of the PE/PPE family, known virulence proteins, or membrane-associated or secreted proteins. While some of the predicted proteins, that is, those unique to MTB or to pathogens, may not be related to pathogenesis, the concentration of potential virulence-related genes in the predicted set suggests that there may be many more genes in this set that could play a role in virulence in MTB. We need to look further at some of the hypothetical proteins, particularly, to determine whether they have a role in the pathogenic lifestyle of the organism.

Unsurprisingly, most of the experimentally predicted potential virulence gene sets are hypothetical proteins. This is expected, since annotation is done based on homology with other proteins in the database, and those that have no hits in other organisms can only initially be annotated as hypothetical. Some may have hits in other organisms (for instance, other pathogens), but the hits may have been to hypothetical proteins. Based on InterPro matches we are able to predict potential functions for some of these hypothetical proteins.

5

Comparative Evolutionary Analysis

While the pathogenesis and epidemiologies of tuberculosis are well studies, relatively little is known about the evolution of the infectious agent *Mycobacterium tuberculosis,* especially at the within-host level...

-Tanaka, 2004

This chapter aims to understand the evolutionary diversity within selected homolog sets including those that contain candidate virulence genes of *Mycobacterium tuberculosis*. It also investigates the evolutionary processes or mechanisms producing such changes. It addresses the objectives by answering the following questions:

- What is the rate of non-synonymous to synonymous substitutions within the selected homologue sets?
- What is the phylogenetic distance between sequences of the selected sets?
- What is the measure of selective pressure acting on the members of the sets?
- How are the above estimated parameters likely to affect functions and structures of the members of the set?

5.1 Overview

In this chapter we investigate some predicted clusters of interest with characterized proteins to determine how these genes evolved. This was achieved via analysis of the evolutionary changes to determine functional and structural constraints and distance between the sequences within the clusters. Some of the clusters investigated were predicted in chapter three to have cellular roles in virulence in pathogens.

We also investigated duplicate copies of proteins in mycobacterial species in these clusters to determine whether expansion of these families of genes is linked with the increase in virulence or evolution of pathogenesis.

5.2 Background

Sequence homology is used to refer to nucleotide or protein sequences that share a common ancestor [Hillis, 1994; de Pinna, 1991]. Homologous sequences can be classified into orthologous and paralogous sequences. Orthologous sequences are formed from a speciation event [Hills, 1994; Patterson, 1988] and become very useful for inferring evolutionary relationships among genes for phylogenetic studies [Filliol *et al.,* 2006]. Paralogous sequences on the other hand occur as a result of a gene duplication event [Fitch, 1970; Groenen, 1993; Fitch, 2000; Patterson, 1988]. Paralogous genes are used to study the history of gene duplication [Hillis, 1994].

The evolutionary relationships between DNA or protein sequences from organisms within and between species can be inferred from the amount of genetic variation found in the different genomes as a result of different types of mutations [Frothingham, 1995]. Sequence polymorphisms and gene duplications are an important source of genetic variation between organisms [Dufayard *et al.,* 2005]. They are also important in acquisition of genes that are essential for adaptability of organisms to their environment [Alland *et al.,* 2007]. For instance, some duplication and mutation processes provide pathogens with a selective advantage to invade the host immune system [Sekiguchi, 2007; Musser, 1995].

After speciation or duplication events, homologous sequences diverge from each other as different mutations occur in the individual sequences [Zmasek and Eddy, 2001]. Depending on the rate of mutation and the selection pressure acting on homologous sequences, certain types of mutations can result in the creation of new functions [Jothi *et al.,* 2006]. Mutations are alterations that occur in DNA sequences due to the change in and deletion of bases; insertions; inversions and substitutions [den Dunnen and Antonarakis, 1999] and can involve a single DNA base pair or a large segment of a chromosome [Cardoso *et al.,* 2004]. Substitutions involve the change of one nucleotide to another [Karboul *et al.,* 2006], that is, a point mutation or single nucleotide polymorphism, and can be divided into transversions and transitions.

A *transversion* is the mutation from a purine to a pyrimidine or vice versa, while a *transition* is the substitution from a pyrimidine to a pyrimidine or purine to a purine [Jothi *et al.,* 2006]. A substitution that leads to no amino acid change is called a synonymous substitution or silent mutation and therefore the protein remains unchanged, whereas an

amino acid replacement substitution is referred to as a non-synonymous substitution and often results in a change in the function of the protein [Karboul *et al.,* 2006; Nei and Gorjobori, 1986].

In order to infer equivalence in gene functionality or maintain gene function within a species, the number of non-synonymous substitutions (*dN*) should therefore remain small. In most cases the number of synonymous substitutions (*dS*) does not affect the functionality of the gene [Kimural, 1983]. The ratio dN/dS (γ, gamma) provide a measure of the selective force or constraints on the genes or genomes being studied. A gamma value less than 1 (γ<1) indicates purifying selection with a value of 1 (γ=1) implying neutral evolution. A gamma value greater than one (γ>1) means diversifying selection. When gamma is significantly greater than one, positive selection can thus be inferred [Yang *et al.*, 2002]. This indicates less functional or structural constraints in the population or set of genes being tested. For example orthologous or paralogous sets of genes with a gamma ratio greater than one may likely not share the same function. On the other hand, since purifying selection is a sign of high functional or structural maintenance or conservation within the group or population, with a gamma value of less than one there is a high possibility of homologous proteins having the same function.

5.3 Material and Methods

Amino acid coding DNA sequences for each protein accession number in different experimentally predicted virulence clusters were downloaded from the Sequence Retrieval System (SRS) searching the EMBL database at http://srs.ebi.ac.uk/srsbin/ .

Pair-wise and multiple DNA sequence alignments were performed using the ClustalW package of BioEdit. For each aligned set of sequences the following analyses were performed:

- Pair-wise, within group synonymous and nonsynonymous single nucleotide substitutions calculated, and DNA polymorphism and divergence between phylogenetic lineages determined using the DNA Sequence Polymorphism package, DNASp 4.10 [Rozas *et al.,* 2003].
- Investigated nature of mutations between members of each cluster by conducting a neutrality test using equations proposed by McDonald and Kreitman [1991] with the DNASp 4.10 package.

- Estimated evolutionary distance and proportion of amino acid differences between the sequences; drew a Bootstrap neighbor-joining tree using the neighbor-joining algorithm with the Molecular Evolutionary genetics Analysis (MEGA3.1) package [Kumar *et al.*, 2004].

5.4 Results and Discussion

5.4.1 Multiple sequence alignments

Pair-wise comparative sequence analysis using ClustalW shows that the amino acid sequences within each cluster are highly similar to each other (Figures 5.1, 5.2 and 5.3). Identical amino acid residues are shown in the alignments with black shading. There are many sites (prolog sites) with a high density of identical amino acid residues (more than 10 residues and a few variable bases followed by other highly identical residues). These conserved terminals or domains are likely to be important for function or conservation of structural folds for genes that are within the same family.

5.4.2 Nucleotide diversity and polymorphic divergence

The pattern of synonymous and non synonymous substitutions can provide information about mutations and selective forces on genes as well as information about the population and recombination events [Fleishmann *et al,* 2002]. We investigated gene members of two clusters that contain MTB complex organisms as well as other phylogenetic lineages, and one MTB complex-specific cluster. This was to check whether there was any evolutionary constraint that might be acting on the synonymous and non synonymous substitutions that might affect the function or structure of members of the cluster.

The number of synonymous nucleotide substitutions per synonymous site, dS, and the number of non-synonymous nucleotide substitutions per non-synonymous site, dN, for each cluster were investigated according to Nei and Gojobori [1986]. Tables 5.1a to 5.1c show the pair-wise nucleotide diversity and dS, dN while Tables 5.2a to 5.2c reveal the within species or lineage nucleotide diversity.

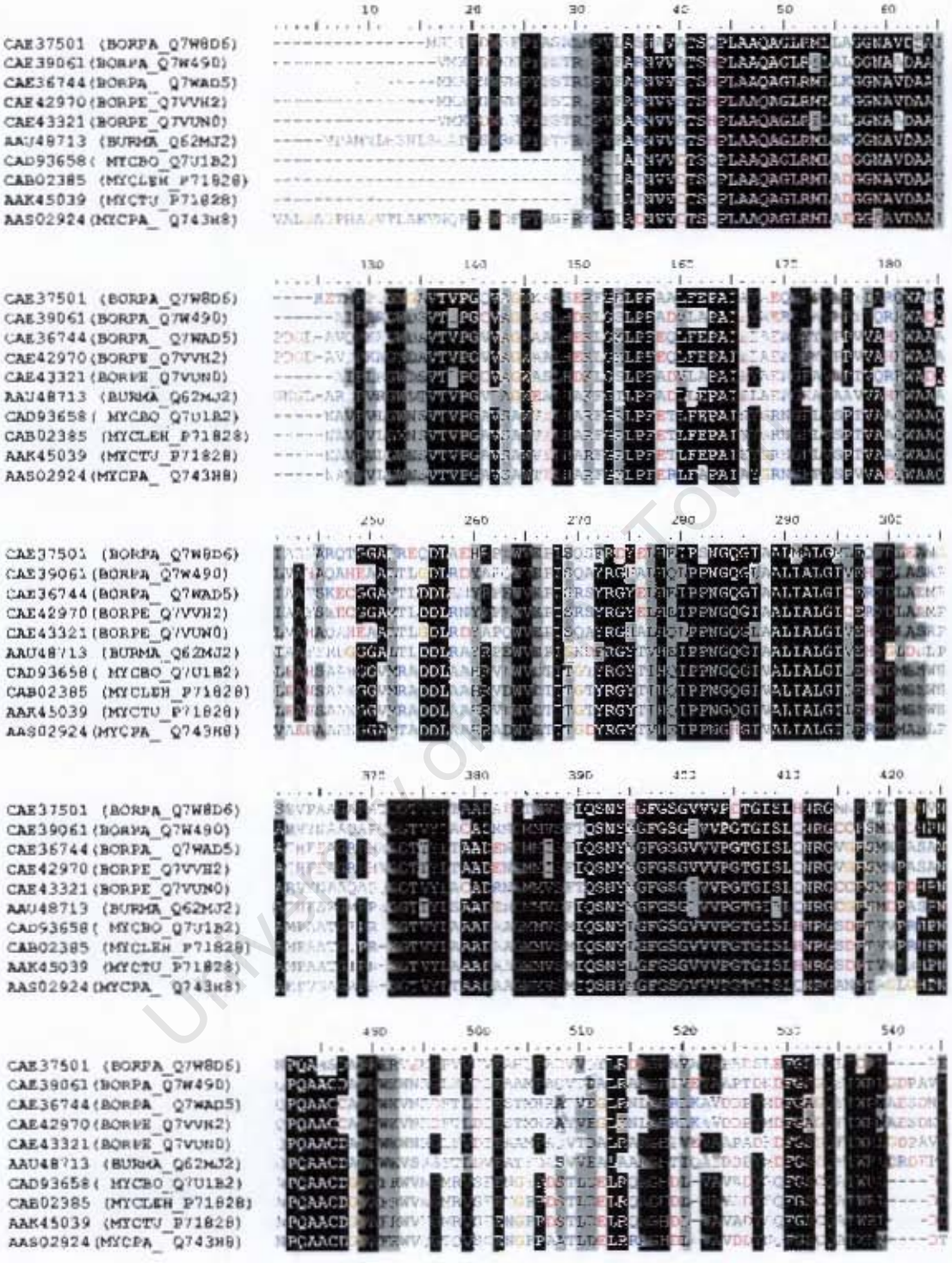

Figure 5.1: Multiple amino acid sequence alignment for the bi-functional Gamma-glutamyltransferase (GGTA) protein cluster.

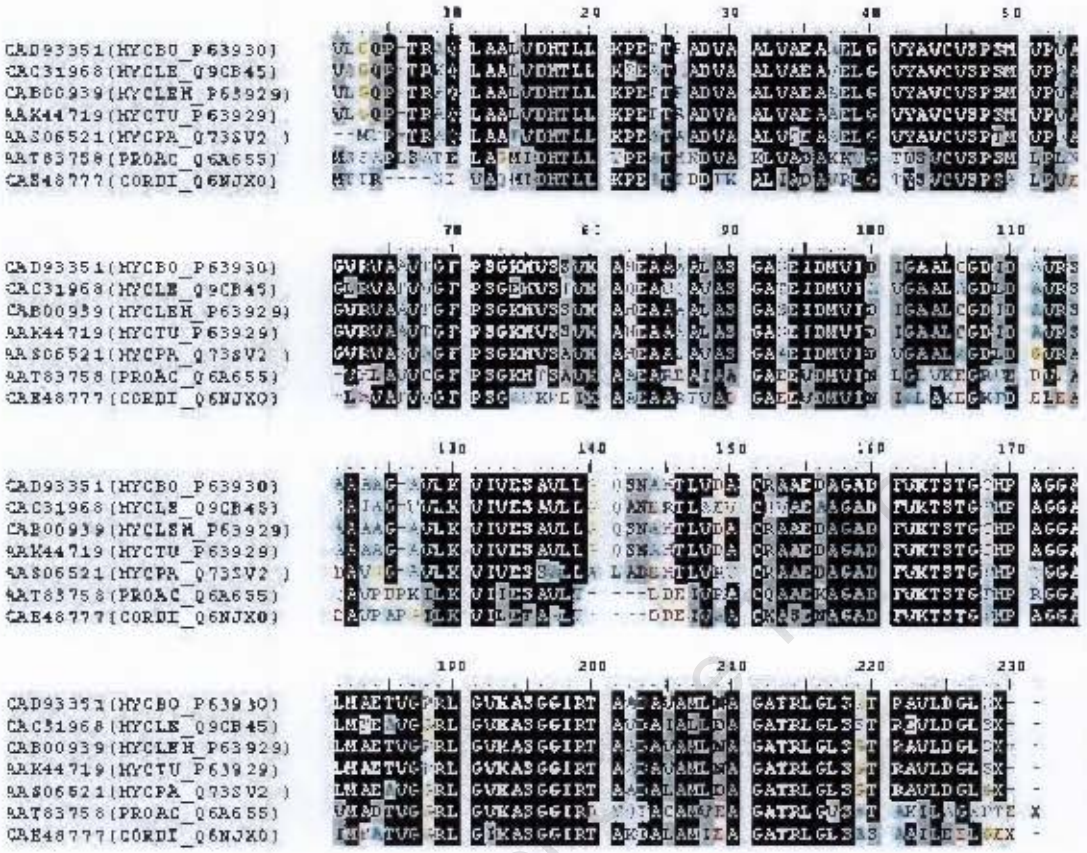

Figure 5.2: Multiple amino acid sequence alignment for the Deoxyribose-phosphate aldolase protein cluster.

The dN/dS ratios for all the families tested were less than 1 (Table 5.1a to 5.1c), which suggests that evolution is under constraint for all the gene members tested. However, the ratio observed among other phylogenetic lineages is always higher than the ratio observed for the MTB copy of the genes. This shows that in each case, the number of non-synonymous substitutions for the MTB complex copy of the gene is lower than that of other phylogenetic lineages tested (Tables 5.1 and 5.2).

For example, it is evident from the nucleotide diversity result that the number of synonymous substitutions per synonymous site (dS) is high for the beta-proteobacteria (0.6) and high-GC (1.9) organisms compared to the MTB complex genes (Table 5.2a, 5.2b), and also compared to non-synonymous replacements.

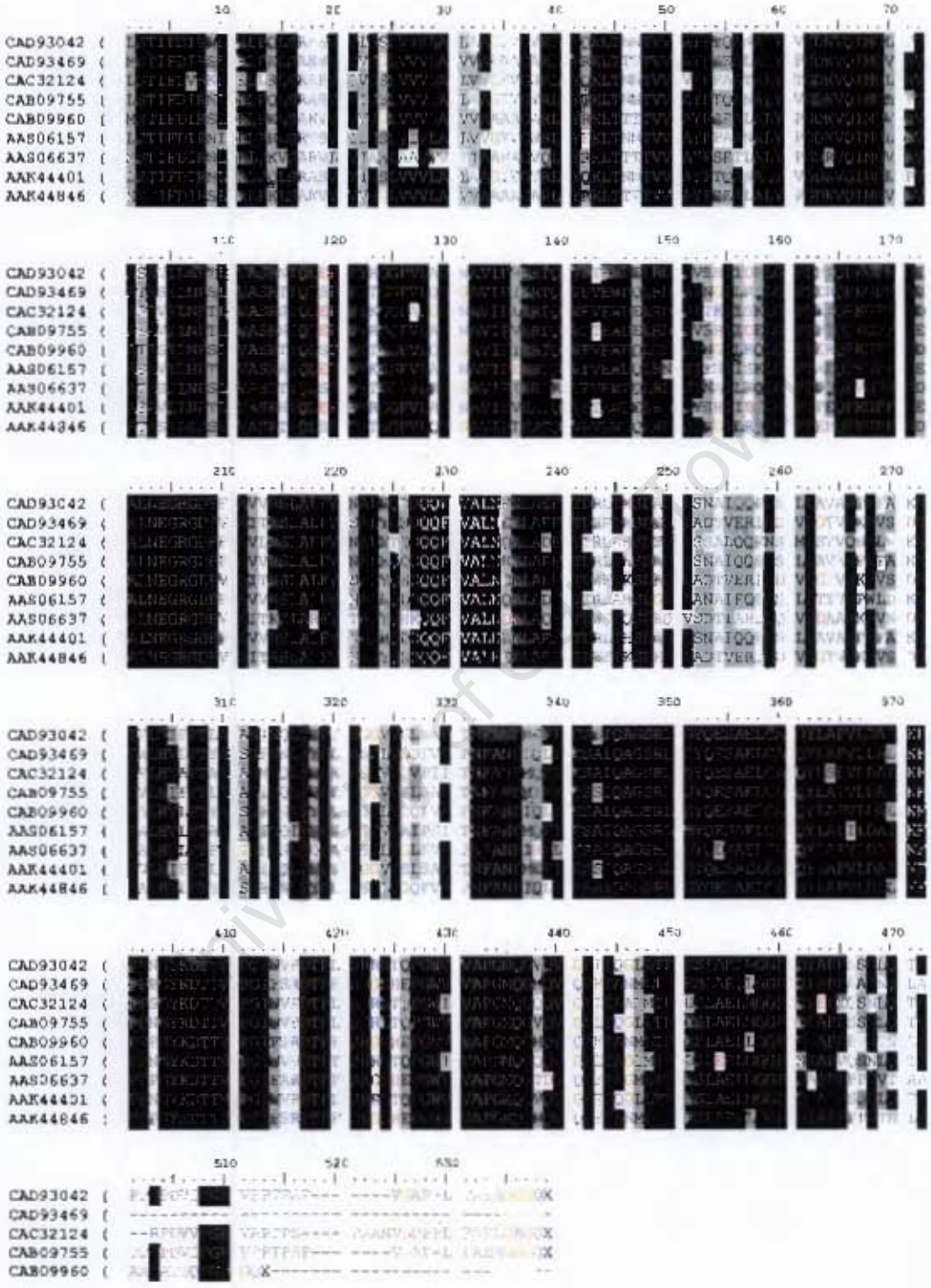

Figure 5.3: Multiple amino acid sequence alignment for the MTB complex virulence factor MCE protein cluster.

Table 5.1a: Pair-wise comparison of synonymous and non-synonymous substitutions within the bi-functional Gamma glutamyltransferase (GGTA) protein cluster.

Seq 1	Seq 2	SynDif	NSynDif	SynPos	NSynPos	*dS*	*dN*	*dN/dS*
CAE39061(BORPA)	CAE36744(BORPA)	177.25	286.75	392	1141	0.6927	0.3061	0.4419
CAE39061(BORPA)	CAE42970(BORPE)	177.75	287.25	392.42	1140.58	0.6947	0.3069	0.4418
CAE39061(BORPA)	CAE43321(BORPE)	2	2	400	1133	0.005	0.0018	0.3600
CAE39061(BORPA)	AAU48713(BURMA)	204.83	281.17	396.58	1136.42	0.8752	0.3002	0.3430
CAE39061(BORPA)	CAD93658(MYCBO)	236.75	348.25	397	1136	1.189	0.3941	0.3315
CAE39061(BORPA)	CAB02385(MYCLEH)	236.75	348.25	397	1136	1.189	0.3941	0.3315
CAE39061(BORPA)	AAK45039(MYCTU)	236.75	348.25	397	1136	1.189	0.3941	0.3315
CAE39061(BORPA)	CAE37501(BORPA)	170.5	364.5	395.67	1137.33	0.641	0.4181	0.6523
CAE39061(BORPA)	AAS02924(MYCPA)	226.58	347.42	398.83	1134.17	1.0625	0.3937	0.3705
CAE36744(BORPA)	CAE42970(BORPE)	11	3	384.75	1148.25	0.0291	0.0026	0.0893
CAE36744(BORPA)	CAE43321(BORPE)	176.25	287.75	392.33	1140.67	0.6853	0.3075	0.4487
CAE36744(BORPA)	AAU48713(BURMA)	168.25	194.75	388.92	1144.08	0.645	0.1931	0.2994
CAE36744(BORPA)	CAD93658(MYCBO)	212.08	355.92	389.33	1143.67	0.9718	0.4020	0.4137
CAE36744(BORPA)	CAB02385(MYCLEH)	212.08	355.92	389.33	1143.67	0.9718	0.4020	0.4137
CAE36744(BORPA)	AAK45039(MYCTU)	212.08	355.92	389.33	1143.67	0.9718	0.402-	0.4137
CAE36744(BORPA)	CAE37501(BORPA)	151.25	348.75	388	1145.00	0.5501	0.3908	0.7104
CAE36744(BORPA)	AAS02924(MYCPA)	182.83	338.17	391.17	1141.83	0.732	0.3767	0.5146
CAE42970(BORPE)	CAE43321(BORPE)	176.75	288.25	392.75	1140.25	0.6873	0.3083	0.4486
CAE42970(BORPE)	AAU48713(BURMA)	169.08	195.92	389.33	1143.67	0.6489	0.1945	0.2997
CAE42970(BORPE)	CAD93658(MYCBO)	214.92	356.08	389.75	1143.25	0.9967	0.4025	0.4038
CAE42970(BORPE)	CAB02385(MYCLEH)	214.92	356.08	389.75	1143.25	0.9967	0.4025	0.4038
CAE42970(BORPE)	AAK45039(MYCTU)	214.92	356.08	389.75	1143.25	0.9967	0.4025	0.4038
CAE42970(BORPE)	CAE37501(BORPA)	150.25	349.75	388.42	1144.58	0.5439	0.3925	0.7216
CAE42970(BORPE)	AAS02924(MYCPA)	183.33	338.67	391.58	1141.42	0.7341	0.3777	0.5145
CAE43321(BORPE)	AAU48713(BURMA)	203.83	282.17	396.92	1136.08	0.8657	0.3017	0.3485
CAE43321(BORPE)	CAD93658(MYCBO)	236.42	348.58	397.33	1135.67	1.1825	0.3948	0.3339
CAE43321(BORPE)	CAB02385(MYCLEH)	236.42	348.58	397.33	1135.67	1.1825	0.3948	0.3339
CAE43321(BORPE)	AAK45039(MYCTU)	236.42	348.58	397.33	1135.67	1.1825	0.3948	0.3339
CAE43321(BORPE)	CAE37501(BORPA)	169.5	364.50	396	1137	0.6342	0.4182	0.6594
CAE43321(BORPE)	AAS02924(MYCPA)	226.25	347.75	399.17	1133.83	1.0571	0.3944	0.3731
AAU48713(BURMA)	CAD93658(MYCBO)	226.17	353.83	393.92	1139.08	1.0878	0.4010	0.3686
AAU48713(BURMA)	CAB02385(MYCLEH)	226.17	353.83	393.92	1139.08	1.0878	0.4010	0.3686
AAU48713(BURMA)	AAK45039(MYCTU)	226.17	353.83	393.92	1139.08	1.0878	0.4010	0.3686
AAU48713(BURMA)	CAE37501(BORPA)	184.67	357.33	392.58	1140.42	0.74	0.4057	0.5482
AAU48713(BURMA)	AAS02924(MYCPA)	196.67	345.33	395.75	1137.25	0.8149	0.3892	0.4776
CAD93658(MYCBO)	CAB02385(MYCLEH)	0	0	394.33	1138.67	0	0	n.a
CAD93658(MYCBO)	AAK45039(MYCTU)	0	0	394.33	1138.67	0	0	n.a
CAD93658(MYCBO)	CAE37501(BORPA)	220	338	393	1140	1.029	0.3773	0.3667
CAD93658(MYCBO)	AAS02924(MYCPA)	202.67	132.33	396.17	1136.83	0.8595	0.1265	0.1472
CAB02385(MYCLEH)	AAK45039(MYCTU)	0	0	394.33	1138.67	0	0	n.a
CAB02385(MYCLEH)	CAE37501(BORPA)	220	338	393	1140	1.029	0.3773	0.3667
CAB02385(MYCLEH)	AAS02924(MYCPA)	202.67	132.33	396.17	1136.83	0.8595	0.1265	0.1472
AAK45039(MYCTU)	CAE37501(BORPA)	220	338	393	1140	1.029	0.3773	0.3667
AAK45039(MYCTU)	AAS02924(MYCPA)	202.67	132.33	396.17	1136.83	0.8595	0.1265	0.1472
CAE37501(BORPA)	AAS02924(MYCPA)	196.42	338.58	394.83	1138.17	0.8164	0.3789	0.4641

Table 5.1b: Pair-wise comparison of synonymous and non-synonymous substitutions within the Deoxyribose-phosphate aldolase protein cluster.

Seq 1	Seq 2	SynDif	NSynDif	SynPos	NSynPos	*dS*	*dN*	*dN/dS*
CAE48777(CORDI	CAD93351(MYCBO)	117.33	141.67	173.17	471.83	0.3835	1.7531	0.21876
CAE48777(CORDI	CAC31968(MYCLE)	123.83	146.17	173.58	471.42	0.4001	2.2649	0.17665
CAE48777(CORDI	CAB00939(MYCLEH)	117.33	141.67	173.17	471.83	0.3835	1.7531	0.218756
CAE48777(CORDI	AAK44719(MYCTU)	117.33	141.67	173.17	471.83	0.3835	1.7531	0.218756
CAE48777(CORDI	AAS06521(MYCPA)	113.17	135.83	176	469	0.366	1.4604	0.25062
CAE48777(CORDI	AAT83758(PROAC)	114.67	134.33	166.67	478.33	0.3518	1.8697	0.18816
CAD93351(MYCBO	CAC31968(MYCLE)	81	56	180.42	464.58	0.1314	0.6846	0.19194
CAD93351(MYCBO	CAB00939(MYCLEH)	0	0	180	465	0	0	0
CAD93351(MYCBO	AAK44719(MYCTU)	0	0	180	465	0	0	0
CAD93351(MYCBO	AAS06521(MYCPA)	87.67	48.33	182.83	462.17	0.1126	0.7648	0.14723
CAD93351(MYCBO	AAT83758(PROAC)	101.75	157.25	173.5	471.5	0.4412	1.1422	0.38627
CAC31968(MYCLE	CAB00939(MYCLEH)	81	56	180.42	464.58	0.1314	0.6846	0.19194
CAC31968(MYCLE	AAK44719(MYCTU)	81	56	180.42	464.58	0.1314	0.6846	0.19194
CAC31968(MYCLE	AAS06521(MYCPA)	102	57	183.25	461.75	0.1349	1.0165	0.13271
CAC31968(MYCLE	AAT83758(PROAC)	118.42	159.58	173.92	471.08	0.4507	1.7882	0.25204
CAB00939(MYCLEH	AAK44719(MYCTU)	0	0	180	465	0	0	0
CAB00939(MYCLEH	AAS06521(MYCPA)	87.67	48.33	182.83	462.17	0.1126	0.7648	0.14723
CAB00939(MYCLEH	AAT83758(PROAC)	101.75	157.25	173.5	471.5	0.4412	1.1422	0.38627
AAK44719(MYCTU	AAS06521(MYCPA)	87.67	48.33	182.83	462.17	0.1126	0.7648	0.14723
AAK44719(MYCTU	AAT83758(PROAC)	101.75	157.25	173.5	471.5	0.4412	1.1422	0.38627
AAS06521(MYCPA	AAT83758(PROAC)	95.75	153.25	176.33	468.67	0.4295	0.9655	0.44485

Table 5.1c: Pair-wise comparison of synonymous and non-synonymous substitutions within the MTB complex virulence factor MCE protein cluster.

Seq 1	Seq2	SynDif	NSynDif	SynPos	NSynPos	*dS*	*dN*	*dN/dS*
CAD93042(MYCBO)	CAD93469(MYCBO)	219.08	224.92	369.33	1067.67	1.1738	0.2473	0.210683
CAD93042(MYCBO)	CAC32124(MYCLE)	217.75	134.25	364.75	1072.25	1.1921	0.1370	0.114923
CAD93042(MYCBO)	CAB09755(MYCLEH)	0	0	365.83	1071.17	0	0	0.000000
CAD93042(MYCBO)	CAB09960(MYCLEH)	216.67	213.33	369.75	1067.25	1.1401	0.2325	0.203929
CAD93042(MYCBO)	AAS06157(MYCPA)	184.25	118.75	370.75	1066.25	0.8149	0.1206	0.147994
CAD93042(MYCBO)	AAS06637(MYCPA)	203.42	227.58	373.75	1063.25	0.9701	0.2520	0.259767
CAD93042(MYCBO)	AAK44401(MYCTU)	0	0	365.83	1071.17	0	0	0.000000
CAD93042(MYCBO)	AAK44846(MYCTU)	216.67	213.33	369.75	1067.25	1.1401	0.2325	0.203929
CAD93469(MYCBO)	CAC32124(MYCLE)	216.75	242.25	368.25	1068.75	1.1521	0.2699	0.234268
CAD93469(MYCBO)	CAB09755(MYCLEH)	219.08	224.92	369.33	1067.67	1.1738	0.2473	0.210683
CAD93469(MYCBO)	CAB09960(MYCLEH)	9.42	19.58	373.25	1063.75	0.0257	0.0186	0.723735
CAD93469(MYCBO)	AAS06157(MYCPA)	183.00	234.00	374.25	1062.75	0.7916	0.2607	0.329333
CAD93469(MYCBO)	AAS06637(MYCPA)	166.75	121.25	377.25	1059.75	0.6675	0.1241	0.185918
CAD93469(MYCBO)	AAK44401(MYCTU)	219.08	224.92	369.33	1067.67	1.1738	0.2473	0.210683
CAD93469(MYCBO)	AAK44846(MYCTU)	9.42	19.58	373.25	1063.75	0.0257	0.0186	0.723735
CAC32124(MYCLE)	CAB09755(MYCLEH)	217.75	134.25	364.75	1072.25	1.1921	0.1370	0.114923
CAC32124(MYCLE)	CAB09960(MYCLEH)	213.67	229.33	368.67	1068.33	1.1113	0.2529	0.227571
CAC32124(MYCLE)	AAS06157(MYCPA)	194.00	103.00	369.67	1067.33	0.9023	0.1033	0.114485
CAC32124(MYCLE)	AAS06637(MYCPA)	206.17	260.83	372.67	1064.33	1.0035	0.2967	0.295665
CAC32124(MYCLE)	AAK44401(MYCTU)	217.75	134.25	364.75	1072.25	1.1921	0.1370	0.114923
CAC32124(MYCLE)	AAK44846(MYCTU)	213.67	229.33	368.67	1068.33	1.1113	0.2529	0.227571
CAB09755(MYCLEH)	CAB09960(MYCLEH)	216.67	213.33	369.75	1067.25	1.1401	0.2325	0.203929
CAB09755(MYCLEH)	AAS06157(MYCPA)	184.25	118.75	370.75	1066.25	0.8149	0.1206	0.147994
CAB09755(MYCLEH)	AAS06637(MYCPA)	203.42	227.58	373.75	1063.25	0.9701	0.2520	0.259767
CAB09755(MYCLEH)	AAK44401(MYCTU)	0	0	365.83	1071.17	0	0	0.000000
CAB09755(MYCLEH)	AAK44846(MYCTU)	216.67	213.33	369.75	1067.25	1.1401	0.2325	0.203929
CAB09960(MYCLEH)	AAS06157(MYCPA)	181.75	221.25	374.67	1062.33	0.7805	0.2440	0.312620
CAB09960(MYCLEH)	AAS06637(MYCPA)	163.17	106.83	377.67	1059.33	0.6436	0.1083	0.168272
CAB09960(MYCLEH)	AAK44401(MYCTU)	216.67	213.33	369.75	1067.25	1.1401	0.2325	0.203929
CAB09960(MYCLEH)	AAK44846(MYCTU)	0	0	373.67	1063.33	0	0	0.000000
AAS06157(MYCPA)	AAS06637(MYCPA)	137.83	244.17	378.67	1058.33	0.4982	0.2757	0.553392
AAS06157(MYCPA)	AAK44401(MYCTU)	184.25	118.75	370.75	1066.25	0.8149	0.1206	0.147994
AAS06157(MYCPA)	AAK44846(MYCTU)	181.75	221.25	374.67	1062.33	0.7805	0.2440	0.312620
AAS06637(MYCPA)	AAK44401(MYCTU)	203.42	227.58	373.75	1063.25	0.9701	0.2520	0.259767
AAS06637(MYCPA)	AAK44846(MYCTU)	163.17	106.83	377.67	1059.33	0.6436	0.1083	0.168272
AAK44401(MYCTU)	AAK44846(MYCTU)	216.67	213.33	369.75	1067.25	1.1401	0.2325	0.203929

Table 5.2a: Nucleotide diversity of the beta-proteobacterial and MTB complex genes.

Type of site	Number of sites	Number of substitutions	Within species diversity (*dS* or *dN*)	dN/dS
BORPA				
Synonymous	403.33	235	0.63179	
Nonsynonymous	1180.67	268	0.38002	0.60150
BORPE				
Synonymous	408.92	156	0.69379	
Nonsynonymous	1199.08	143	0.32073	0.46229
MTB COMPLEX				
Synonymous	396.58	218	0.43101	
Nonsynonymous	1142.42	117	0.06297	0.14610
BRUMA				
Synonymous	none	none	none	
Nonsynonymous	none	none	none	na

This seems to indicate a decreased selective pressure against synonymous substitutions, or purifying selective pressure for the MTB complex genes. There is also a higher evolutionary pressure on non-synonymous substitutions per non-synonymous site (dN) for the proteobacteria and high-GC organisms than for MTB complex members of the same cluster. This low frequency of mutations in *M. tuberculosis* genes compared to other organisms has previously been observed by Sreevatsan *et al.* (1997).

Table 5.2b: Nucleotide diversity of the high-GC and MTB complex genes.

Type of site	Number of Sites	Number of substitutions	Within species diversity(*dS* or *dN*)	*dN/dS*
High GC bacterial				
Synonymous	168.83	131	1.90712	
Nonsynonymous	488.17	122	0.34935	0.183182
Mycobacterial complex				
Synonymous	188.43	143	0.54712	
Nonsynonymous	480.57	61	0.08781	0.160495

Table 5.2c: Nucleotide diversity of the MTB complex MCE genes

Type of site	Number of Sites	Number of substitutions	Within species diversity (*dS* or *dN*)	*dN/dS*
MYCTU				
Synonymous	395.58	231.75	1.139	
Nonsynonymous	1131.42	246.25	0.257	0.22562
MYCTU2				
Synonymous	395.58	231.75	1.139	
Nonsynonymous	1131.42	246.25	0.257	0.22562
MYCBO				
Synonymous	406.5	247.75	1.256	
Nonsynonymous	1174.5	147.25	0.137	0.10923
MYCLE				
Synonymous	none	none	none	
Nonsynonymous	none	none	none	na
MYCPA				
Synonymous	395.67	147.42	0.515	
Nonsynonymous	1101.33	269.58	0.296	0.57533

Looking more specifically at the different protein clusters, the dN/dS values for the bi-functional Gamma-glutamyltransferase (GGTA) protein cluster signify that the beta-proteobacterial copies of the gene seem to have greater diversity than the MTB complex genes (Table 5.2a), whereas the dN/dS values for High-GC and MTB complex organisms are much closer for the Deoxyribose-phosphate aldolase protein cluster (Table 5.2b). For the latter cluster the similarity in the values is a result of a combination of a much higher rate of synonymous substitutions in the High-GC organism gene and the much lower rate of non-synonymous substitutions in the MTB complex genes.

The different observations for the 2 clusters could be due to the fact that in the first cluster, the beta-proteobacteria copies contain paralogs from same species while MTB complex and high-GC members of same family (second cluster) contain one copy of the gene per species. It could therefore be possible that the duplicated copy of the gene may be losing functionality within the species. There are hypotheses of evolutionary scenarios where paralogs either acquire a new function that is different to the original function or where the paralogous copies share the function of the original protein [Nembaware *et. al.*, 2002].

Investigation of the dN/dS values for the virulence factor MCE protein family (MCE1D), containing MTB complex-specific genes shows that the dN/dS values are almost equal among different species of MTB complex organisms with the exception of *Mycobacterium avium paratuberculosis* having higher dN/dS values close to 0.6. Interestingly, all the members of the cluster have two copies of this gene except *Mycobacterium leprae* with only one copy (Table 5.1c). This could imply that more than one copy of this gene is important for increased virulence of the species or that one copy of these genes was lost in *Mycobacterium leprae* due to reductive evolution. This is consistent with the findings of Cole *et al.* [2001].

Comparative sequence analysis of *M. leprae* and *M. tuberculosis* provided evidence for reductive evolution in *M. leprae.* Many of the genes being lost in *M. leprae* tend to become pseudogenes or are inactivated as their functions are no longer required in highly specific niches [Cole *et al.*, 2001].

It is evident from the dN/dS ratios for the paralogous MTB complex genes that these genes generally show a significantly low value of dN/dS, suggesting that purifying

selection is acting on the genes. The values observed for the MTB complex specific cluster are close to what is observed in other clusters that are shared between MTB complex and other phylogenetic lineages.

The trends observed above could be due to the different evolutionary constraints that affect MTB complex organisms compared to other bacteria. This is in agreement with previous work that reported the ratio of non-synonymous substitutions to synonymous substitutions to be silent and much lower than 1.0 among MTB complex organisms compared to an average ratio close to 1 that is generally observed among other bacterial species [Scorpio and Zhang, 1996].

5.4.3 Polymorphic divergence

Polymorphic divergence is a measure of the amount of DNA variation between populations taking into account the effect of the DNA polymorphism. *KA/kS* is the ratio of non synonymous polymorphic divergence to synonymous polymorphic divergence between two populations. Here a population is defined as a set of sequences from the same species or phylogenetic lineage.

The divergence ratios, *kA/kS,* are generally very close between different species in each cluster tested (Table 5.3a, 5.3b and 5.3c). In Table 5.3a, the value of *kA/kS* between the MTB complex and beta-proteobacterial species only differs slightly (0.38846 and 0.38577, respectively). This seems to reject the previous hypothesis that the beta-proteobacteria gene copy might be losing functionality. When the same phylogenetic lineage species are compared the difference is higher (0.597) than the divergence between different species of beta-proteobacteria (Table 5.3a). This suggests that paralogs from the same species are more divergent compared to orthologs from different species, which is expected.

However, the number of net substitutions per site (DA values) indicates that the beta-proteobacterial lineage and MTB complex copies of the gene have the longest divergence time, and there is an almost equal divergence time within proteobacterial lineage. This means that the distance between the MTB complex and beta-proteobacterial lineage copy of the gene is greater than the distance between the different species of beta-proteobacterial lineage.

Table 5.3a: Nucleotide divergence between species of MTB complex and Beta-proteobacterial genes.

Type of site	Number of substitutions (S or N)	Divergence (kS or kA)	Number of net substitution per site (DA)	$\frac{kA}{kS}$
BORPA vs BORPE				
Synonymous	239	0.37133	-0.12911	0.59731
Non-synonymous	270	0.2218		
BORPA vs MTBCOMPLEX				
Synonymous	358	1.00186	0.15003	0.38847
Non-synonymous	330	0.38919		
BORPE vs MTBCOMPLEX				
Synonymous	391	1.02554	0.16585	0.38578
Non-synonymous	345	0.39563		

Table 5.3b: Divergence of the high-GC and *M. tuberculosis* complex genes.

Type of site	Number of substitutions (S or N)	Divergence (kS or kA)	Number of net substitution per site (DA)	$\frac{kA}{kS}$
Synonymous	165	1.34865		
Non-synonymous	60	0.41512	0.13426	0.3078

Table 5.3c: Divergence of the MTB complex genes.

Type of site	Number of substitutions (S or N)	Divergence (kS or kA)	Number of net substitutions per site (DA)	$\frac{kA}{kS}$
MYCTU vs MYCBO				
Synonymous	254	0.67791		
Nonsynonymous	220	0.15983	0.05847	0.23577
MYCTU vs MYCLE				
Synonymous	254	1.19409		
Nonsynonymous	220	0.15983	0.16344	0.13385
MYCTU vs MYCPA				
Synonymous	250	0.79219		
Nonsynonymous	214	0.18867	0.04325	0.23816
MYCBO vs MYCLE				
Synonymous	260	0.39176		
Nonsynonymous	135	0.06543	0	0.16702
MYCBO vs MYCPA				
Synonymous	172	0.92962		
Nonsynonymous	245	0.19803	0.12331	0.21302
MYCPA vs MYCLE				
Synonymous	241	0.96609		
Nonsynonymous	128	0.20351	0.13427	0.2106533

The divergence result allows an assumption to be made about when the duplication event occurred that created more than one copy of the gene in the beta-proteobacterial lineage. The assumption is that each lineage obtained a single copy of this gene from their ancestral gene and duplications occurred there after in beta-proteobacterial lineage copy of the gene.

5.4.4 Neutrality test

We conducted the McDonald and Kreitman [1991] test to investigate the hypothesis that all mutations are selectively neutral within and between the phylogenetic lineages within the same cluster [Kimura, 1983]. From the result in Table 5.4, we observed the ratio of non-synonymous fixed substitutions to synonymous fixed substitutions between lineages to be greater than the ratio of non-synonymous to synonymous polymorphisms within a lineage in all the clusters tested. Our results suggest that all mutations are not selectively neutral within and between the phylogenetic lineages within same cluster. Under the neutrality test as proposed by McDonald and Kreitman [1991], the ratio of non-synonymous fixed substitutions to synonymous fixed substitutions, or differences between phylogenetic lineages should be equal to the ratio of non-synonymous to synonymous polymorphisms within a phylogenetic lineage.

The results obtained above reject the neutral hypothesis tested and further confirm our observation in diversity (*dN/dS*) and divergence (k*A/kS*) results that there is an evolutionary constraint on non-synonymous mutations to maintain the functionality within members of the same cluster.

5.4.5 Phylogenetic trees

Generating phylogenetic trees for pairs of sequences follows the principles that the extent of sequence differences is proportional to the length of the autonomous sequence. The results of estimated evolutionary distance and bootstrap neighbour-joining evolutionary distance trees of genes members of the three clusters tested are shown in Figures 5.4, 5.5 and 5.6. The trees show that the distance between orthologs from different species within the same cluster are shorter than the distance between paralogs from the same species. Interestingly homologs (paralogs and orthologs) from same phylogenetic lineage cluster together.

Table 5.4: Phylogenetic lineage neutrality test

	Beta-proteomes and MTB complex	High G+C and MTB complex
Fixed differences between species		
Synonymous	34	39
Nonsynonymous	58	60
Fixed differences ratio	1.70588	1.53846
Polymorphic changes within species		
Synonymous	334	237
Nonsynonymous	238	160
Polymorphic changes ratio	0.71258	0.67511

For example, Figure 5.4 reveals the relationship between the bi-functional gamma-glutamyltransferase (**GGTA)** protein from 3 beta-proteobacteria species and 4 members of MTB complex organisms. Gene members from the beta-proteobacteria lineage are shown in BLUE and MTB Complex in RED. In the tree, MTB complex genes cluster together and their homologs from beta-proteobacteria also cluster together separately. However, paralogous genes from within species (BORPA and PORPE) are more phylogenetically distant from each other than the orthologs from these two species, which is to be expected.

Generally the trees followed phylogenetic profiles of the gene members of the proteins rather than the species phylogeny. This is further confirmed by Figure 5.6 which shows the relationships between duplicated virulence factor mce proteins of the MTB complex. This tree also reveals that paralogs from the same species are more distant to each other than the orthologs in different species.

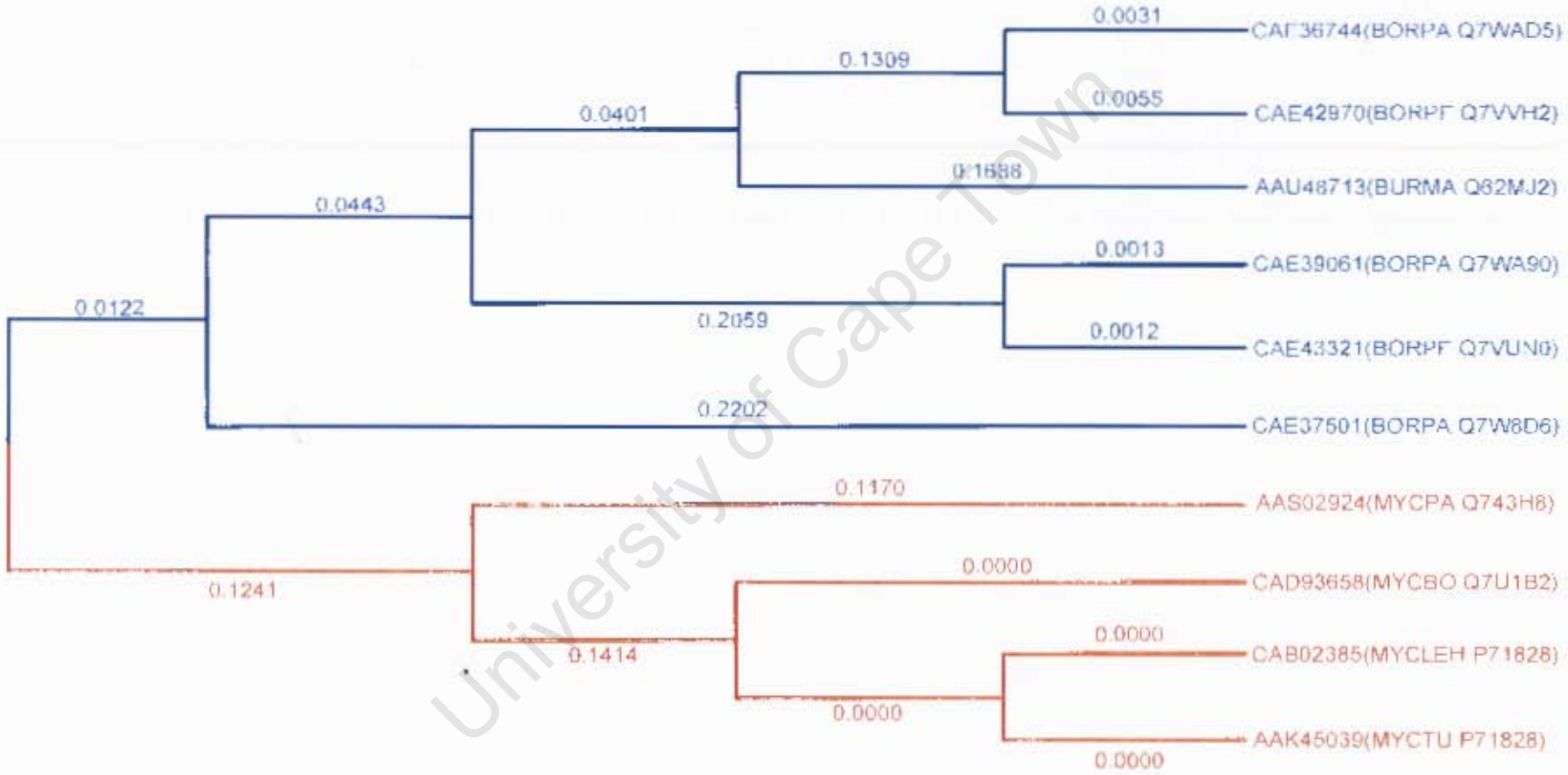

Figure 5.4: Evolutionary distance and bootstrap neighbour-joining evolutionary distance tree for the bi-functional gamma-glutamyltransferase (GGTA) protein cluster.

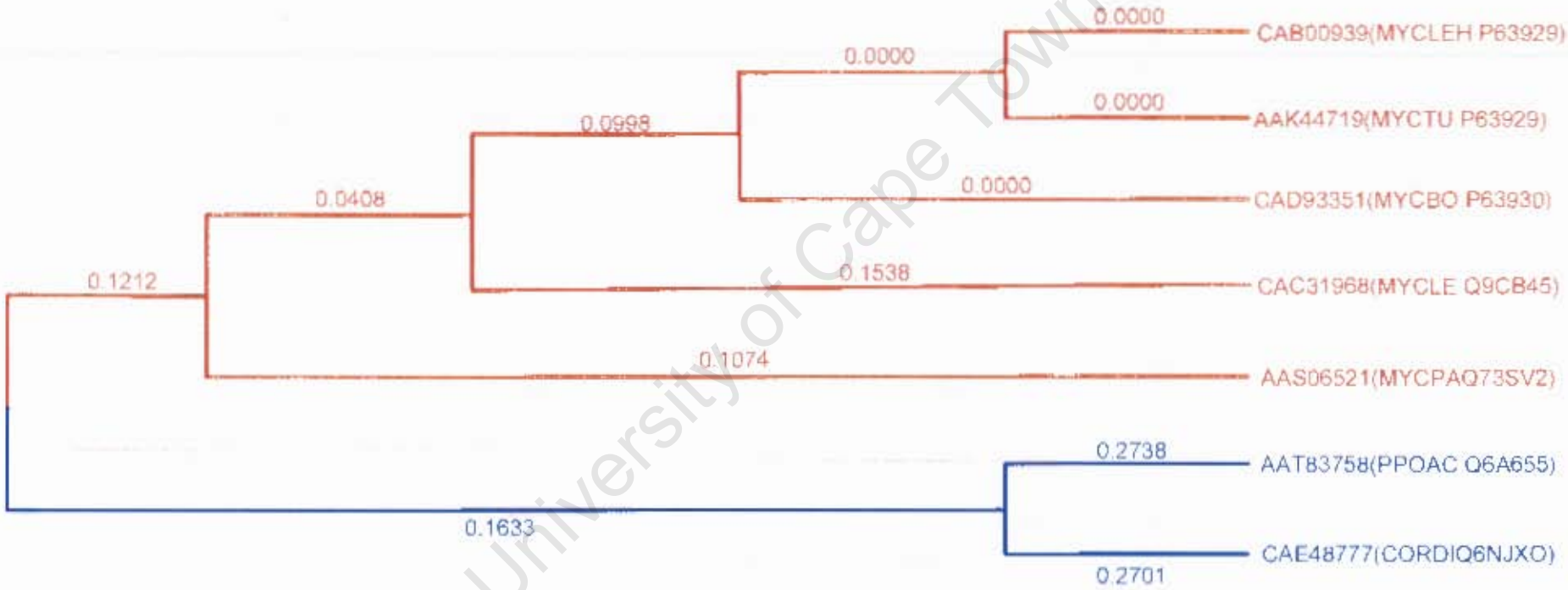

Figure 5.5: Evolutionary distance and bootstrap neighbour-joining evolutionary distance tree for the deoxyribose-phosphate aldolase protein cluster

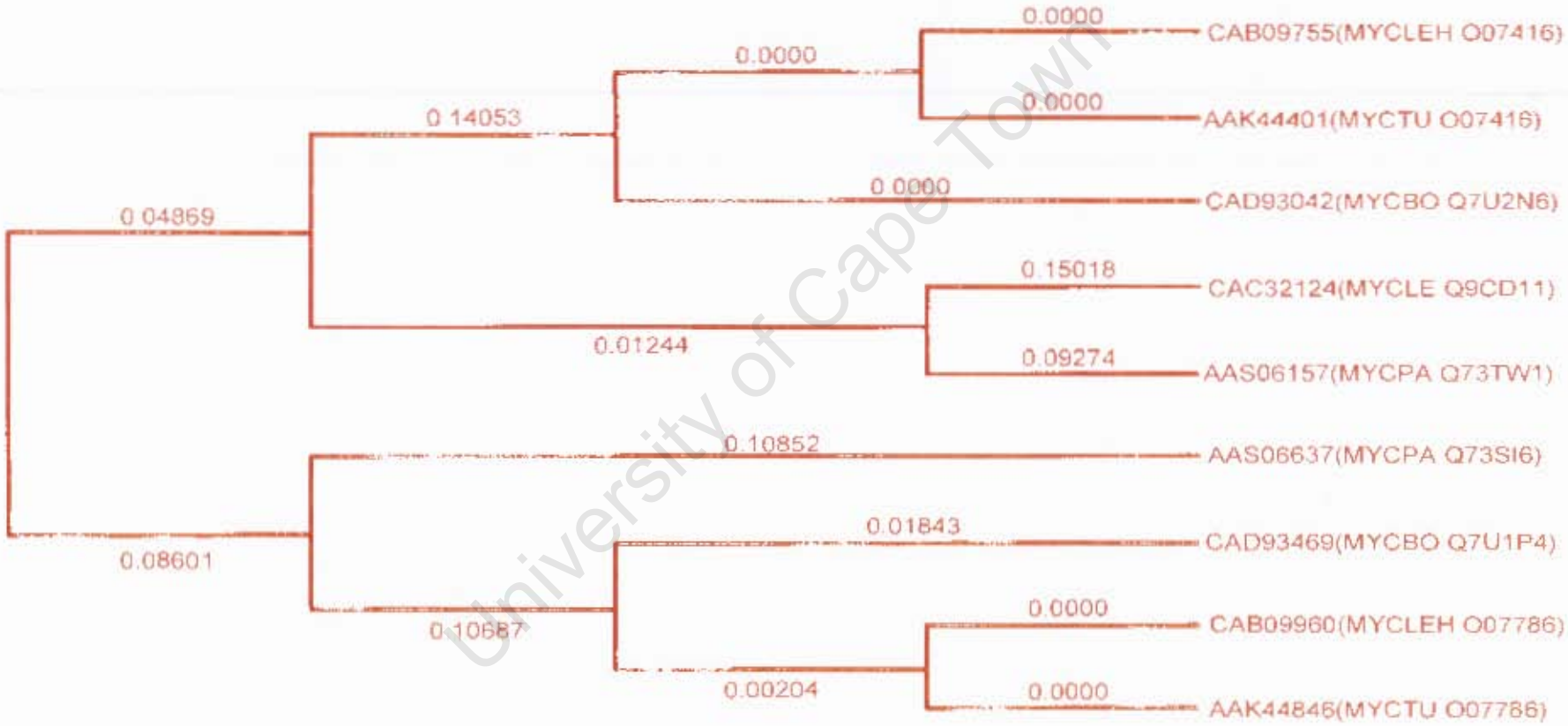

Figure 5.6: Evolutionary distance and bootstrap neighbour-joining tree for MTB complex virulence factor MCE protein cluster.

5.5 Conclusions

Comparative evolutionary analysis studies are routinely employed for phylogeny studies. However, it has been shown (in this chapter) that it can also be utilized for identification of unique proteins at different levels between different organisms. A detailed investigation of functional or structural equivalence among homologous proteins from different organisms can also be revealed by evolutionary analysis.

The evolutionary analysis results of different clusters confirmed that protein members from the same cluster are evolutionary related to each other and there is evidence of purifying selection against non-synonymous mutations and substitutions to maintain the same function or structure among the members of same cluster created by the experiment conducted in chapter three.

6

General Conclusions and Future Work

Simplicity is the final achievement. After one has played a vast quantity of notes and more notes, it is simplicity that emerges as the crowning reward of art.

- Frederic Chopin

The study of genome offers high prospects towards a better understanding of bacteria in general. Whole genome comparative genomic analysis of microbial pathogens and their associated disease processes has facilitated identification of new and better drugs and therapeutic targets. There is still the urgent need for functional annotation for most of the recently sequenced proteins and those pathogen proteins that are formerly annotated as unknown or hypothetical proteins. This will enable researchers to apply these data to understand what distinguishes pathogenic from non-pathogenic species and develop a faster and more effective drugs target for known pathogens and/or their proteins.

In this work, we have produced a microbial phylogenetic profile for 84 genomes including 56 pathogens and 28 non pathogens. We identified and started characterizing a set of genes from *M. tuberculosis* that were unique to pathogens or MTB complex organisms, and thus might be involved in the virulence and pathogenesis of *these* pathogens.

The set of predicted proteins included many surface or secreted proteins, as well as some that are known to be involved in functions related to pathogenesis. The set also included a large number of hypothetical proteins that may be of interest for further study.
We predicted potential functions for some of these proteins that were previously unknown or annotated as hypothetical proteins.

We further perform evolutionary analysis to investigate some predicted virulence gene clusters to determine how these genes evolved in pathogens. We detected evidence of

purifying selection against non-synonymous polymorphisms and short evolutionary distance among related sequences from different pathogens.

This evidence suggested that the same function or structure are likely to be maintained among members of the tested clusters, and allow us to further refine and to broaden annotation of protein functions in assisting drug development.

However, to further confirm the predicted functions for all predicted virulence genes, more research efforts are needed to check whether they are also involved in similar pathways and whether they share similar expression profiles. In particular, among other things, future work should:

- Investigate the potential role of these genes in virulence by understanding the stage and level of their expressions and identification of co-expressed genes.

- Perform *in silico* comparative analysis with various host genomes to understand and gain insights into the mechanisms surrounding virulence of these pathogens, with emphasis on how they adapt to and evade the host immune response and how they interact with host genes. This can be achieved using functional information and available expression data.

References

Adams, M.B. [2003] Ecological issues related to N deposition to natural ecosystems: research needs, *Environ. Int.* **29** pp. 189–199.

Alland, D., Lacher, D.W.,Hazbon, M.H., Motiwala, A.S.,Qi, W., Fleischmann, R.D. and Whittam, T.S. [2007] Role of large sequence polymorphisms (LSPs) in generating genomic diversity among clinical isolates of *Mycobacterium tuberculosis* and the utility of LSPs in phylogenetic analysis. *J Clin Microbiol.* **45** (1) pp 39-46.

Altschul, S.F., Gish, W., Miller, W., Meyers, E.W. and Lipman, D.J. [1990] Basic local alignment search tool. *Journal of Molecular Biology.* **215** pp 403-410.

Amos, L.A., van den Ent, F. and Lowe, J. [2004] Structural/functional homology between the bacterial and eukaryotic cytoskeletons. *Curr. Opin. Cell Biol.* **16** pp 24–31

Andreas, S., Francisco, S.D., Jörg, R. and Thomas, L. [2006] A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics* **7** p 302. doi:10.1186/1471-2105-7-302

Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, R., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou, Y., Lopez, R., Marx, B., Mulder, N.J., Oinn, T. M., Pagni, M., Servant, F., Sigrist, C.J.A and Zdobnov, E.M.[2001] The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Research,* **29** (1) pp 37-40.

Banu, S., Honoré, N., Saint-Joanis, B., Philpott, D., Prévost, M. and Cole, S.T [2002] Are the PE-PGRS proteins of *Mycobacterium tuberculosis* variable surface antigens? *Molecular Microbiology*, **44** (1) pp 9-19(11)

Barnes, P.F., Bloch, A.B., Davidson, P.T. and Snider, D.E. [1991] Tuberculosis in patients with human immunodeficiency virus infection. *N Engl J Med*; **32** pp 1644–50

Bardarov, S., Kriakov, J., Carriere, C., Yu, S., Vaamonde, C. and McAdam, R.A. [1997] Conditionally replicating mycobacteriophages: a system for transcriptase delivery to *Mycobacterium tuberculosis*, *Proc Natl Acad Sci USA* 94 pp. 10965 -10966.

Bateman, A. , Coin, L., Durbin, R., Finn, R. D.,Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L. L., Studholme, D. J., Yeats, C. and Eddy, S. R. [2004] The Pfam protein families database. *Nucleic Acids Res.* **32** Database Issue D138–D141.

Bates, J.H. and Stead, W.W [1993] The history of tuberculosis as a global epidemic. *Med Clin North Am* **77** (6) pp 1205-1217

Baumann, S., Eddine, A.N. and Kaufmann, H.E. [2006] Progress in tuberculosis vaccine development. *Current Opinion in Immunology* **18** (4) pp 438-448.

Becker, G. J. [2005] Bioinformatics, Computational Biology, and the National Centers for Biomedical Computing, Journal of the American College of Radiology, **2** (5) pp 398-400.

Beck-Sague, C., Dooley S.W., Hutton, M.D., Otten, J., Breeden, A., Crawford, J.T., Pitchenik, A.E., Woodley, C., Cauthen, G. and Jarvis, W.R. [1992] Hospital outbreak of multidrug-resistant *Mycobacterium tuberculosis* infections: factors in transmission to staff and HIV-infected patients. *JAMA* 268 pp 1280-1286.

Betts, J.C., Dodson, P., Quan, S., Lewis, A.P., Thomas, P.J., Duncan, K., and McAdam, R.A. [2000] Comparison of the proteome of *Mycobacterium tuberculosis strain* H37Rv with clinical isolate CDC 1551. *Microbiology* **146** pp 3205–3216.

Bininda-Emonds, O.R.P., Gittleman, J. L. and Steel, M.A. [2002] The super tree of life: Procedures, problems, and prospects. *Annu Rev Ecol Syst* **33** pp 265–289.

Binnewies, T.T., Y. Motro, P.F. Hallin, O. Lund, D. David Dunn, T. La, D.J. Hampson, M. Bellgard, T.M. Wassenaar and D.W. Ussery [2006] Ten years of bacterial genome sequencing: 10 comparative-genomics-based discoveries, *Funct Integr Genomics* **6** pp. 165–185.

Biswas, M., O'Rourke, J.F., Camon, E., Fraser, G., Kanapin, A., Karavidopoulou, Y., Kersey, P., Kriventseva, E., Mittard, V., Mulder, N., Phan, I, Servant, F and Apweiler, R. [2002] Applications of InterPro in protein annotation and genome analysis. *Briefings in Bioinformatics.* **3** (3) pp 285–295.

Blanchard, J. L [2004] Bioinformatics and Systems Biology, rapidly evolving tools for interpreting plant response to global change. *Field Crops Research* **90** (1) pp 117-131.

Blower, S. M., Small, P.M. and Hopewell, P. C. [1996] Control Strategies for Tuberculosis Epidemics: New Models for Old Problems. *Science* **273** (5274) pp. 497 – 500.

Blower, S.M., McLean, A.R., Porco, T.C., Small, P.M., Hopewell, P.C., Sanchez, M.A. and Moss, A.R. [1995]. The intrinsic transmission dynamics of tuberculosis epidemics. *Nat Med.* **1** (8) pp 815-21.

Böttger, E. C., Springer, B., Pletschette, M., and Sander, P. [1998] Fitness of antibiotic resistant microorganisms and compensatory mutations *Nature Medicine* 4 pp 1343 – 1344.

Bradford, W.Z., Jeffrey N.M., Arthur L.R., Gisela F.S., Philip, C.H. and Peter, M.S. [1996] The changing epidemiology of acquired drug-resistant tuberculosis in San Francisco, USA, *The Lancet*, **348** pp 928-931.

Brosch, R., Pym, A.S., Gordon, S.V., and Cole, S.T. [2001] The evolution of mycobacterial pathogenicity: clues from comparative genomics. *Trends in Microbiology* **9** (9) pp 452 – 458.

Brosch, R., Gordon, S.V., Eiglmeier, K., Garnier, T., Takaia, F., Yeramian, E. and Cole, S.T. [2000] Genomics, biology, and evolution of the *Mycobacterium tuberculosis* complex. In: Hatfull, G.F. and Jacobs Jr., W.R., Eds Molecular Genetics of Mycobacteria, ASM Press, Washington, DC.

Brosch, R., Pym, A.S., Gordon, S.V. and Cole, S.T. [2001] The evolution of mycobacterial pathogenicity: clues from comparative genomics. *Trends Microbiol* **9** pp 452– 458.

Brosch, R., Gordon, S.V., Marmiesse, M., Brodin, P., Buchrieser, C., Eiglmeier, K., Garnier, T., Gutierrez, C., Hewinson, G., Kremer, K., Parsons, L.M. Pym, A.S., Samper, S., van Soolingen, D. and Cole, S.T. [2002] A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proc Natl Acad Sci USA* **99**, pp 3684-3689.

Burgos, M.V. and Pym, A.S. [2002] Molecular epidemiology of tuberculosis. *Eur Respir J.* **20** pp 54S-65S.

Calamita, H., Ko, C., Tyagi, S., Yoshimatsu, T., Morrison, N.E. and Bishai, W.R. (2005) The *Mycobacterium tuberculosis* SigD sigma factor controls the expression of ribosome-associated gene products in stationary phase and is required for full virulence. *Cellular Microbiology* **7** (2) pp 233–244.

Camacho, L.R., Ensergueix, D., Perez, E., Gicquel, B. and Guilhot, C. [1999] Identification of a virulence gene cluster of *Mycobacterium tuberculosis* by signature-tagged transposon mutagenesis. *Molecular Microbiology* **34** (2) pp 257-267.

Camus, J.C., Pryor, M.J., Medigue, C. and Cole, S.T. [2002] Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv. *Microbiology* **148** pp 2967- 2973

Cannon, S.B. and Young, N.D. [2003] OrtholParaMap: distinguishing orthologs from paralogs by integrating comparative genome data and gene phylogenies. *BMC Bioinformatics*, **4** (1) p 35.

Cardoso, R.F., Cooksey, R.C., Morlock, G.P., Barco, P., Cecon, L., Forestiero, F., Leite, C.Q., Sato, D.N., Shikama Mde, L., Mamizuka, E.M., Hirata, R.D. and Hirata, M.H.[2004] Screening and characterization of mutations in isoniazid-resistant *Mycobacterium tuberculosis* isolates obtained in Brazil. *Antimicrob. Agents Chemother.* **48** (9) pp 3373-3381.

CDCP, Centers for Disease Control and Prevention [1992] National Action Plan to Combat Multidrug-Resistant Tuberculosis. *MMWR*, **41** (RR-11). pp 1-48.
http://www.cdc.govmmwr/ review/mwrhtml/ 0031159.htm

CDCP, Centers for Disease Control and Prevention [1999] Reported Tuberculosis in the United States. August 1999.

Chen, F., Mackey, A.J., Stoeckert, C.J. Jr. and Roos, D.S. [2006] OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res* **34** (Database) pp D363-368.

Clarke, S.C. [2005] Pyrosequencing: nucleotide sequencing technology with bacterial genotyping applications, *Expert Rev Mol Diagn* **5** pp. 947–953.

Cole S.T., Brosch R., Parkhill J., Garnier T.; Churcher C., Harris, D., Gordon, S. V., Eiglmeier, K., Gas, S., Barry, C. E., Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Krogh, A., McLean, J., Moule, S., Murphy, L.,Oliver, K., Osborne, J., Quail, M.A., Rajandream, M.A, Rogers, J., Rutter, S., Seeger, K.,

Skelton, J., Squares, R., Squares, S., Sulston, J. E., Taylor, K., Whitehead, S. and Barrell, B.G. [1998]. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**, pp. 537–544.

Cole, S.T., Eiglmeier, K., Parkhill, J., James, K.D., Thomson, N.R., Wheeler, P.R., Honoré, N., Garnier, T., Churcher, C., Harris, D., Mungall, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R.M., Devlin, K., Duthoy, S., Feltwell, T., Fraser, A., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Lacroix, C., Maclean, J., Moule, S., Murphy, L., Oliver, K., Quail, M.A, Rajandream, M.A., Rutherford, K.M., Rutter, S., Seeger, K., Simon, S., Simmonds, M., Skelton, J., Squares, R., Squares, S., Stevens, K., Taylor, K., Whitehead, S., Woodward, J.R. and B. G. Barrell [2001] Massive gene decay in the leprosy bacillus. *Nature* **409** pp 1007–1011.

Cole, S.T. [1998] Comparative mycobacterial genomics. *Curr. Opin. Microbiol.* **1** p 567–571.

Cole, S.T. [1999] Learning from the genome sequence of *Mycobacterium tuberculosis* H37Rv. *FEBS Letters* **452** pp 7–10.

Cole, S.T. [2002] Comparative mycobacterial genomics as a tool for drug target and antigen discovery. *Eur Respir J* **20** Suppl. 36 pp 78s–86s

Comstock, G.W. [1999] How much isoniazid is needed for prevention of tuberculosis among immunocompetent adults? *Int J Tuberc Lung Dis* **3** pp 847-850.

Cooksey, R.C., Morlock, G.P., McQueen, A., Glickman, S.E. and Crawford, J.T. [1996] Characterization of streptomycin resistance mechanisms among *M. tuberculosis* isolates from patients in New York City. *Antimicrob. Agents Chemother.* **40** pp 1186−1188.

Coronado, V.G., Beck-Sague, C.M., Hutton, M.D., Coronado, V.G., Beck-Sague, C.M., Hutton, M.D., Davis, B.J., Nicholas, P., Villareal, C., Woodley, C.L., Kilburn, J.O., Crawford, J.T. and Frieden, T.R. [1993] Transmission of multidrug-resistant *Mycobacterium tuberculosis* among persons with human immunodeficiency virus infection in an urban hospital: epidemiologic and restriction fragment length polymorphism analysis. *J Infect Dis.* **168** pp 1052-1055.

Daffel, M. and Draper, P. [1998] The envelope layers of mycobacteria with reference to their pathogenicity. *Adv Microb Physiol* **39**, 131-203.

Daniel, T.M [2006] The history of tuberculosis *Respir Med.* **100** (11) pp 1862-70.

Daniel, T.M [2000] Pioneers of Medicine and their Impact on Tuberculosis. Rochester, N.Y.: University of Rochester Press.

Daubin, V., Gouy, M. and Perriere, G. [2002] A phylogenomic approach to bacterial phylogeny: Evidence of a core of genes sharing a common history. *Genome Res* **12** pp 1080–1090.

Davidson, W. M [2006] Structure and Function of Cells and Viruses. Cell Biology and Microbiology. Molecular Expressions. http://micro.magnet.fsu.edu/cells/bacteriacell.html

Deng, W., Liou, S.R., Plunkett, G., Mayhew, G.F., Rose, D.J., Burland, V., Kodoyianni, V, Schwartz, D.C. and Blattner, F.R [2003] Comparative genomics of *Salmonella enterica serovar* Typhi strains Ty2 and CT18. *Bacteriol.* **185** (7) pp 2330-2337.

den Dunnen, J.T. and Antonarakis, S.E. [1999] Mutation nomenclature extensions and suggestions to describe complex mutations: A discussion. *Human Mutation.* **15** (1) pp 7-12

De Pinna, M.C. [1991] Concepts and tests of homology in the cladistic paradigm. *Cladistics.* **7** pp 367-394.

Devulder, G., Pérouse de Montclos, M. and Flandrois, J.P. [2005] A multi-gene approach to phylogenetic analysis using the genus *Mycobacterium* as a model. *Int J Syst Evol Microbiol* 55 pp 293-302; DOI 10.1099/ijs.0.63222-0

Dietrich, J., Lundberg, C.V. and Andersen, P. [2006] TB vaccine strategies-what is needed to solve a complex problem? *Tuberculosis.* **86** (3-4) pp 163-168.

Dlodlo, R.A., Fujiwara, P.I. and Enarson, D.A. [2005] Should tuberculosis treatment and control be addressed differently in HIV-infected and -uninfected individuals? *Eur Respir J.* **25** pp 751-757.

Dolin, R.M. and Kochi, A. [1994] Global tuberculosis incidence and mortality during 1990–2000, *Bull World Health Organ* **72**, pp. 213–220.

Doetsch, R. N. [1978] Benjamin Marten and his New Theory of Consumptions. *Microbiol. Rev.* **42** (3) pp 521–528.

Dooley, S.W., Villarino, M.E., Lawrence, M., Salinas, L., Amil, S., Rullan, R.V., Jarvis, W.R., Bloch, A.B. and Cauthen, G.M. [1992] Nosocomial transmission of tuberculosis in a hospital unit for HIV-infected patients. *JAMA* **267** pp 2632-2634.

Dooley, S.W., Jarvis, W.R., Martone, W.J. and Snider, D.E. [1992] Multidrug-resistant tuberculosis. *Ann. Intern. Med.* **117** pp 257-259.

Doolittle, W.F. [1999] Phylogenetic classification and the universal tree. *Science* **284** pp 2124– 2129

Doolittle, W.F. [2000] Uprooting the tree of life. *Sci. Am.* **282** pp 90–95.

Dufayard, J.F., Duret, L., Penel, S., Gouy, M., Rechenmann, F. and Perriere, G. [2005] Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics.* **21** (11) pp 2596 – 2603.

Durbin, R., Eddy, S.R., Krogh, A. and Mitchison, G. [1998] *Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids*, Cambridge, England Cambridge University Press.

Dye, C., Suzanne, S., Paul, D., Vikram, P. and Mario, C.R. [1999] Global Burden of Tuberculosis- Estimated Incidence, Prevalence, and Mortality by Country. WHO Global Surveillance and Monitoring Project. *JAMA.* **282** pp 677-686.

Edlin, B.R., Tokars, J.I., Grieco, M.H., Crawford, J.T., Williams, J., Sordillo, E.M., Ong, K.R. Kilburn, J.O., Dooley, S.W. and Castro, K.G. [1992] An outbreak of multidrug-resistant tuberculosis among hospitalized patients with the acquired immunodeficiency syndrome. *N. Engl. J. Med.* **326** pp 1514-1521.

Edwards, R.A., Rodriguez-Brito, B., Wegley, L., Haynes, M., Breitbart, M., Peterson, D.M., Saar, M.O., Alexander, S., Alexander, C.E. Jr. and Rohwer, F. [2006] Using pyrosequencing to shed light on deep mine microbial ecology under extreme hydrogeologic conditions, *BMC Genomics* **7** p. 57.

Efstratiadis, A., Kafatos, F. C. and Maniatis, T. [1977]. The primary structure of rabbit beta-globin mRNA as determined from cloned DNA. *Cell* **10,** pp 571-585.

Eiglmeier, K., Parkhill, J., Honore, N., Garnier, T., Tekaia, F., Telenti, A., Klatser, P., James, K.D., Thomson, N.R., Wheeler, P.R., Churcher, C., Harris, D., Mungall, K., Barrell, B.G. and Cole, S.T. [2001] The decaying genome of *Mycobacterium leprae. Lepr Rev* **72** pp 387-398.

Elad Ziv, Charles, L.D. and Sally, M.B. [2001] Early Therapy for Latent Tuberculosis Infection. *American Journal of Epidemiology* **153** (4) pp 381-385.

El Sahly, H.M., Teeter, L.D., Pawlak, R.R., Musser, J.M. and Graviss, E.A. [2006] Drug-resistant tuberculosis: A disease of target populations in Houston, Texas. *J. of Infection* **53** pp 5–11

Eulenstein, O., Mirkin, B. and Vingron, M. [1998] Duplication-based measures of difference between gene and species trees. *J. Comput. Biol.* **5** pp 35–48.

Ferebee, S [1970] Controlled chemoprophylaxis trials in tuberculosis. A general review. *Bibl. Tuberc.* **26** pp 28–106.

Ferretti, J.J., Ajdic, D. and McShan, W.M. [2004] Comparative genomics of streptococcal species. *Indian J. Med. Res.* **119** Suppl:1-6.

Field, M., Wilson, G., and van der Gast, W. [2006] How do we compare hundreds of bacterial genomes?*Current Opinion in Microbiology*, Article in Press. http://dx.doi.org/ 10.1016/j.mib.2006.08.008

Filliol, I., Motiwala, A.S., Cavatore, M., Qi, W., Hazbon, M.H., Bobadilla del Valle, M., Fyfe, J., Garcia-Garcia, L., Rastogi, N., Sola, C., Zozio, T., Guerrero, M.I., Leon, C.I., Crabtree, J., Angiuoli, S., Eisenach, K.D., Durmaz, R., Joloba, M.L., Rendon, A., Sifuentes-Osornio, J., Ponce de Leon, A., Cave, M.D., Fleischmann, R., Whittam, T.S. and Alland, D. [2006] Global phylogeny of *Mycobacterium tuberculosis* based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set. *J. Bacteriol.* **188** (2) pp 759-72

Fischl, M.A., Uttamchandani, R.B., Daikos, G.L., Poblete, R.B., Moreno, J.N., Reyes, R.R., Boota, A.M., Thompson, L.M., Cleary, T.J. and Lai, S. [1992] An outbreak of tuberculosis caused by multiple-drug-resistant tubercle bacilli among patients with HIV infection. *Ann. Intern Med.* **117** pp177-183.

Fitch, W.M [1970] Distinguishing homologous from analogous proteins. *Syst. Zool.* **19** pp 99–106.

Fitch, W.M [2000] Homology a personal view on some of the problems. *Trends Genet* **16** pp 227–31

Fleischmann, R.D., Alland, D., Eisen, J.A., Carpenter, L., White, O., Peterson, J., DeBoy, R., Dodson, R., Gwinn, M., Haft, D., Hickey, E. , Kolonay, J.F., Nelson, W.C., Umayam, L.A., Ermolaeva, M., Salzberg, S.L., Delcher, A., Utterback, T., Weidman, J., Khouri, H., Gill, J., Mikula, A., Bishai, W., Jacobs Jr, W.R., Venter, J.C. and Fraser, C. M. [2002]. Whole-Genome Comparison of *Mycobacterium tuberculosis* Clinical and Laboratory Strains. *Journal of Bacteriology* **184** pp 5479–5490.

Frieden, T.R. and Munsiff, S.S. [2005] The DOTS strategy for controlling the global tuberculosis epidemic. *Clinics in Chest Medicine* **26** (2) pp 197-205.

Frothingham, R. [1995] Differentiation of strains in *Mycobacterium tuberculosis* complex by DNA sequence polymorphisms, including rapid identification of *M. bovis* BCG. *J Clin Microbiol.* **33** (4) pp 840–844.

Fruth, U. and Young, D. [2004] Prospects for new TB vaccines: stop TB working group on TB vaccine development, *Int J Tuberc Lung Dis* **8**, pp. 151–155

Garcia-Vallve, S., Romeu, A. and Palau, J. [2000] Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res* **10** pp 1719-1725.

Gene Ontology Consortium [2006] The Gene Ontology (GO) project in 2006 *Nucleic Acids Research*, **34**, Database issue pp. D322-D326

Gey van Pittius, N.C., Sampson, S.L. Lee, H., Kim, Y., van Helden, P.D. and Warren, R.M. [2006] Evolution and expansion of the *Mycobacterium tuberculosis* PE and PPE multigene families and their association with the duplication of the ESAT-6 (esx) gene cluster regions. *BMC Evolutionary Biology* **6** p95 doi:10.1186/1471-2148-6-95

Gil, R., Latorre, A. and Moya, A. [2004] Bacterial endosymbionts of insects: insights from comparative genomics. *Environ Microbial.* **6** (11) pp 1109-22.

Glickman, M.S. and Jacobs, W.R. [2001] Microbial Pathogenesis of *Mycobacterium tuberculosis:* Dawn of a Discipline. *Cell* **104** (4) pp 477-485.

Global Tuberculosis Institute [GTI] TB History: A History of Tuberculosis Treatment. http://www.umdnj.edu/ntbcweb/tbhistory.htm

Gordon, S.V., Heym, B., Parkhill, J., Barrell, B. and Cole, S.T. [1999] New insertion sequences and a novel repeated sequence in the genome of *Mycobacterium tuberculosis* H37Rv. *Microbiology* **145** pp 881-892

Gordon, S.V., Eiglmeier, K, Garnier, T., Brosch, R., Parkhill, J., Barrell, B., Cole, T., and Hewinson, R.G. [2001] Genomics of *Mycobacterium bovis. Tuberculosis (Edinb).* **81** (1-2) pp 157-163.

Gradmann, C.[2001] Robert Koch and the Pressures of Scientific Research, *Tuberculosis and Tuberculin, Medical History* **45** pp. 1–32.

Gradmann, C. [2006] Robert Koch and the white death: from tuberculosis to tuberculin. *Microbes Infect.* **8** (1) pp 294-301.

Grannich, R.M., Balandrano, S., Santaella, A.J., Binkin, N.., Castro, K.G., Marquez-Fiol, A., Anzaldo, G., Zarate, M., Jaimes, M.L., Velazquez-Monroy, O., Salazar, L., Alvarez-Lucas, C., Kuri, P., Flisser, A., Santos-Preciado, J., Ruiz-Matus, C., Tapia-Conyer, R. and Tappero, J.W. [2000] Survey of Drug Resistance of *Mycobacterium tuberculosis* in 3 Mexican States, 1997. *Arch Intern Med* **160** pp 639-644.

Griffiths, R.I., Mark, J.B., Niall, P.M. and Andrew, S.W. [2006] The functions and components of the Sourhope soil microbiota. *Applied Soil Ecology* **33** (2) pp 114-126.

Groenen, P., Hulsen, T., Huynen, M.A., and de Vlieg, J. [2006] Benchmarking ortholog identification methods using functional genomics data. *Genome Biology* 2006, **7** R31. Available online http://genomebiology.com/2006/7/4/R31

Groenen, P.M., Bunschoten, A.E., van Soolingen, D., and van Embden, J.D. [1993] Nature of DNA polymorphism in the direct repeat cluster of *Mycobacterium tuberculosis*; application for strain differentiation by a novel typing method. *Mol Microbiol.* **10**(5) pp 1057–1065.

Haas, F., and Haas, S.S [1996] The origins of *Mycobacterium tuberculosis* and the notion of its contagiousness In W. N. Rom and S. Garay (ed.), Tuberculosis. Little, Brown and Co., Boston, Mass. p 3–19.

Heifets, L.B. and Good, R.C. [1994] Current laboratory methods for the diagnosis of tuberculosis, *In*. Bloom, B.R (ed.), Tuberculosis: pathogenesis, protection, and control. ASM Press, Washington, D.C. pp 85-110.

Hillis, D.M. [1994] Homology in molecular biology. In: Homology: the hierarchical basis of comparative biology. Hall, B.K. (ed.) Academic press, San Diego.

Holt, J.G. [1994] Bergey's *Manual of Determinative Bacteriology*. 9th ed.. Williams and Wilkins, Baltimore, Maryland.

Hutton, M.D., Stead, W.W., Cauthen, G.M., Bloch, A.B. and Ewing, W.M. [1990] Nosocomial transmission of tuberculosis associated with a draining abscess. *J. Infect. Dis*. **161** pp 286-295.

Jothi, R., Zotenko, E., Tasneem, A. and Przytycka, T.M. [2006] COCO-CL: hierarchical clustering of homology relations based on evolutionary correlations. *Bioinformatics*, **22** (7) pp 779 - 788.

Karboul, A., van Pittius,N.C., Namouchi, A., Vincent, V., Sola, C., Rastogi, N., Suffys, P., Fabre, M., Cataldi, A., Huard, R.C., Kurepina, N., Kreiswirth, B., Ho, J.L., Gutierrez, M.C. and Mardassi, H. [2006] Insights into the evolutionary history of tubercle bacilli as disclosed by genetic rearrangements within a PE-PGRS duplicated gene pair *BMC Evol Biol.* **6** p 107. doi: 10.1186/1471-2148-6-107.

Kim, S.J. [2005] Drug-susceptibility testing in tuberculosis: methods and reliability of results. *Eur Respir J* **25**, pp 564-569.

Mirkin, B., Muchnik, I and Smith, T.F. [1995] A biologically consistent model for comparing molecular phylogenies. *J. Comput. Biol.* **2** pp 493–507

Mora, M., Donati, C., Medini, D., Covacci, A and Rappuoli, R. [2006] Microbial genomes and vaccine design: refinements to the classical reverse vaccinology approach, *Current Opinion in Microbiology*. Article in Press, doi:10.1016/j.mib.2006.07. 003.

Moreira, L.M., de Souza, R.F., Almeida, N.F., Setubal, J.C., Oliveira, J.C., Furlan, L.R., Ferro, J.A. and da Silva, A.C. [2004] Comparative genomics analyses of citrus-associated bacteria. *Annu Rev Phytopathol.* **42** pp 163-84.

Mount, D.W. [2001] Bioinformatics: Sequence and Genome analysis. Cold Spring Harbor Laboratory, Pr.

Mulder, N.J., Rolf, A., Teresa, K. A., Amos, B., Alex, B., Binns, D., Paul, B., Peer, B., Phillip, B., Lorenzo, C., Richard, C., Emmanuel, C., Ujjwal, D., Richard, D., Fleischmann, W., Gough, J., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lonsdale, D., Lopez, R., Letunic, I, Madera, M., Maslen, J., McDowall, J., Mitchell, A., Nikolskaya, A.N., Orchard, S., Pagni, M, Ponting, C.P., Quevillon, E., Selengut, J., Sigrist, C.J.A, Silventoinen, V., Studholme, D.J., Vaughan, R. and Wu, C [2005] InterPro, progress and status in 2005. *Nucleic Acids Research*, **33**, Database issue D201-D205.

Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Barrell, D., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P., Bucher, P, Copley, R.R., Courcelle, E., Das, U., Durbin, R., Falquet, L., Fleischmann, W., Griffiths-Jones, S., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lopez, R., Letunic, I., Lonsdale, D., Silventoinen, V., Orchard, S.E., Pagni, M., Peyruc, D., Ponting, C.P., Selengut, J.D., Servant, F., Sigrist, C.J., Vaughan, R. and Zdobnov, E.M. [2003] The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res* **31** pp 315-318.

Murray, J.F. [2004] *Mycobacterium tuberculosis* and the Cause of Consumption From Discovery to Fact . *American Journal of Respiratory and Critical Care Medicine* Vol 169. pp. 1086-1088

Murray, M and Nardell, E. [2002] Molecular epidemiology of tuberculosis: achievements and challenges to current knowledge. *Bull World Health Organ*, **80** (6), pp 477-482.

Murray, M., Christopher, J.L. and Salomon, J.A. [1998] Modeling the impact of global tuberculosis control strategies *Medical Sciences* **95** (23) pp 13881-13886

Musser, J.M. [1995] Antimicrobial agent resistance in mycobacteria: molecular genetic insights *Clinical Microbiology Reviews* **8** (4) pp 496-514.

Nahid, P. and Daley, C. [2006] Prevention of tuberculosis in HIV-infected patients. *Curr Opin Infect Dis.* **19**, pp 189-193.

Nair, J., Rouse, D.A., Bai, G.H. and Morris, S.L. [1993] The *rpsL* gene and streptomycin resistance in single and multiple drug-resistant strains of *Mycobacterium tuberculosis*. *Mol. Microbiol.* **10** pp 521−527

Narain, J.P., Raviglione, M.C. and Kochi, A. [1992] HIV-associated tuberculosis in developing countries: epidemiology and strategies for prevention. *Tuberc Lung Dis,* **73** pp 311-21.

Nascimento, A.L., Ko, A.I., Martins, E.A., Monteiro-Vitorello, C.B., Ho, P.L., Haake, D.A., Verjovski-Almeida, S., Hartskeerl, R.A., Marques, M.V., Oliveira, M.C., Menck, C.F., Leite, L.C., Carrer, H., Coutinho, L.L., Degrave, W.M., Dellagostin, O.A., El-Dorry, H., Ferro, E.S., Ferro, M.I., Furlan, L.R., Gamberini, M., Giglioti, E.A., Goes-Neto, A., Goldman, G.H., Goldman, M.H., Harakava, R., Jeronimo, S.M., Junqueira-de-Azevedo, I.L., Kimura, E.T., Kuramae, E.E., Lemos, E.G., Marino, C.L., Nunes, L.R., de Oliveira, R.C., Pereira, G.G., Reis, M.S., Schriefer, A., Siqueira, W.J., Sommer, P., Tsai, S.M., Simpson, A.J., Ferro, J.A., Camargo, L.E., Kitajima, J.P., Setubal, J.C. and Van Sluys, M.A [2004] Comparative genomics of two *Leptospira interrogans serovars* reveals novel insights into physiology and pathogenesis. *J Bacteriol.* **186** (7) pp 2164-2172.

Nei, M. and Gojobori, T. [1986] Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3** pp 418-426.

Nei, M. and Miller, J.C. [1990] A simple method for estimating average number of nucleotide substitutions within and between populations from restriction data. *Genetics* **125** pp 873-879.

Nembaware, V., Crum, K., Kelso, J. and Seoighe, C. [2002] Impact of the Presence of Paralogs on Sequence Divergence in a Set of Mouse-Human Orthologs. *Genome Res.*. **12** pp 1370-1376

Nobrega, M. A. and Pennacchio, L.A. [2004] Comparative genomic analysis as a tool for biological discovery *J Physiol.* **554** pp 31-39.

Nunn, P., Williams, B., Floyd, K., Dye, C., Elzinga, G. and Raviglione, M. [2005] Tuberculosis control in the era of HIV. *Nat Rev Immunol.* **5**, pp 819-26.

Ormerod, L.P. [2005] Multidrug-resistant tuberculosis (MDR-TB): epidemiology, prevention and treatment. *British Medical Bulletin* **73-74** (1) pp 17-24

Overbeek, R., Fonstein,M., D'Souza, M., Pusch, G. D. and Maltsev, N. [1999] The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci.* USA, **96** pp 2896–2901.

Overbeek, R., Begley T., Butler R.M., Choudhuri, J.V., Chuang, H.Y., Cohoon, M., de Crecy-Lagard, V., Diaz, N., Disz, T., Edwards, R., Fonstein, M., Frank, E.D., Gerdes, S., Glass, E.M., Goodman, A., Hanson, A., Iwata-Reuyl, D., Jensen, R., Jamshidi, N., Krause, L., Kubal, M., Larsen, N., Linke, B., Chards, A.C., Meyer, F., Neuweger, H., Olsen, G., Olson, R., Osterman, A., Portnoy, V., Pusch, G.D., Rodionov, D.A., Ruckert, C., Steiner, J., Stevens, R., Thiele, I., Vassieva, O., Ye,Y., Zagnitko, O. and Vonstein, V. [2005] The Subsystems Approach to Genome Annotation and its use in the Project to Annotate 1000 Genomes. *Nucleic Acids Research* **33** (17) pp 5691-5702.

Pablos-Mendez, A., Raviglione, M. C., Laszlo, A., Binkin, N., Rieder, H.L., Bustreo, F., Cohn, D.L., Lambregts-van, C.S., Weezenbeek, S., Kim, J., Chaulet, P. and Nunn, P [1998] Global surveillance for antituberculosis-drug resistance, 1994 -1997. World Health Organization-International Union against Tuberculosis and Lung Disease Working Group on Anti-Tuberculosis Drug Resistance Surveillance. *N. Engl. J. Med.* **338**, pp 1641-1649.

Pace, N.R., Hugenholtz, P. and Goebel, B.M [1998] Impact of Culture-Independent Studies on the Emerging Phylogenetic View of Bacterial Diversity. *Journal of Bacteriology* **180** (18) pp. 4765-4774.

Page, R.D. and Charleston, M.A [1997] From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Mol. Phylogenet. Evol.* **7** pp 231–240

Park, M.M., Davis, A.L., Schluger, N.W., Cohen, H. and Rom, W.N [1996] Outcome of MDR-TB patients, 1983–1993: prolonged survival with appropriate therapy. *Am J. Respir. Crit. Care Med* **153** pp 317–24.

Patterson, C. [1988] Homology in classical and molecular biology. *Mol. Ecol. Evol.* **5** (6) pp 603-625.

Pearson, M.L., Jereb, J.A., Frieden, T.R. Crawford, J.T., Davis, B.J., Dooley, S.W., Jarvis, W.R. [1992] Nosocomial transmission of multidrug-resistant *Mycobacterium tuberculosis*: a risk to patients and health care workers. *Ann. Intern. Med.* **117** pp191-196.

Pelicic, V., Jackson, M., Reyrat, J.M., Jacobs, W.R., Gicquel, B. and Guilhot, C. [1997] Efficient allelic exchange and transposon mutagenesis in *M. tuberculosis*, *Proc Natl Acad Sci USA* **94** pp. 10955–10960.

Poulet, I. and Cole, S.T. [1995] Characterization of the highly abundant polymorphic GC-rich-repetitive sequence (PGRS) present in *Mycobacterium tuberculosis. Archives of Microbiology.* **163** (2) pp 87-95

Poux, C., van Rheede, T, Madsen, O., and de Jong, W. W [2002]. Sequence Gaps Join Mice and Men: Phylogenetic Evidence from Deletions in Two Proteins. *Molecular Biology and Evolution* **19** pp 2035-2037

Prentice, M.B [2004] Bacterial comparative genomics. *Genome Biol.* **5**(8) p 338 Epub.

Ragan, M.A. [2001] Detection of lateral gene transfer among microbial genomes. *Curr. Opin. Genet. Dev.* **11** (6) pp 620-6.

Raskin, D.M., Seshadri, R., Pukatzki, S.U. and Mekalanos, J.J.[2006] Bacterial Genomics and Pathogen Evolution. *Cell* **124** (4) pp 703-714

Raviglione, M.C., Snider, D. and Kochi, A [1995] Global epidemiology of tuberculosis: morbidity and mortality of a worldwide epidemic. *JAMA* **273** pp 220-226.

Remm, M., Storm, C.E. and Sonnhammer, E.L [2001] Automatic clustering of orthologs and inparalogs from pairwise species comparisons. *J. Mol.Biol* **314** pp 1041–52.

Rieder, H.L., Cauthen, G.M., Comstock, G.W. and Snider, D.E [1989] Epidemiology of tuberculosis in the United States. *Epidemiol Rev* **11** pp 79-98.

Riska, Paul F., Ya, S., Bardarov, S., Freundlich, L., Sarkis,G., Carrière, C., Kumar, V., Chan, J., and Jacobs, W.R [1999] Rapid Film-Based Determination of Antibiotic Susceptibilities of *Mycobacterium tuberculosis* strains by using a Luciferase Reporter Phage and the Bronx Box. *Journal of Clinical Microbiology*, **37**(4) pp 1144-1149.

Robinson, P.N., Wollstein, A., Bohme, U. and Beattie, B. [2004] Ontologising gene-expression micro-arrary data: characterising clusters with Gene Ontology. *Bioinformatics* **20** (6) pp 979-981.

Kimura, M. [1983] The neutral theory of Molecular Evolution. Cambridge University Press, Cambridge, Massachusetts.

Koonin, E.V. [2005] Orthologs, Paralogs, and Evolutionary Genomics. *Annual Review of Genetics* **39** pp 309-338

Koonin, E.V. and Galperin, M.Y. [2002] Sequence Evolution Function. *Computational Approaches in Comparative Genomics*. New York: Kluwer

Koonin, E.V., Makarova, K.S. and Aravind, L. [2001] Horizontal gene transfer in prokaryotes: quantification and classification. *Annu. Rev. Microbiol.* **55** pp 709–42

Koonin, E.V. [2001] An apology for orthologs - or brave new memes. *Genome Biol* **2** comment 1005.1-1005.2.

Kumar, S., Tamura, K., and Nei, M. [2004] MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Briefings in Bioinformatics* **5** pp150-163.

Kurup, S.H., and Chan, C.C. [2006] Mycobacterium-related Ocular Inflammatory Disease: Diagnosis and Management, *Ann Acad Med,* Singapore, **35** pp 203-207.

Lara-Tejero, M., Sutterwala, F.S., Ogura, Y., Grant, E.P., Bertin, J., Coyle, A.J., Flavell, R.A. and Galan, J.E. [2006]. Role of the caspase-1 inflammasome in *Salmonella typhimurium* pathogenesis. J *Exp Med.* **203** pp 1407-1412.

Li, L., Stoeckert, C.J. and Roos, D.S. [2003] OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13** pp 2178-2189.

Liolios, K., Nektarios, T., Philip, H. and Kyrpides, N.C. [2006] The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide, *Nucleic Acids Res.* **34** (Database issue): D332–D334.

Liu, J., Hegyi, H., Acton, T.B., Montelione, G.T. and Rost, B. [2004] Automatic target selection for structural genomics on eukaryotes. *Proteins: Structure, Function, and Bioinformatics* **56** (2) pp 188–200

Lonnroth, K., Thuong, L.M., Linh, P.D. and Diwan, V.K. [1999] Delay and discontinuity – a survey of TB patients' search of a diagnosis in a diversified health care system. *Int. J. Tuberc. Lung Dis.* **3** pp 992-1000.

Lowrie, D.B. [2006] DNA vaccines for therapy of tuberculosis: Where are we now?, *Vaccine* **24** (12), pp 1983-1989.

Luscombe, N.M., Greenbaum, D. and Gerstein, M. [2001] What is bioinformatics? A proposed definition and overview of the field. *Methods Inf. Med.* **40** (4) pp 346-58.

McDonald, J. H. and Kreitman, M. [1991] Adaptive protein evolution at the Adh locus in Drosophila. *Nature* 351: 652-654.

Mazumder, R., Natale, D.A., Murthy, S., Thiagarajan, R. and Wu, C.H. [2005] Computational identification of strain-, species- and genus-specific proteins. *BMC Bioinformatics* **6** pp 279. doi:10.1186/1471-2105-6-279

Rozas, J., Sánchez-Delbarrio, J. C., Messeguer, X. and Rozas, R. [2003] DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19** pp 2496-2497.

Ryan, F. [1992] The forgotten plague: how the battle against tuberculosis was won and lost. Boston, MA: Little, Brown. p. 342–364.

Saeed, A.I., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., Braisted, J., Klapa, M., Currier, T., Thiagarajan, M., Sturn, A., Snuffin, M., Rezantsev, A., Popov, D., Ryltsov, A., Kostukovich, E., Borisovsky, I., Liu, Z., Vinsavich, A., Trush, V. and Quackenbush, J. [2003] TM4: a free, open-source system for microarray data management and analysis. *Biotechniques*. **34** (2) pp 374-378.

Sakula A. Robert Koch: centenary of the discovery of the tubercle bacillus [1982]. *Thorax* **37** pp 246–251.

Sambandamurthy, V.K. and Jacobs, W.R Jr [2005] Live attenuated mutants of *Mycobacterium tuberculosis* as candidate vaccines against tuberculosis, *Microbes Infect.* **7** pp. 955–961.

Sarrel, M [2006] A history of tuberculosis. *Communicable Disease Service Tuberculosis Control Program* online http://www.state.nj.us/health/cd/tbhistry.htm

Scarselli, M., Giuliani, M.M., Adu-Bobie, J., Pizza, M. and Rappuoli, R [2005] The impact of genomics on vaccine design, *Trends in Biotechnology*, **23** (2) pp 84-91.

Schleifer, K.H [2004] Microbial diversity: facts, problems and prospects. *Syst. Appl. Microbiol.* **27** pp 3-9.

Schleifer, K.H. and Ludwig, W [1999] Phylogeny of bacteria. In: Microbial Evolution and Infection (Eds. Göbel, B. and R. Ruf), Einhorn-Presse Verlag GmbH, Reinbek. pp. 94- 100.

Schultz, C.J., Rumsewicz, M. P., Johnson, K.L., Jones, B.J., Gaspar, Y.M. and Bacic, A. [2002] Using genomic resources to guide research directions. The arabinogalactan protein gene family as a test case, *Plant Physiol.* **129**, pp. 1448–1463.

Scorpio, A. and Zhang, Y [1996] Mutations in pnc A, a gene encoding pyrazinamidase/ nicotinamidase, cause resistance to the bacillus. *Nat. Med.* **2** pp 662-667.

Sekiguchi, J.., Miyoshi-Akiyama, T., Augustynowicz-Kopec, E., Zwolska, Z., Kirikae, F., Toyota, E., Kobayashi, I., Morita, K., Kudo, K., Kato, S., Kuratsuji, T., Mori, T. and Kirikae, T. [2007]. Detection of Multidrug Resistance in *Mycobacterium tuberculosis. J. Clin. Microbiol.* **45** pp 179-192.

Selwyn, P.A., Sckell, B.M., Alcabes, P., Friedland, G. H., Klein, R. S., and Schoenbaum, E. E. [1992] High risk of active tuberculosis in HIV-infected drug users with cutaneous anergy. *JAMA* **268** (4) pp 504-509.

Selwyn, P. A., Hartel, D., Lewis, V. A., Schoenbaum, E. E., Vermund, S. H., Klein, R. S., Walker, A. T. and Friedland, G. H [1989]. A prospective study of the risk of tuberculosis among intravenous drug users with human immunodeficiency virus infection. *N. Engl. J. Med.* **320** pp 545-550.

Shafer, R.W. and Edlin, B.R [1996] Tuberculosis in patients infected with human immunodeficiency virus: perspective on the past decade. *Clin Infect Dis.* **22**, pp 683-704.

Shah, I. and Hunter, L. [1997] Predicting enzyme function from sequence: a systematic appraisal. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **5** pp 276–283

Shen, Y., Ling, S., Prabhat, S., Dan, H., Liyou, Q., George, D., Norman, L. L., and Chen, Z.W [2004] Clinical Latency and Reactivation of AIDS-Related Mycobacterial Infections. *Journal of Virology* 78 (24), pp 14023-14032.

Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D. and Church, G.M. [2005] Accurate multiplex polony sequencing of an evolved bacterial genome, *Science* **309** pp. 1728–1732

Siew, N., Azaria, Y. and Fischer, D [2004] The ORFanage: an ORFan database. *Nucleic Acids Res* **32** Database issue:D281-283.

Sigrist, C.J.A., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A. and Bucher, P. [2002] PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief. Bioinform.* **3** (3) pp 265–274.

Skeiky, A. W. and Sadoff, J.C [2006] Advances in tuberculosis vaccine strategies. *Nat Rev Microbiol.,* **4** pp 469-476.

Small, P. M.,. Shafer, R. W, Hopewell, P. C., Singh, S. P., Murphy, M. J., Desmond, E., Sierra, M. F. and Schoolnik, G. K. [1993] Exogenous reinfection with multidrug-resistant *Mycobacterium tuberculosis* in patients with advanced HIV infection. *N. Engl. J. Med.* **328** pp1137-1144.

Smith, I. [2003] *Mycobacterium tuberculosis* Pathogenesis and Molecular Determinants of Virulence, *Clinical Microbiology Reviews*, **16** (3) pp. 463–496.

Smolinski, S. M., Hamburg, M.A. and Lederberg, J. Eds [2003], Committee on Emerging Microbial Threats to Health in the 21st Century. Microbial Threats to Health: Emergence,Detection, and Response. Online at http://www.nap.edu/catalog/ 10636.html

Sonnhammer, E.L., Remm, M. and Storm, C.E. [2001] Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* **314** (5) pp 1041-1052.

Spratt, B.G. [1996] Antibiotic resistance: counting the cost. *Curr. Biol.* **6** pp 1219−1221.

Sreevatsan, S., Pan, X., Stockbauer, K.E., Connell, N.D., Kreiswirth, B.N., Whittam, T.S. and Musser, J.M [1997] Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. *Proc. Natl. Acad. Sci. USA* **94** pp 9869 - 9874.

Sterling, T.R., Bethel, J., Goldberg, S., Weinfurter, P., Yun, L. and Horsburgh, C.R. [2006] The scope and impact of treatment of latent tuberculosis infection in the United States and Canada. *American Journal of Respiratory and Critical Care Medicine*; **173** pp 927-931.

Storm, C.E. and Sonnhammer, E.L. [2002] Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics* **18** pp 92–99

Sturino, J.M. and Klaenhammer, T.R [2006] Engineered bacteriophage-defence systems in bioprocessing, *Nat. Rev. Microbiol.* **4** pp. 395–404.

Styblo, K. [1985] The relationship between the risk of tuberculous infection and the risk of developing infectious tuberculosis. *Bulletin of the International Union Against Tuberculosis*, **60** pp117-119.

Sudre, P., ten Dam, G. and Kochi, A. [1992] Tuberculosis: a global overview of the situation today. *Bull World Health Organ.* **70** pp 149-159.

Supply, P., Mazars, E., Lesjean, S., Vincent, V., Gicquel, B., and Locht, C. [2000] Variable human minisatellite-like regions in the *Mycobacterium tuberculosis* genome. *Mol. Microbiol.* **36** pp 762-771

Tanaka, M.M. [2004] Evidence for positive selection on *Mycobacterium tuberculosis* within patients. *BMC Evol Biol.* 4 p 31

Tatusov, R.L., Koonin, E.V. and Lipman, D.J [1997] A Genomic Perspective on Protein Families Science 24. **278** (5338) pp. 631 – 637.

Tatusov, R.L., Mushegian, A.R., Bork, P., Brown, N.P., Hayes, W.S., Borodovsky, M., Rudd, K.E. and Koonin, E.V. [1996] Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*. *Curr. Biol.* **6** pp 279–291.

Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D. and Koonin, E.V [2001] The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.,* **29** pp 22–28

Tekaia, F., Gordon, S.V., Garnier, T., Brosch, R., Barrell, B.G. and Cole, S.T. [1999] Analysis of the proteome of *Mycobacterium tuberculosis* in silico. *Tubercle Lung Disease* **79** pp 329–342.

Tekaia, F. and Yeramian, E [2005] Genome trees from conservation profiles. *PLoS Comput Biol* **1** (7) p75.

The Gene Ontology Consortium [2001], Creating the gene ontology resource: design and implementation, *Genome Res.,* **11**,pp. 1425–1433.

TØnjum, T., Welty, D. B., Jantzen, E. and Small, P. L [1998]. Differentiation of *M. ulcerans, M. marinum,* and *M. haemophilum*: Mapping of their relationships to *M. tuberculosis* by fatty acid profile analysis, DNA-DNA hybridization, and 16S rRNA gene sequenceanalysis. *J Clin Microbiol* **36** pp 918-925.

Ussery, X.T., Bierman, J.A., Valway, S.E., Seitz, T.A., DiFerdinando, G.T. Jr. and Ostroff, S.M. [1995] Transmission of multidrug-resistant *Mycobacterium tuberculosis* among persons exposed in a medical examiner's office, New York. *Infect Control Hosp Epidemiol* **16** pp 160-165.

van den Braak, N., Simons, G., Gorkink, R., Reijans, M., Eadie, K., Kremers, K., van Soolingen, D., Savelkoul, P., Verbrugh, H., and van Belkum, A. [2004] A new high-throughput AFLP approach for identification of new genetic polymorphism in the genome of the clonal microorganism *Mycobacterium tuberculosis*, *Journal of Microbiological Methods*, **56**(1) pp 49-62.

Wade, M.M., Volokhov, D., Peredelchuk, M., Chizhikov, V. and Zhang, Y [2004] Accurate mapping of mutations of pyrazinamide-resistant *Mycobacterium tuberculosis* strains with a scanning-frame oligonucleotide microarray, *Diagnostic Microbiology and Infectious Disease*, **49** (2) pp 89-97.

Ward, N. and Fraser, C.M [2005] How genomics has affected the concept of microbiology. *Current Opinion in Microbiology* **8** (5) pp 564-571.

Weinstock, G.M., Smajs, D., Hardham, J. and Norris, S.J [2000] From microbial genome sequence to applications, *Research in Microbiology*, **151** (2) pp 151-158.

Wilson, J. W., Schurr, M, J., LeBlanc, C. R., Ramamurthy, R., Buchanan, K .L. and Nickerson, C.A [2002] Mechanisms of bacterial pathogenicity. *Postgrad Med. J.* **78** pp 216–224.

Woese, C. R. [1987] Bacterial evolution. *Microbiol. Rev.* **51** pp 221-271.

Wolf, Y.I., Rogozin, I.B, Grishin, N.V., Tatusov, R.L. and Koonin, E.V [2001]. Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evolutionary Biology* **1** p 8.

Wolf, Y.I., Rogozin, I.B., Grishin, N.V. and Koonin, E.V [2002] Genome trees and the tree of life. *Trends Genet.* **18** pp 472–79.

World Health Organization [1998] The world health report 1998 - Life in the 21st century: A vision for all. *WHO* Report. Online http://www.who.int/whr/previous/en/index.html

World Health Organization [2002] Global tuberculosis control: surveillance, planning, finances. *WHO Report* (WHO/CDC/TB/2003. 316) Geneva. Available online http://www.who.int/whr/previous/en/index.html.

World Health Organization [2003] Global tuberculosis control: surveillance, planning, finances. WHO Report 2003 (WHO/CDC/TB/2003. 316). Geneva

World Health Organization [2004] A global emergency: a combined response. *WHO Report.* Available online http://www.who.int/whr/previous/en/index.html.

World Health Organization [2006]. *Tuberculosis Fact Sheet No 104* - Global and regional incidence. Available online http://www.who.int/whr/en.

Yanay, O., Marco, P., Reinhard, S. and Burkhard, R. [2005] Beyond annotation transfer novel protein functional prediction methods to assist drug discovering DDT **10** (21)

Yang, J.K., Yoon, H-.J., Ahn, H.J., Lee, B.I., Cho, S.H., Waldo, G.S., Park, M.S. and Suh, S.W. [2002] Crystallization and preliminary X-ray crystallographic analysis of the Rv2002 gene product from *Mycobacterium tuberculosis*, a beta-ketoacyl carrier protein reductase homologue. *Acta Crystallogr* **D58** pp. 303 - 305

Yuri, I.W, Rogozin, I.B., Grishin,N.V., Tatusov, R.L. and Koonin, E.V [2001] Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evolutionary Biology* **1** p 8 doi:10.1186/1471-2148-1-8.

Zmasek, C.M. and Eddy, S.R [2001] A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics* **17** (9) pp 821-828

Zmasek, C.M. and Eddy, S.R [2002] Analyzing proteomes by automated phylogenomics using re-sampled inference of orthologs. *BMC Bioinformatics* **3** p 14.

APPENDIX

Table A1-a: The result of known virulence bacterial genes 1.0S

Organisms Code	Cluster Number																		
	14	29	131	341	366	522	1052	2470	4241	5166	6264	8270	10321	10631	10744	11052	16892	30401	36375
MYCTU	2	2	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MYCLEH	2	2	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MYCBO	2	2	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MYCPA	2	2	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MYCLE	2	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CORDI	2	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PROAC	2	2	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TROWT	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TROW8	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CLOPE	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CLOTE	1	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BACAN	2	1	1	0	1	0	1	0	0	0	0	1	1	0	0	0	0	0	0
BACC1	2	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
STAAM	1	1	1	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
STAES	1	1	1	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
LISIN	1	1	1	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0
LISMO	1	1	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
MYCGE	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MYCPN	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MYCPE	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
STRR6	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
STRP1	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
STRA5	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
STRMU	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
ENTFA	1	1	1	1	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0
THEMA	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CHLTR	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CHLMU	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CHLPN	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CHLCV	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ECO57	1	1	1	1	1	1	1	0	0	1	1	0	0	1	1	0	0	0	0
YERPE	1	1	1	0	1	1	0	0	0	1	0	0	0	0	0	0	0	1	0
YERPS	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
SALTY	1	1	1	1	1	1	1	0	0	1	1	0	0	1	1	0	0	0	1
SALTI	1	1	1	1	1	1	1	0	0	1	1	0	0	0	0	0	0	0	0
SALCH	1	1	1	1	1	1	1	0	0	1	1	0	0	1	1	0	0	0	0
SHIFL	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
HAEIN	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
HAEDU	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
PASMU	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
VIBCH	1	1	1	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
VIBPA	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
VIBUY	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0

Table A-1a Contd.

Organisms Code	Cluster Number																		
	14	29	131	341	366	522	1052	2470	4241	5166	6264	8270	10321	10631	10744	11052	16892	30401	36375
METCA	2	1	1	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0
NEIMBE	1	1	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
NEIG1	1	1	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
BORPA	1	1	1	1	1	1	0	1	1	0	0	0	0	0	0	1	1	0	0
BORPE	1	1	1	1	1	1	0	1	1	0	0	0	0	0	0	2	1	0	0
BURMA	1	1	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0
CHRVO	1	1	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0
RICPR	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RICCN	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BRUSU	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BRUME	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CAMJE	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HELPJ	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AQUAE	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DEIRA	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CORGL	2	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
COREF	2	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
STRCO	5	3	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
STRAW	5	3	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BIFLO	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CLOAB	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BACSU	1	1	0	0	0	0	2	0	0	0	0	1	1	0	0	0	0	0	0
BACHD	2	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OCEIH	1	1	0	1	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0
LACJO	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LACPL	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LACLA	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
THETN	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PSEAE	1	1	1	2	1	1	0	1	3	0	0	0	0	0	0	0	0	0	0
PSEPK	1	1	1	3	0	1	0	1	2	0	0	0	0	0	0	0	0	0	0
BUCAP	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
BUCAI	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
ECOLI	1	1	1	1	1	1	1	0	0	1	1	0	0	1	1	0	0	0	0
WIGBR	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
NITEU	1	1	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
CAUCR	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BRAJA	2	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DESVH	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GEOSL	1	2	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
WOLSU	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SILPO	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table A1b: The result of known virulence bacterial genes 1.0S

Cluster Number	Total Protein in pathogens	AVE P	Number of Pathogens	Present In TB?	Virulence TB?	Virulence in others?	Total Protein in non pathogens	AVE N	Number of non pathogens	Total genomes	TP in cluster
14	1117	19.9464	54	y	y in mtb	y in SALTY*	959	34.25	26	80	2076
29	617	11.0179	47	y		y in SALTY*	506	18.071	25	72	1123
131	271	4.839286	42	y		y in Vibch*	209	7.4643	22	64	480
341	128	2.28571	55	y		y in LISMO*	67	2.3929	25	80	195
366	76	1.35714	53	y		y in LISMO*	46	1.6429	28	81	122
522	72	1.28571	51	y		y in SHIFL	39	1.3929	27	78	111
1052	73	1.30357	39	n		y in STRR6	29	1.0357	19	58	102
2470	51	0.91071	51	y		y in PASMU	27	0.9643	27	78	78
4241	34	0.60714	21	y		y in BACAN	27	0.9643	15	37	61
5166	34	0.60714	29	y		y in PSEAE	16	0.5714	14	43	50
6264	24	0.42857	22	y		y in ENTFA	11	0.3929	11	34	35
8270	271	4.83929	42	y		y in VBCH*	209	7.4643	22	64	480
10321	31	0.55357	6	n		y in YERPS	1	0.0357	1	7	32
10631	4	0.07143	3	n		y in BORPE	7	0.25	6	9	11
10744	4	0.07143	3	n		y in ECO57	2	0.0714	2	5	6
11052	4	0.07143	3	n		y in ECO57	2	0.0714	2	5	6
16892	2	0.03571	2	n		y in BACAN	2	0.0714	2	4	4
30401	2	0.03571	2	n		y in BACAN	2	0.0714	2	4	4
36375	3	0.05357	3	n		y in ECO57	1	0.0357	1	4	4

Table A1c: The result of known virulence bacterial genes 1.0S

Cluster No	Description line	INTERPRO					
14	ATP-dependent_Clp_protease_proteolytic_subunit_ (EC_3.4.21.92)_(Endopeptidase_Clp)	IPR001907					
29	RNA polymerase sigma factor rpoD (Sigma-42)	IPR011991	IPR007624	IPR007627	IPR007630	IPR000943	PR012760
131	Chaperone_clpB	IPR003959	IPR004176	IPR001270	IPR003593		
341	Aldehyde_dehydrogenase_(EC_1.2.1.3)	IPR002086	IPR012303	P23240			
366	Potassium-transporting_ATPase_B_chain_(EC_3.6.3.12)_(ATP_phosphohydrolase_potassium-transporting_B_chain)_(Potassium_binding_and_translocating_subunit_B	IPR008250	IPR005834	IPR001757	IPR006391		
522	ATP_synthase_gamma_chain_(EC_3.6.3.14)_(ATP_synthase_F1_sector_gamma_subunit	IPR000131					
1052	Zinc-binding_lipoprotein_adcA_precursor	IPR006127	IPR006128	IPR006129			
2470	Phosphomannomutase/phosphoglucomutase_(EC_5.4.2.8)_(EC_5.4.2.2)_(PMM_/_PGM)	IPR005845	IPR005841	IPR005843	IPR005846	IPR005844	
4241	Virulence_factors_putative_positive_transcription_regulator_bvgA	IPR011991	IPR001789	IPR000792	IPR011006		
5166	Transcriptional_regulatory_protein_basR/pmrA	IPR001789	IPR001867	IPR011006			
6264	Sensor_protein_basS/pmrB_(EC_2.7.3.-)	IPR003661	IPR003660	IPR003594	IPR004358	IPR005467	IPR009082
8270	Capsule_biosynthesis_protein_capB	IPR008337					
10321	Capsule_biosynthesis_protein_capC	IPR008338					
10631	Cytotoxic_protein_ccdB_(Protein_letB)_(Protein_G)_(LynB)	IPR002712	IPR011067				
10744	Protein_ccdA_(Protein_letA)_(Protein_H)_(LynA)	IPR009956					
11052	Probable_tonB-dependent_receptor_bfrD_precursor_(Virulence-associated_outer_membrane_protein_Vir-90)	IPR000531	IPR012910	IPR010105			
16892	Virulence_sensor_protein_bvgS_precursor_(EC_2.7.3.-)	IPR001789 IPR004358 IPR005467	IPR001638	IPR003661	IPR000014	IPR008207	IPR003594
30401	Attachment_invasion_locus_protein_precursor	IPR000758	IPR000700	IPR001311	IPR009082	IPR011006	
36375	Putative_surface-exposed_virulence_protein_bigA_precursor	IPR005546					

Table A-2a: The result of Known virulence bacterial genes 0.5 S

Organisms Code	Cluster Number																				
	2	5	14	60	144	176	201	435	675	814	1116	1236	1237	2985	5484	5485	7143	8282	8629	12119	25593
MYCTU	17	7	10	2	2	2	0	1	2	1	1	0	0	0	0	0	0	0	0	0	0
MYCLEH	17	7	10	2	2	2	0	1	2	1	1	0	0	0	0	0	0	0	0	0	0
MYCBO	17	7	10	2	2	2	0	1	2	1	1	0	0	0	0	0	0	0	0	0	0
MYCPA	21	10	22	2	2	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0
MYCLE	6	4	2	2	2	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
CORDI	13	4	4	2	2	0	3	1	0	1	0	0	0	0	0	0	0	0	0	0	0
PROAC	15	2	2	2	2	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
TROWT	3	1	0	2	2	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0
TROW8	3	1	0	2	2	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0
CLOPE	25	14	2	2	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
CLOTE	36	20	1	3	2	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0
BACAN	49	38	9	2	3	1	3	1	1	0	1	2	0	0	1	0	1	1	0	0	1
BACC1	50	38	10	2	3	1	2	1	0	0	1	2	0	0	1	0	0	0	0	0	0
STAAM	19	6	5	3	2	1	2	1	1	0	2	0	0	0	0	0	0	0	0	0	0
STAES	18	6	6	2	1	1	3	1	2	0	0	0	0	0	0	0	1	1	0	0	0
LISIN	25	9	2	3	2	1	3	1	0	1	2	0	0	0	0	0	0	0	0	0	0
LISMO	25	10	2	3	2	1	3	1	0	1	1	0	0	0	0	0	0	0	0	0	0
MYCGE	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
MYCPN	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
MYCPE	1	0	2	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
STRR6	13	4	1	3	1	2	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0
STRP1	11	3	2	3	1	2	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0
STRA5	21	8	2	3	1	2	4	1	0	0	0	0	0	0	2	0	0	0	0	0	0
STRMU	15	3	2	4	1	2	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0
ENTFA	21	7	1	3	1	2	4	1	0	1	1	0	0	0	1	0	0	0	0	0	0
THEMA	16	4	0	2	1	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
CHLTR	2	0	0	2	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CHLMU	2	0	0	2	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CHLPN	2	0	0	2	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CHLCV	2	0	0	2	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ECO57	46	27	8	3	2	2	2	1	1	2	1	1	10	1	0	2	0	0	1	0	0
YERPE	31	16	6	7	1	1	2	1	1	2	1	0	4	0	0	0	0	0	0	0	0
YERPS	4	3	1	0	1	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0
SALTY	41	24	9	3	1	2	3	1	1	1	1	2	7	0	0	1	0	0	1	1	0
SALTI	41	25	6	3	1	2	3	1	1	1	1	2	4	0	0	0	0	0	0	0	0
SALCH	42	25	6	2	1	2	3	1	1	2	1	2	5	0	0	1	0	0	1	0	0
SHIFL	5	3	0	1	1	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
HAEIN	6	3	0	1	1	1	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0
HAEDU	4	1	0	1	1	1	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0
PASMU	9	5	2	1	1	1	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0
VIBCH	46	28	5	3	1	1	1	1	2	1	0	0	0	0	0	0	0	0	0	0	0
VIBPA	54	41	10	4	1	2	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
VIBUY	66	47	10	3	1	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
METCA	15	19	1	2	2	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0
NEIMBE	4	2	3	2	1	2	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
NEIG1	3	2	2	2	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0
BORPA	31	19	19	3	1	1	1	1	5	2	1	2	0	1	0	0	0	0	0	0	0
BORPE	27	11	12	2	1	1	1	1	3	2	1	0	0	2	0	0	0	0	0	0	0
BURMA	40	24	19	5	1	2	0	1	3	1	1	0	0	0	0	0	0	0	0	0	0
CHRVO	64	45	9	3	1	2	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
RICPR	3	3	0	1	1	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
RICCN	5	3	1	1	1	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
BRUSU	24	14	16	2	1	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
BRUME	27	14	18	2	1	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
CAMJE	9	0	1	2	1	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0
HELPJ	5	0	0	2	1	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0
AQUAE	8	8	1	2	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0

Table A-2a contd.

Organisms Code	Cluster Number																				
	2	5	14	60	144	176	201	435	675	814	1116	1236	1237	2985	5484	5485	7143	8282	8629	12119	25593
MYCTU	17	7	10	2	2	2	0	1	2	1	1	0	0	0	0	0	0	0	0	0	0
MYCLEH	17	7	10	2	2	2	0	1	2	1	1	0	0	0	0	0	0	0	0	0	0
MYCBO	17	7	10	2	2	2	0	1	2	1	1	0	0	0	0	0	0	0	0	0	0
MYCPA	21	10	22	2	2	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0
DEIRA	33	7	7	3	1	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0
CORGL	15	6	8	2	2	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0
COREF	15	5	5	2	2	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0
STRCO	92	19	19	4	5	4	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
STRAW	78	20	20	4	5	5	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
BIFLO	11	5	1	2	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
CLOAB	49	22	1	4	2	2	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0
BACSU	37	11	11	2	1	1	3	1	2	0	0	0	0	0	1	0	1	1	0	0	0
BACHD	43	13	13	2	2	1	3	1	1	0	0	0	0	0	0	0	0	0	0	0	0
OCEIH	26	8	15	2	1	1	3	1	3	0	0	0	0	0	0	0	1	1	0	0	0
LACJO	9	5	1	3	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
LACPL	13	6	2	4	1	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
LACLA	10	4	0	3	2	1	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0
THETN	27	9	0	1	1	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0
PSEAE	67	52	19	6	2	2	1	1	3	1	1	3	0	2	0	0	0	0	0	0	0
PSEPK	67	62	26	4	2	1	1	1	2	2	0	5	0	1	0	0	0	0	0	0	0
BUCAP	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
BUCAI	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
ECOLI	44	28	9	2	1	1	2	1	1	1	1	1	1	1	0	1	0	0	1	1	0
WIGBR	1	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
NITEU	12	12	1	2	1	1	0	1	1	1	0	1	0	1	0	1	0	0	0	0	0
CAUCR	44	42	13	2	1	1	0	1	3	1	1	0	0	1	0	0	0	0	0	0	0
BRAJA	92	62	15	3	2	2	1	1	4	2	1	5	0	1	0	0	0	0	0	0	0
DESVH	47	30	2	2	2	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0
GEOSL	59	41	3	2	1	1	1	1	0	0	1	2	0	0	0	0	0	0	0	0	0
WOLSU	37	17	0	2	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0
SILPO	23	12	17	2	1	2	2	1	2	1	0	0	0	0	0	0	0	0	0	0	0

Table A-2b: The result of Known virulence bacterial genes 0.5 S

Cluster No.	Description line	InterPro
2	Virulence_factors_putative_positive_transcription_regulator_bvgA	IPR011991 IPR001789 IPR001867 IPR005829 IPR011006
5	Sensor_protein_basS/pmrB_(EC_2.7.3.-)	IPR003661 IPR003660 IPR003594 IPR004358 IPR005467 IPR009082
14	Aldehyde_dehydrogenase_(EC_1.2.1.3)	IPR002086 IPR012303
60	Chaperone_clpB	IPR004176 IPR003959 IPR003593 IPR013093 IPR001270
144	ATP-dependent_Clp_protease_proteolytic_subunit_(EC_3.4.21.92)_(Endopeptidase_Clp)	IPR001907
176	Acyl_carrier_protein_(ACP)	IPR009081 IPR006163 IPR003231
201	Zinc-binding_lipoprotein_adcA_precursor	IPR006127 IPR006128 IPR006129
435	ATP_synthase_gamma_chain_(EC_3.6.3.14)_(ATP_synthase_F1_sector_gamma_subunit)	IPR000131
675	Capsule_biosynthesis_protein_capD	IPR000101
814	Phosphomannomutase/phosphoglucomutase_(EC_5.4.2.8)_(EC_5.4.2.2)_(PMM_/_PGM)	IPR005845 IPR005841 IPR005843 IPR005846 IPR005844
1116	Potassium-transporting_ATPase_B_chain_(EC_3.6.3.12)_(Potassium-translocating_ATPase_B_chain)_(ATP_phosphohydrolase_potassium-transporting_B_chain)	IPR008250 IPR005834 IPR001757 IPR006391
1236		IPR000758
1237	Attachment_invasion_locus_protein_precursor	IPR000531 IPR012910 IPR010105
2985	Probable_tonB-dependent_receptor_bfrD_precursor_(Virulence-associated_outer_membrane_protein_Vir-90)	
5484		IPR002712 IPR011067
5485	Cytotoxic_protein_ccdB_(Protein_letB)_(Protein_G)_(LynB)	IPR008337
7143	Capsule_biosynthesis_protein_capB	IPR008338
8282	Capsule_biosynthesis_protein_capC	IPR009956
8629	Protein_ccdA_(Protein_letA)_(Protein_H)_(LynA)	IPR005546
12119	Putative_surface-exposed_virulence_protein_bigA_precursor	IPR011991 IPR011608
25593	Anthrax_toxin_expression_trans-acting_positive_regulator	

Table A-3a: The result of Known virulence *Mycobacterium tuberculosis* genes 1.0 S

Organisms Code	Cluster Number																		
	29	341	366	652	1583	2441	2613	2818	2820	2822	2823	2931	3332	7120	8232	8234	8935	11724	12944
MYCTU	2	1	1	3	1	2	1	2	2	4	2	2	2	1	1	1	1	1	1
MYCLEH	2	1	1	3	1	2	1	2	2	4	2	2	2	1	1	1	1	1	1
MYCBO	2	1	1	3	1	2	1	2	2	1	2	2	1	1	1	1	1	0	1
MYCPA	2	1	1	3	1	3	1	2	2	0	2	2	2	1	1	1	1	1	0
MYCLE	2	0	0	0	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0
CORDI	2	1	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
PROAC	2	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TROWT	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TROW8	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CLOPE	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CLOTE	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BACAN	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BACC1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
STAAM	1	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
STAES	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LISIN	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LISMO	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MYCGE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MYCPN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MYCPE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
STRR6	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
STRP1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
STRA5	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
STRMU	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENTFA	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
THEMA	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CHLTR	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CHLMU	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CHLPN	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CHLCV	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ECO57	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
YERPE	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
YERPS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SALTY	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SALTI	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SALCH	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HAEIN	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HAEDU	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PASMU	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table A-3a contd

Organisms Code	Cluster No																		
	29	341	366	652	1583	2441	2613	2818	2820	2822	2823	2931	3332	7120	8232	8234	8935	11724	12944
VIBCH	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VIBPA	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VIBUY	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
METCA	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NEIMBE	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NEIG1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BORPA	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BORPE	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RICPR	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RICCN	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BRUSU	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BRUME	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CAMJE	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HELPJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AQUAE	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DEIRA	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CORGL	2	1	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
STRCO	3	1	1	1	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0
STRAW	3	1	1	4	3	0	1	0	0	0	0	0	0	0	0	0	0	0	0
BIFLO	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CLOAB	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BACSU	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BACHD	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OCEIH	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LACJO	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LACPL	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LACLA	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BUCAP	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BUCAI	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ECOLI	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
WIGBR	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NITEU	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CAUCR	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BRAJA	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DESVH	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GEOSL	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
WOLSU	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SILPO	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table A-3b: The result of Known virulence *Mycobacterium tuberculosis* genes 1.0 S

Cluster Number	Total Protein in pathogen	Number of Pathogen	Present inTB?	Virulence TB?	Virulence others?	Total Protein in non-pathogens	Number of non pathogens	Total genomes	Total Protein in Cluster
29	57	49	Y	Y mtb	Y in ENTIFA	33	26	75	90
341	22	21	Y	Y mtb	Y in VIBCH	15	12	34	37
366	24	22	Y	Y mtb	Y in ENTIFA	11	11	34	35
652	16	8	Y	Y mtb		8	5	13	24
1583	7	7	Y	Y mtb		7	4	11	14
2441	10	5	Y	Y mtbc		0	0	5	10
2613	6	6	Y	Y mtb		4	4	10	10
2818	9	5	Y	Y mtbc		0	0	5	9
2820	9	5	Y	Y mtbc		0	0	5	9
2822	9	3	Y	Y mtbc		0	0	3	9
2823	9	5	Y	Y mtbc		0	0	5	9
2931	9	5	Y	Y mtbc		0	0	5	9
3332	8	5	Y	Y mtbc		0	0	5	8
7120	5	5	Y	Y mtbc		0	0	5	5
8232	4	4	Y	Y mtbc		0	0	4	4
8234	4	4	Y	Y mtbc		0	0	4	4
8935	4	4	Y	Y mtbc		0	0	5	5
11724	3	3	Y	Y mtbc		0	0	3	3
12944	3	3	Y	Y mtbc		0	0	3	3

Table A-3c: The result of Known virulence *Mycobacterium tuberculosis* genes 1.0 S

Cluster No.	Description line	InterPro					
29	RNA_polymerase_sigma_factor_rpoD_(Sigma-A)	IPR011991	IPR007624	IPR007627	IPR007630	IPR000943	IPR012760
341	Aldehyde_dehydrogenase_family_protein/Hypothetical_protein	IPR002086	IPR012303				
366	Potassium-transporting_ATPase_B_chain_(EC_3.6.3.12)	IPR008250	IPR005834	IPR001757	IPR006391		
652	DNA-binding_response_regulator /Possible 2 components system response transcriptional positive regulator PHOP	IPR001789	IPR001867	IPR011991	IPR005829	IPR011006	
1583	Mycobacterial persistence regulator MRPA P (2 components response transcriptional regiulatory protein (DNA-binding_response_regulator	IPR001789	IPR001867	IPR005829	IPR011006		
2441	Virulence_factor/MCE-FAMILY_PROTEIN	IPR003399	IPR005693				
2613	Cytotoxin_/haemolysin_homologue_(CYTOTOXIN|HAEMOLYSIN_HOMOLOGUE_TLYA)_(Cytotoxin/hemolysin)	IPR002942	IPR002877	IPR004538			
2818	Virulence_factor_mce_family_protein	IPR003399	IPR005693				
2820	MCE-FAMILY_PROTEIN_MCE2D_(Virulence_factor_mce_family_protein)	IPR003399	IPR005693				
2822	Phospholipase_C_1_precursor_(EC_3.1.4.3)_(MTP40_antigen)	IPR007312	IPR006311				
2823	Virulence_factor_mce_family_protein	IPR003399	IPR008360	IPR005693			
2931	MCE-FAMILY_PROTEIN_MCE1B_(Virulence_factor_mce_family_protein)	IPR003399	IPR005693				
3332	Virulence_factor_mce_family_protein	IPR003399	IPR005693				
7120	Heparin-binding_hemagglutinin_(Adhesin)	PR000897					
8232	Sulfatase family protein	IPR000917					
8234	Virulence_factor_mce_family_protein	IPR003399	IPR005693				
8935	Exported_repetitive_protein_precursor_(Cell_surface_protein_pirG)_(EXP53)	IPR008165					
11724	Virulence_factor_mce_family_protein	IPR003399	IPR005693				
12944	VIRULENCE-REGULATING_TRANSCRIPTIONAL_REGULATOR_VIRS_(ARAC/XYLS_FAMILY)	IPR012287	IPR000005	IPR009057			

Table A-4a: The result of Known virulence *Mycobacterium tuberculosis* genes 0.5 S

Organisms Code	Cluster Number												
	2	49	887	1679	1681	1699	1748	1823	1883	1905	5060	5995	6392
MYCTU	17	3	1	4	4	4	4	4	4	4	2	1	1
MYCLEH	17	3	1	4	4	4	4	4	4	4	2	1	1
MYCBO	17	3	1	3	3	3	3	3	1	2	2	1	1
MYCPA	21	4	1	8	8	8	7	6	0	6	0	1	1
MYCLE	6	2	1	1	1	1	1	1	0	1	0	1	1
CORDI	13	2	1	0	0	0	0	0	0	0	0	0	0
PROAC	15	2	1	0	0	0	0	0	0	0	0	0	0
TROWT	3	1	0	0	0	0	0	0	0	0	0	0	0
TROW8	3	1	0	0	0	0	0	0	0	0	0	0	0
CLOPE	25	5	1	0	0	0	0	0	0	0	0	0	0
CLOTE	36	6	1	0	0	0	0	0	0	0	0	0	0
MYCPE	1	1	1	0	0	0	0	0	0	0	0	0	0
STRR6	13	1	1	0	0	0	0	0	0	0	0	0	0
STRP1	11	1	1	0	0	0	0	0	0	0	0	0	0
STRA5	21	1	1	0	0	0	0	0	0	0	0	0	0
STRMU	15	1	1	0	0	0	0	0	0	0	0	0	0
ENTFA	21	1	1	0	0	0	0	0	0	0	0	0	0
THEMA	16	1	1	0	0	0	0	0	0	0	0	0	0
CHLTR	2	1	0	0	0	0	0	0	0	0	0	0	0
CHLMU	2	1	0	0	0	0	0	0	0	0	0	0	0
CHLPN	2	1	0	0	0	0	0	0	0	0	0	0	0
CHLCV	2	1	0	0	0	0	0	0	0	0	0	0	0
ECO57	46	3	0	0	0	0	0	0	0	0	0	0	0
YERPE	31	3	0	0	0	0	0	0	0	0	0	0	0
YERPS	4	0	0	0	0	0	0	0	0	0	0	0	0
SALTY	41	3	0	0	0	0	0	0	0	0	0	0	0
SALTI	41	3	0	0	0	0	0	0	0	0	0	0	0
SALCH	42	3	0	0	0	0	0	0	0	0	0	0	0
SHIFL	5	0	0	0	0	0	0	0	0	0	0	0	0
HAEIN	6	2	0	0	0	0	0	0	0	0	0	0	0
HAEDU	4	2	0	0	0	0	0	0	0	0	0	0	0
PASMU	9	2	0	0	0	0	0	0	0	0	0	0	0
VIBCH	46	3	0	0	0	0	0	0	0	0	0	0	0
VIBPA	54	3	0	0	0	0	0	0	0	0	0	0	0
VIBUY	66	3	0	0	0	0	0	0	0	0	0	0	0
METCA	15	2	0	0	0	0	0	0	0	0	0	0	0
NEIMBE	4	2	0	0	0	0	0	0	0	0	0	0	0

Table A-4a contd.

Organisms Code	Cluster Number												
	2	49	887	1679	1681	1699	1748	1823	1883	1905	5060	5995	6392
NEIG1	3	2	0	0	0	0	0	0	0	0	0	0	0
BORPA	31	2	0	0	0	0	0	0	1	0	0	0	0
BORPE	27	2	0	0	0	0	0	0	1	0	0	0	0
BURMA	40	4	0	0	0	0	0	0	2	0	0	0	0
CHRVO	64	3	1	0	0	0	0	0	1	0	0	0	0
RICPR	3	2	1	0	0	0	0	0	0	0	0	0	0
RICCN	5	2	1	0	0	0	0	0	0	0	0	0	0
BRUSU	24	3	1	0	0	0	0	0	0	0	0	0	0
BRUME	27	3	1	0	0	0	0	0	0	0	0	0	0
CAMJE	9	1	1	0	0	0	0	0	0	0	0	0	0
HELPJ	5	1	1	0	0	0	0	0	0	0	0	0	0
AQUAE	8	2	1	0	0	0	0	0	0	0	0	0	0
DEIRA	33	0	0	0	0	0	0	0	0	0	0	0	0
CORGL	15	2	1	0	0	0	0	0	0	0	0	0	0
COREF	15	2	1	0	0	0	0	0	0	0	0	0	0
STRCO	92	13	1	1	1	1	1	1	1	1	0	0	0
STRAW	78	12	1	1	1	1	1	1	1	1	0	0	0
BIFLO	11	1	1	0	0	0	0	0	0	0	0	0	0
CLOAB	49	5	1	0	0	0	0	0	0	0	0	0	0
BACSU	37	5	1	0	0	0	0	0	0	0	0	0	0
BACHD	43	6	1	0	0	0	0	0	0	0	0	0	0
OCEIH	26	6	1	0	0	0	0	0	0	0	0	0	0
LACJO	9	1	1	0	0	0	0	0	0	0	0	0	0
LACPL	13	1	1	0	0	0	0	0	0	0	0	0	0
LACLA	10	1	1	0	0	0	0	0	0	0	0	0	0
THETN	27	6	1	0	0	0	0	0	0	0	0	0	0
PSEAE	67	3	0	0	0	0	0	0	2	0	0	0	0
PSEPK	67	3	0	0	0	0	0	0	0	0	0	0	0
BUCAP	0	2	0	0	0	0	0	0	0	0	0	0	0
BUCAI	0	2	0	0	0	0	0	0	0	0	0	0	0
ECOLI	44	3	0	0	0	0	0	0	0	0	0	0	0
WIGBR	1	2	0	0	0	0	0	0	0	0	0	0	0
NITEU	12	2	0	0	0	0	0	0	0	0	0	0	0
CAUCR	44	2	1	0	0	0	0	0	1	0	0	0	0
BRAJA	92	4	1	0	0	0	0	0	0	0	0	0	0
DESVH	47	2	1	0	0	0	0	0	0	0	0	0	0
GEOSL	59	3	1	0	0	0	0	0	0	0	0	0	0
WOLSU	37	1	1	0	0	0	0	0	0	0	0	0	0
SILPO	23	3	0	0	0	0	0	0	0	0	0	0	0

Table A-4b: The result of Known virulence *Mycobacterium tuberculosis* genes 0.5 S

Cluster No	Description Line	INTERPRO				
2	MYCOBACTERIAL_PERSISTENCE_REGULATOR_MRPA_(TWO_COMPONENT_RESPONSE_TRANSCRIPTIONAL_REGULATORY_PROTEIN)_(DNA-binding_response_regulator)	IPR001789	IPR001867	IPR005829	IPR011006	IPR011991
49	RNA_polymerase_sigma_factor_rpoD_(Sigma-A)	IPR009042	IPR007624	IPR007627	IPR007630	IPR000943 IPR012760
887	Cytotoxin_/haemolysin_homologue_(CYTOTOXIN|HAEMOLYSIN_HOMOLOGUE_TLYA)_(Cytotoxin/hemolysin)	IPR002942	IPR002877	IPR004538		
1679	MCE-FAMILY_PROTEIN_MCE2D_(Virulence_factor_mce_family_protein)	IPR003399	IPR005693			
1681	Virulence_factor_mce_family_protein	IPR003399	IPR008360	IPR005693		
1699	MCE-FAMILY_PROTEIN_MCE1B_(Virulence_factor_mce_family_protein)	IPR003399	IPR005693			
1748	Virulence_factor_mce_family_protein_(MCE-FAMILY_PROTEIN_MCE4A)	IPR003399	IPR005693			
1823	Virulence_factor_mce_family_protein	IPR003399	IPR005693			
1883	Phospholipase_C_1_precursor_(EC_3.1.4.3)_(MTP40_antigen)/	IPR007312	IPR006311			
1905	POSSIBLE_MCE-FAMILY_LIPOPROTEIN_LPRM_(MCE-FAMILY_LIPOPROTEIN_MCE3E)_(Virulence_factor_mce_family_protein)	IPR003399	IPR008995	IPR005693		
5060	Possible_virulence-regulating_38_kDa_protein	IPR012287	IPR000005	IPR009057		
5995	Exported_repetitive_protein_precursor_(Cell_surface_protein_pirG)_(EXP53)	IPR008165				
6392	Heparin-binding_hemagglutinin_(Adhesin)	IPR000897				

Table A-4c: The result of Known virulence *Mycobacterium tuberculosis* genes 0.5 S

Cluster Number	Total Protein in pathogens	Number of pathogens	Present inTB?	Virulence In TB?	Virulence in others?	Total Protein in non pathogens	Number of non pathogens	Total genomes	Total Protein in cluster
2	1117	54	y	y in mtb	y in salty*	959	26	80	2076
49	124	54	y	y in mtb		95	27	81	219
887	27	27	y	y in mtb		19	19	46	46
1679	20	5	y	y in mtbc		2	2	7	22
1681	20	5	y	y in mtbc		2	2	7	22
1699	20	5	y	y in mtbc		2	2	7	22
1748	19	5	y	y in mtbc		2	2	7	21
1823	18	5	y	y in mtbc		2	2	7	20
1883	14	7	y	y in mtb		5	4	12	19
1905	17	5	y	y in mtbc		2	2	7	19
5060	6	3	y	y in mtbc		0	0	4	6
5995	5	5	y	y in mtbc		0	0	5	5
6392	5	5	y	y in mtbc		0	0	5	5