THE MYCOSINS, A FAMILY OF SECRETED SUBTILISIN-LIKE SERINE PROTEASES ASSOCIATED WITH THE IMMUNOLOGICALLY-IMPORTANT ESAT-6 GENE CLUSTERS OF *MYCOBACTERIUM TUBERCULOSIS*

Nicolaas Claudius Gey van Pittius VI

Dissertation presented for the degree of Doctor of Philosophy at the University of Stellenbosch

Promoters: Prof. A. D. Beyers

and Dr. R. M. Warren

Co-promoter: Prof. P. D. van Helden

Stellenbosch

December 2002

Declaration

I, the undersigned, hereby declare that the work contained in this dissertation is my own original work, and has not, to my knowledge, previously in its entirety or in part been submitted at any university for a degree.

Date

Summary

Pathogenic organisms frequently utilize proteases to perform specific functions related to virulence. There is little information regarding the role of proteolysis in *Mycobacterium tuberculosis* and no studies on the potential involvement of these enzymes in the pathogenesis of tuberculosis. The present study initially focused on the characterization of a family of membrane anchored, cell wall associated, subtilisin-like serine proteases (mycosins-1 to 5) of *Mycobacterium tuberculosis*. These proteases were shown to be constitutively expressed in *M. tuberculosis*, to be located in the cell wall of the organism and to be potentially shed (either actively or passively) from the wall. Relatively high levels of gamma interferon secretion by T-cells in response to these proteases were observed in Mantoux positive individuals. The absence of any detectable protease activity lead to a protein sequence analysis which indicated that the mycosins are probable mycobacterial-specific proprotein processing proteases.

To identify possible substrates for these proteases, the genome sequence regions surrounding the mycosin genes were analyzed. This revealed that the mycosin genes are in fact part of a cluster of 6 to 12 genes which have been duplicated multiple times in the genome of *M. tuberculosis*. Due to the presence of members of the previously described ESAT-6 T-cell antigen family within this duplicated region, the five gene cluster regions were named the ESAT-6 loci. *In silico* analysis of finished and unfinished genome sequencing data revealed the presence of orthologues of the *Mycobacterium tuberculosis* H37Rv ESAT-6 loci in the genomes of other mycobacteria, e.g. *M. tuberculosis* CDC1551, *M. tuberculosis* 210, *M. bovis, M. leprae*, *M. avium*, and the avirulent strain *M. smegmatis*. Phylogenetic analyses done on the resulting sequences have established the duplication order of the gene clusters and demonstrated that gene cluster region 4 (Rv3444c-3450c) is ancestral. Region 4 is also the only region for which an orthologue could be found in the genomes of *Corynebacterium diptheriae* and *Streptomyces coelicolor*. Thus, the comparative genomic analyses revealed that the presence of the ESAT-6 gene cluster seems to be a unique characteristic shared by members of the high G+C gram-positive bacteria and that multiple duplications of this cluster have occurred and have been maintained only within the genomes of members of the genus *Mycobacterium*.

The ESAT-6 gene cluster regions were shown to consist of the members of the ESAT-6 gene family (encoding secreted T-cell antigens that lack detectable secretion signals), the mycosins (secreted, cell wall-associated subtilisin-like serine proteases) as well as genes encoding putative ABC transporters, ATP-binding proteins, and other membrane-associated proteins. Thus, from the observation that members of the ESAT-6 family are secreted without the normal *sec*-dependent secretion signals, it was hypothesized that the membrane-associated and energy-providing proteins function together to form a transport system for the secretion of the members of the ESAT-6 protein family. Supporting this hypothesis, one of the ESAT-6 gene clusters was shown to be expressed as a single polycistronic RNA, forming an operon structure. The promoter for this operon, P_{ESREG3}, was also identified and its activity characterized. Subsequent secretion analyses results have shown that secretion of members of the ESAT-6 protein family is dependent on the presence of the proteins encoded by the ESAT-6 gene cluster regions, confirming the putative transport-associated functions of the ESAT-6 gene cluster-encoded proteins. The mycobacterial ESAT-6 gene clusters contain a number of features of quorum sensing and lantibiotic operons, and an extensive review of the literature have led to the hypothesis that the members of the ESAT-6 family may be secreted as signaling molecules and may be involved in the regulation of expression of genes during intracellular residence of the bacterium. In the final part of this study, the evolutionary history of the PE and PPE gene families (members of which is found situated in the ESAT-6 gene clusters) were investigated. This investigation revealed that the expansion of these families are linked to the duplications of the ESAT-6 gene clusters, which is supported by the absence of the multiple copies of the PE and PPE families in the genome of the fast-growing mycobacterium *M. smegmatis*. Furthermore, dot blot analyses showed that the PPE gene present in ESAT-6 gene cluster region 5 is able to distinguish between mycobacteria belonging to the slow-growing or fast-growing species, indicating a function for the genes of these two families and/or the ESAT-6 gene clusters in the phenotypical differences distinguishing these two groups of mycobacteria.

In conclusion, this study has highlighted numerous important aspects of mycobacterial genomics and has greatly contributed to the current body of knowledge concerning the role of proteases, gene duplication and mechanisms of antigen expression and secretion in *M. tuberculosis*.

Opsomming

Patogeniese organismes gebruik gereeld proteases om spesifieke funksies te verrig wat te doene het met virulensie. Daar is egter baie min inligting beskikbaar aangaande die rol van proteolise in *Mycobacterium tuberculosis* en geen studies is al gedoen om die invloed van hierdie ensieme in die patogenese van tuberkulose te bestudeer nie. Die huidige studie het oorspronklik gefokus op die karakterisering van `n familie van membraan-geankerde, selwand-geassosieërde, subtilisien-agtige serien proteases (mycosins-1 tot 5) van *Mycobacterium tuberculosis*. Daar is aangetoon dat hierdie proteases deurgans uitgedruk word in *M. tuberculosis*, dat hulle in die selwand van die organisme geleë is, en dat hulle potensieël afgeskilfer word (of aktief of passief) vanaf die wand. Relatiewe hoë vlakke van gamma interferon uitskeiding deur T-selle is verkry in respons tot die proteases in Mantoux positiewe persone. Die afwesigheid van enige opspoorbare protease aktiwiteit het gelei tot die analisering van die protein volgordes van hierdie ensieme, wat daarop gedui het dat die mycosins heel moontlik mikobakteriële-spesifieke proprotein prosesserende proteases is.

Die genoom volgorde gebiede rondom die mycosin gene is geanaliseer om die identiteit van die moontlike substrate van die proteases te identifiseer. Dit het gewys dat die mycosin gene in werklikheid deel vorm van `n groep van 6 tot 12 gene wat veelvuldiglik in die genoom van *M. tuberculosis* gedupliseer is. As gevolg van die aanwesigheid van lede van die voorheen beskryfde ESAT-6 T-sel antigeen familie binne hierdie gedupliseerde geen groep, is die vyf geen groep gebiede die ESAT-6 lokusse genoem. *In silico* analises van voltooide en onvoltooide genoom volgorde data het die teenwoordigheid van ortoloë van die *Mycobacterium tuberculosis* H37Rv ESAT-6 lokusse in die genome van ander mikobakterië, bv. *M. tuberculosis* CDC1551, *M. tuberculosis* 210, *M. bovis, M. leprae*, *M. avium*, en die nie-virulente ras *M. smegmatis*, bevestig. Filogenetiese analises wat op die geenvolgordes uitgevoer is het die volgorde van duplisering van die geen groepe vasgestel, en het aangetoon dat die geen groep gebied 4 (Rv3444c-3450c) die voorouer is van die ander duplikate. Gebied 4 is ook die enigste gebied waarvoor daar ortoloë gebiede in die genome van *Corynebacterium diptheriae* en *Streptomyces coelicolor* gevind kon word. Die vergelykende genomiese analises het dus aangetoon dat die teenwoordigheid van die ESAT-6 geen groepe `n unieke kenmerk is van die lede van die hoë G+C gram-positiewe bakterië en dat die veelvuldige duplisering van hierdie gebiede slegs plaasgevind het en behou is in die genome van die lede van die genus *Mikobakterium*.

Die ESAT-6 geen groep gebiede bestaan uit die gene van die ESAT-6 geen familie (wat kodeer vir gesekreteerde T-sel antigene sonder enige bespeurbare sekreteringsseine), die mycosins (wat gesekreteerde, selwand-geassosieërde subtilisien-agtige serien proteases is) sowel as gene wat kodeer vir moontlike ABC tranporters, ATP-bindingsproteine, en ander membraan-geassosieërde proteine. Aangesien die lede van die ESAT-6 familie gesekreteer word sonder enige normale sec-afhanklike sekreteringsseine, is daar gehipotetiseer dat hierdie membraan-geassosieërde en energie-verskaffende proteine saam funksioneer om `n transport sisteem te vorm vir die sekretering van die lede van die ESAT-6 protein familie. Hierdie hipotese is ondersteun deur die resultate wat aangedui het dat een van die ESAT-6 geen groep gebiede uitgedruk word as een enkele polisistroniese RNA en dus `n operon vorm. Die promoter vir hierdie operon, P_{ESREG3}, is ook geidentifiseer en die aktiwiteit daarvan gekarakteriseer. Daaropvolgende sekresie analise resultate het getoon dat die sekresie van die lede van die ESAT-6 protein familie afhanklik is van die teenwoordigeheid van die proteine wat uitgedruk word deur die ESAT-6 geen groep gebiede, wat die moontlike transport-geassosieërde funksies van die ESAT-6 geen groep-geënkodeerde proteine bevestig. Die mikobakteriële ESAT-6 geen groepe vertoon `n hele aantal eienskappe van kworum aanvoelings en lantibiotiese operone, en `n omvattende oorsig van die literatuur het gelei tot die hipotese dat die lede van die ESAT-6 familie moontlik as sein molekules gesekreteer mag word en moontlik betrokke is by die regulering van die uitdrukking van gene gedurende die intrasellulêre verblyf van die bakterium. In die finale gedeelte van hierdie studie, is die evolusionêre geskiedenis van die PE en PPE geen families (waarvan lede teenwoordig is in die ESAT-6 geen groep gebiede) ondersoek. Hierdie ondersoek het openbaar dat die uitbreiding van hierdie families gekoppel kan word aan die duplisering van die ESAT-6 geen groepe, wat ondersteun word deur die afwesigheid van die veelvuldige kopieë van die PE en PPE families in die genoom van die vinnig-groeiende mikobakterium *M. smegmatis*. Dot blot analises het verder aangetoon dat die PPE geen wat geleë is in die ESAT-6 geen groep gebied 5 die vermoë het om te kan onderskei tussen mikobakterieë wat behoort aan die vinnig-groeiende of die stadig-groeiende spesies. Dit dui daarop dat die gene van hierdie twee families en/of die ESAT-6 geen

groepe een of ander funksie verrig wat `n invloed het op die fenotipiese verskille wat die twee groepe mikobakterië van mekaar onderskei.

Om saam te vat, hierdie studie het `n hele aantal belangrike aspekte van mikobakteriële genomika aangeraak en het grootliks bygedra tot die huidige kennis aangaande die rol van proteases, geenduplisering en die meganismes van antigeen uitdrukking en sekresie in *M. tuberculosis*.

Publications and Presentations

Portions of this thesis have been published as:

1) **Chapter Two:** Brown, G.D., Dave, J.A., Gey van Pittius, N.C., Stevens, L., Ehlers, M.R.W., and Beyers, A.D., **The mycosins of *M. tuberculosis* H37Rv: A family of Subtilisin-Like Serine Proteases,** Gene, 2000, Aug 22, 254 (1-2): 147-155.
2) **Chapter Three:** Gey van Pittius, N.C., Gamieldien, J., Hide, W., Brown, G.D., Siezen, R.J., and Beyers, A.D., **The ESAT-6 gene cluster of *Mycobacterium tuberculosis* and other high G+C gram-positive bacteria,** Genome Biology 2001 2(10): research0044.1-0044.18

The following manuscripts that contain portions of this thesis will be submitted:

1) **Chapter Two:** Dave, J.A., Gey van Pittius, N.C., Beyers, A.D., Ehlers, M.R.W., and Brown, G.D., **Mycosin-1, a Subtilisin-like Serine Protease of *Mycobacterium tuberculosis*, is Cell Wall-Associated and Expressed during Infection of Macrophages,**
2) **Chapter Three:** Gey van Pittius, N.C., **An Acid Fast Guide to Web Resources for Comparative Mycobacterial Genomics** (Tutorial),
3) **Chapter Three:** Gey van Pittius, N.C., Warren, R.M., and Van Helden, P.D., **The mycobacterial ESAT-6 gene clusters** (Review),
4) **Chapter Three:** Gey van Pittius, N.C., **ESAT-6 and CFP-10: What is the diagnoses?** (Counterpoint),
5) **Chapter Four:** Gey van Pittius, N.C., Warren, R.M., and Van Helden, P.D., **Mycosin-3, a Subtilisin-like Serine Protease *of Mycobacterium tuberculosis*, is expressed as part of the ESAT-6 gene cluster region 3 operon along with members of the ESAT-6, CFP-10, PE and PPE multigene families,**
6) **Chapter Five:** Gey van Pittius, N.C., Daugelat, S., Warren, R.M., Kaufmann, S.H.E., and Van Helden, P.D., **The ESAT-6 gene cluster of *Mycobacterium tuberculosis* forms a multi-component transport system for the secretion of members of the immunologically important ESAT-6 and CFP-10 multigene families,**

7) **Chapter Five:** Gey van Pittius, N.C., Warren, R.M., Siezen, R.J., and Van Helden, P.D., **Are the immunologically important mycobacterial ESAT-6 gene clusters involved in quorum sensing and cell-cell signaling?** (Hypothesis),
8) **Chapter Six:** Gey van Pittius, N.C., Sampson, S.L., Lee, H., Warren, R.M., and Van Helden, P.D., **The evolutionary history of the expansion of the *Mycobacterium tuberculosis* PE and PPE multigene families and its association with the duplication of the ESAT-6 gene cluster.**

Presentations of work contained in this thesis have been made by the author at the following international meetings:

1) **1st International Union of Biochemistry and Molecular Biology and South African Society of Biochemistry and Molecular Biology Special Meeting on Biochemical and Molecular Basis of Disease**, Cape Town, South Africa, November 2001 (Poster presentation)
2) **Genomes 2000 Conference: International Conference on Microbial and Model Genomes**, Paris, France, April 2000 (Poster presentation)
3) **2nd International Congress of the Federation of African Societies of Biochemistry and Molecular Biology in conjunction with the 15th Congress of the South African Society of Biochemistry and Molecular Biology**, Potchefstroom, South Africa, September 1998 (Oral presentation)

Presentations of work contained in this thesis have been made by the author at the following national meetings:

1) **AstraZeneca Joint UCT, US, UWC, MRC 4th Annual Medical Research Day**, Cape Town, October 2001 (Oral presentation)
2) **45th Academic Yearday**, Faculty of Medicine, University of Stellenbosch, Tygerberg, August 2001 (Oral presentation)
3) **AstraZeneca Joint UCT, US, UWC, MRC 3rd Annual Medical Research Day**, Cape Town, September 2000 (Oral presentation)

4) **11th Biennial Congress of the South African Society for Microbiology as part of the BioY2K Combined Millennium Meeting**, Grahamstown, January 2000 (Oral presentation)
5) **AstraZeneca Joint UCT, US, UWC, MRC 2nd Annual Medical Research Day**, Cape Town, August 1999 (Oral presentation)
6) **43rd Academic Yearday**, Faculty of Medicine, University of Stellenbosch, Tygerberg, August 1999 (Oral presentation)
7) **42nd Academic Yearday**, Faculty of Medicine, University of Stellenbosch, Tygerberg, August 1998 (Poster presentation)

Invited Oral Presentations of the work contained in this thesis have been made by the author at the following International and National Institutions:

1) **Fogarty Foundation Workshop**, Lung Institute, University of Cape Town Medical School, Cape Town, South Africa (September 2001)
2) Department of Tuberculosis Immunology, **Statens Serum Institut**, Copenhagen, Denmark (May 2000)
3) Department of Immunology, **Max-Planck-Institute for Infection Biology**, Berlin, Germany (May 2000)
4) Department of Immunology, **University of Cape Town Medical School**, Cape Town, South Africa (January 2000)
5) Oral Presentations at Various Departmental Research Meetings at the **Department of Medical Biochemistry**, University of Stellenbosch, Tygerberg, South Africa (1998 - 2001)

Presentations of work contained in this thesis have been made by collaborators on the following occasions:

1) **8th International Conference on Small Genomes**, Lake Arrowhead, USA, September 2000 (Poster presentation)
2) **Keystone Symposium on Macrophage Biology**, Keystone, USA, January 1999 (Poster presentation)
3) **Keystone Symposium on TB: Molecular Mechanisms and Immunological Aspects**, Keystone, USA, April 1998 (Poster presentation)

Acknowledgements

I would like to acknowledge the contributions made by the following persons, without whom I would not have been able to finish this study:

Eerstens wil ek my Here, die God wat die hemel en aarde geskape het, bedank vir Sy teenwoordigheid in my lewe. Dankie Heer dat U my beskerm en bewaar het deur hierdie tydperk in my lewe en dat ek altyd op U kan vertrou, ook vir die toekoms. Al wat ek kan sê is: "Soli Deo Gloria!"

My Méshie, my vrou, my lewe. Baie dankie vir jou liefde, jou hulp en ondersteuning, ek weet dit was vir jou swaar met tye en jy het baie opgeoffer, en ek is baie dankbaar daarvoor. Ek is baie lief vir jou.

Pa en Ma. Dankie dat julle my belangstelling in die wetenskap altyd ondersteun het en dit nooit van my weerhou het nie al het julle baie goed geweet dat daar geen geld in is nie! Ek is lief vir julle.

My familie - my sussie Marissa, my broer Hugo, Ouma Maaike, Ma Lida en Ouma Driekie. Ek is baie lief vir julle almal en baie dankbaar vir julle ondersteuning.

Albert. Ek het nooit die kans gekry om vir jou dankie te se dat jy my die geleentheid gegee het om vir jou te kon werk nie. Jy het my baie geleer van die lewe en van hoe dit is om `n passie vir die wetenskap te he. Ek mis jou baie en is net jammer dat jy nie kon sien hoe ons werk geblom het nie.

Rob, I would like to extend my deepest gratitude towards you for all you have done for me in the past one and a half years. I have really learnt a lot from you and am indebted to you for being willing to help me finish off the work started by myself and Albert. You have been an excellent promoter, a stimulating colleague and a great friend, and I am looking forward to building upon this in the future.

Gordon, Paul and Joel. I am indebted to you all for shaping my career, for continued advice and support, be it intellectually or financially. It has been a pleasure to be associated with you.

The members of the "A-team" and also other past and present friends in the department - Kevin, Debbie, Sam, Glenda, Joanita, Carmen, Lorraine and Siobhán. Thank you all for your friendship.

All my other colleagues in the Department of Medical Biochemistry, University of Stellenbosch.

Tannie Arja, Tannie Rika, Basil, Shahied and Monica - you have played an essential part in the successful completion of my work over the past few years. I thank you for your dedication and friendship.

I would also like to thank the following people for their contributions, which incudes reagents, clones and constructs used during the extent of this study: Francis Nano and Jamie Doran for the *M. tuberculosis* PhoA bank, Koen de Smet and Douglas Young for mycobacterial shuttle vectors, John Belisle for various mycobacterial products and fractions (generated through the National Institutes of Health Contract No1-AI-75320, entitled "Tuberculosis Research Materials and Vaccine Testing."), Samantha Sampson for PPE primers and help with RT-PCR, Heyoung Lee for the PPE dot blot analyses, Ian Wiid for *M. bovis* BCG Tokyo, Stewart Cole for *M. leprae* cosmids, Carel van Heerden for all DNA sequencing, Karin Eiglemeier for *M. tuberculosis* cosmids, William Bourn for valuable advice, comments and discussions as well as for sharing a mycobacterial promoter probe vector, Yossef Av-Gay and Brigitte Giquel for mycobacterial promoter probe vectors, Helen Kieser for the *Streptomyces coelicolor* cosmid, Aruni de Zoysa for *Corynebacterium diphtheriae* genomic DNA, Maryke Henton for *Corynebacterium*, *Nocardia* and *Rhodococcus* strains, Sabine Daugelat and Miriam Braunstein for various plasmids, constructs and advice, Franz Christoph Bange for the integrating *M. tuberculosis* H37Rv and *M. smegmatis* genomic DNA cosmid libraries, Peter Andersen, Ida Rosenkrans and Rikke Skjot for anti-ESAT-6 and anti-TB10.4 antibodies and purified ESAT-6 dimer and TB10.4 protein, Laurens van Pinxteren, Peter Andersen, Stefan Kaufmann and Sabine Daugelat for their immense hospitality during my visits to their laboratories and Roland Siezen, Neil Rawlings, Ken Duncan and Douglas Rawlings for valuable comments.

Also, thank you to everyone who have read parts of this dissertation and have provided constructive critisism. Here I would like to once again extend a very special thank you to Rob Warren, my

promoter. Also to my collaborators Roland Siezen, Win Hide and Junaid Gamieldien, and to the anonymous reviewers of chapters 2 and 3.

Parts of this project was financially supported by the following institutions: Glaxo Smithkline Action TB Initiative, Harry Crossley Foundation, MRC of South Africa, and the University of Stellenbosch.

Preliminary sequence data for *Mycobacterium tuberculosis* 210, *Mycobacterium avium* 104 and *Mycobacterium smegmatis* MC^2 155 was obtained from The Institute for Genomic Research (TIGR). Preliminary sequence data for *Mycobacterium paratuberculosis* K10 was obtained from the University of Minnesota. Preliminary sequence data for *Mycobacterium bovis* AF2122/97(spoligotype 9), *Corynebacterium diphteriae* NCTC13129 and *Streptomyces coelicolor A3(2)*, was obtained from the Sanger Centre.

"The more I study nature, the more I stand in awe before the Creator"

Louis Pasteur

"Ever since God created the world, his invisible qualities, both his eternal power and his divine nature, have been clearly seen; they are perceived in the things that God has made"

The Holy Bible - Romans 1: 20

This work is dedicated to the memory of two very special men who, through their unwavering support for me and through their own greatness have shaped my view of the world and of life......

.....to my grandfather, Nicolaas Claudius Gey van Pittius IV

Your respect for studiousness and the unquenched thirst for knowledge that you awakened in me are only equaled by my memory of the proud look in your eyes when you held my small face

.... and to my mentor and dearest friend, Albert Beyers

You taught me to love science and nature through the eyes of God, to marvel at the intricacies of the smallest microbe and to cry for the pain in the eyes of the sick

I will never forget you...

Table of Contents

List of Abbreviations

aa	amino acids
ABC	ATP-binding casette
AIDS	aquired immune-deficiency syndrome
ALLM	N-acetyl-Leu-Leu-methional
ALLN	N-acetyl-Leu-Leu-norleucinal
Amp	Ampicillin
Asp	asparagine
BCG	Bacille de Calmette et Guerin
BLAST	basic local alignment search tool
bp	basepair
BSA	bovine serum albumin
CF	culture filtrate
CFP-10	culture filtrate protein 10
CMI	cell mediated immune response
3, 4-DCI	3, 4-dichloroisocoumarin
°C	degrees Celsius
DNA	deoxyribonucleic acid
DTH	delayed-type hypersensitivity
DTT	dithiothreitol
E.	*Escherichiae*
EDTA	ethylenediaminotetraacetic acid
ELISA	enzyme-linked immunosorbent assay
ESAT-6	6 kDa early-secreted antigenic target
E-64	L-trans-epoxysuccinyl-leucylamido-[4-guanidino]-butane
FITC	fluorescein isothiocyanate
G+C	guanine + cytosine
g/l	grams per liter
GST	glutathione-S-transferase

HA	hemagglutinin
His	histidine
HIV	human immunodeficiency virus
hr	hours
HRPO	horseradish peroxidase
Hyg	Hygromycin
IFN-γ	gamma interferon
IS	insertion sequence
IU	international units
Kan	Kanamycin
kbp	kilobasepairs
kDa	kilodalton
LB	Luria-Bertani (medium)
M.	*Mycobacterium*
Mb	megabases
MBP	maltose binding protein
MHC	major histocompatibility complex
μg	microgram
μM	micromolar
μl	microliter
min	minutes
M	molar
MPTR	major polymorphic tandem repeat
ng	nanogram
nm	nanometer
OADC	oleic acid/albumin/dextrose/catalase
OD	optical density
ONPG	*o*-nitrophenyl-β-D-galactoside
ORF	open reading frame
PAGE	polyacrylamide gel electrophoresis

PBMC	peripheral blood mononuclear cells
PBS	phosphate buffered saline
PCR	polymerase chain reaction
PE	proline-glutamic acid
%	percent
P_{ESREG3}	ESAT-6 gene cluster region 3 promoter
PGRS	polymorphic GC-rich sequence
PHA	phytohemagglutinin
PhoA	bacterial alkaline phosphatase
PMSF	phenylmethanesulfonyl fluoride
PPD	purified protein derivative
PPE	proline-proline-glutamic acid
^{r}	resistance
RBS	ribosome binding site
RD	region of difference
RNA	ribonucleic acid
rpm	revolutions per minute
RT-PCR	reverse transcriptase polymerase chain reaction
SDS	sodium dodecyl sulphate
Ser	serine
ST-CF's	short-term culture filtrates
subsp.	subspecies
TAE	tris/acetic acid/EDTA buffer
TB	tuberculosis
TCA	trichloroacetic acid
TGF-β	transforming growth factor beta
Th1	T-helper 1
3D	three-dimensional
TIGR	The Institute for Genomic Research
TLCK	L-1-chloro-3-[4-tosylamido]-7-amino-2-heptanone HCl

TPCK	L-1-chloro-3-[4-tosylamido]-4-phenyl-2-butanone
TU	tuberculin units
U	units
WCL	whole cell lysate
X-Gal	5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside

CHAPTER ONE

INTRODUCTION

"He died of consumption, died forgotten, died withered and blighted like the flowers a lover has given to his mistress, which she leaves to die secreted in a drawer where she has hidden them from the world."

The Man in the Iron Mask (Chapter V: Two Friends) - Dumas

1.1. Preamble

Mycobacterium tuberculosis, which causes the human disease tuberculosis, is an extremely slow-growing rod-shaped bacterium with a thick cell wall (Shinnick and Good, 1994). Despite more than a hundred years of research on this organism, the mechanisms of its pathogenicity is still poorly understood (Ehlers, 1993). This paucity of knowledge may be due to a number of factors, including the slow growth of the organism (it may take up to six weeks to form a colony on a plate) leading to problems with contamination and making experiments extremely time consuming (Parish and Stoker, 1999). Other difficulties encountered by mycobacteriologists includes the extensive clumping of the organisms when grown in liquid culture due to the thick, lipid-rich cell wall, the resistance of the organism to standard chemical lysis procedures, the fact that experiments have to be carried out under category 3 biosafety conditions, the high G+C content of the mycobacteria (leading to difficulty in molecular techniques such as the polymerase chain reaction), the spontaneous acquirement of antibiotic resistance, and the absence of general mycobacterial genetic research tools (Parish and Stoker, 1999). While scientists are struggling to improve the basic tools required for the provision of the necessary foundation for future progress in molecular analysis of the mycobacteria, the pathogen remains the most important cause of mortality due to a single infectious agent, and the incidence of disease is on the increase worldwide (Ehlers, 1993). Thus, there is an urgent need for an increase in the understanding of the metabolism of the pathogen, its interactions with the host as well as the host responses to infection. The recent sequencing of the whole genome sequences of two strains of *M. tuberculosis* (Cole *et al.*, 1998) and the causative agent of leprosy, *M. leprae* (Cole *et al.*, 2001), as well as the ongoing sequencing projects of other members of the genus *Mycobacterium*, has opened the way for the unraveling of the molecular basis of the pathogenesis, host range specificity, evolution and phenotypic differences involved in growth characteristic and virulence (Brosch *et al.*, 2001).

In the present study a family of five subtilisin-like serine proteases of *M. tuberculosis* were identified, cloned and characterized. As proteases are frequently involved in important mechanisms of pathogenicity (Maeda and Yamamoto, 1996), the significance of the duplications of these genes was studied. Using the recently available whole genome sequence data of several members of the genus *Mycobacterium*, including the reference strain *M. tuberculosis* H37Rv (Cole *et al.*, 1998), other genes were found to be associated with these proteases, including the secreted, immunologically-important, potent T-cell antigens ESAT-6 and CFP-10 (Sørensen *et al*., 1995, Berthet *et al*., 1998). This in turn lead to the identification of an association between the duplication of these genes and the expansion of the newly described PE and PPE multigene families (novel genes of unknown function that have been associated with pathogenicity, and which makes up 10 % of the coding potential of the genome of *M. tuberculosis*)(Cole *et al*., 1998).

1.2. Background

1.2.1. Tuberculosis

Tuberculosis is an infectious disease caused by the bacillus *Mycobacterium tuberculosis* and remains the greatest cause of death worldwide due to a single pathogen (Ehlers, 1993). It usually enters the body through inhalation of aerosol droplets carrying bacteria into the lungs of an individual (Riley *et al.*, 1959, McKinney *et al.*, 1998), after which disease progression involves phagocytosis of the organisms by alveolar macrophages, and subsequent survival and replication within these cells. Over 80% of tuberculosis cases present as pulmonary tuberculosis, which is also the most infectious form of the disease, and leads to coughing, fever, weight loss, tiredness and the coughing up of blood (Hopewell, 1994). Extra-pulmonary tuberculosis results from the spread of the organism to other parts and organs of the body.

In 1993, the World Health Organization (WHO) took an unprecedented step and declared tuberculosis a global emergency, so great was the concern about the modern dangerously growing tuberculosis epidemic (WHO, 1993). The registered number of new cases of TB worldwide roughly correlates with economic conditions: the highest incidences are seen in the countries of Africa, Asia, and Latin America with the lowest gross national products (NJMS National Tuberculosis Center, http://www.umdnj.edu/~ntbcweb/history.htm). The breakdown in health services (Kochi, 1994), the variability in the efficacy of BCG vaccination (Colditz *et al.*, 1994, Fine, 1995, Roche *et al.*, 1995), the spread of HIV/AIDS (Barnes *et al.*, 1991, Schulzer *et al.*, 1992, Haas and Des Prez, 1994, Wendel *et al.*, 2001), the failure to complete treatment due to the long duration of chemotherapy and noncompliance (Lipsitch and Levin, 1998, Agrawal *et* al., 2001) and the subsequent emergence of multidrug-resistant tuberculosis (Willcox, 2000) are all factors contributing to the worsening impact of this disease. It is estimated that between the years 2000 and 2020, nearly one billion people will be newly infected, 200 million people will become sick, and 35 million will die from tuberculosis - if control is not further strengthened (WHO Fact Sheet N°104, Revised April 2000). Approximately 2 billion people are thought to be presently infected worldwide, with about 3 million deaths occurring annually due to this disease. This is alarming, but is further compounded by the fact that the increasing incidence does not only occur in developing countries, but also in industrialized countries due to the emergence of HIV infection and drug resistance (Yew and Chau, 1995). The identification of novel drug targets and the development of an effective vaccine against *M. tuberculosis* is thus of utmost importance to combat this disease.

1.2.2. The history of tuberculosis

Tuberculosis in humans (otherwise known as consumption, phthisis or chronic wasting) was documented in history as early as Ancient Greece, when the well-known physician Hippocrates identified phthisis as the most widespread disease of the times, and noted that it was almost always fatal ("Of the Epidemics", by Hippocrates, ca 400 B.C.). We now know that *Mycobacterium tuberculosis* was infecting humans much earlier, as pathological signs of tubercular decay was found in fragments of the spinal column from Egyptian mummies from 2400 B.C. (NJMS National Tuberculosis Center, http://www.umdnj.edu/~ntbcweb/history.htm). Salo and coworkers (1994) identified *M. tuberculosis* DNA in 1000-year-old lung tissue of a pre-Columbian Peruvian mummy and spinal deformities in a pre-dynastic Egyptian mummy, shown to be specific to an *M. tuberculosis*-complex bacterial infection of the spine (Pott's disease), were identified to be even more ancient at around 5400 years old (Crubezy E, *et al.*, 1998). It is commonly though that *M. tuberculosis* evolved from saprophytic soil bacteria, which firstly found a niche infecting animals and after the domestication of cattle subsequently spread to humans (Stead, 1997).

The term tuberculosis was first described by Franciscus de la Boe (Dr. Silvius) in his *Opera Medica* of 1679, and relates to the tubercles, tuberculous cavities and tuberculous lymph nodes that are associated with the disease (http://www.wits.ac.za/myco/noframe/no_history.htm#people). The first great step forward in the combat of tuberculosis came with the recognition by an English physician, Benjamin Marten, that tuberculosis may be caused by an airborne organism (McKinney *et al.*, 1999). In his work, *A New Theory of Consumption*, published in 1722, he hypothesized that the etiological agent ".....may possibly be some certain Species of *Animalculae* or wonderfully minute living creatures that, by their peculiar Shape or disagreeable Parts are inimicable to our Nature; but, however, capable of subsisting in our Juices and Vessels......" (Dubos and Dubos, 1952). These "wonderfully minute creatures" were only first isolated more than 100 years later in the late 1800's by the German physician Robert Koch (Koch, 1882). He revealed his discovery of the identification of

Mycobacterium tuberculosis as the causative agent of tuberculosis in a historical address to the Berlin Physiological Society on March 24, 1882 (McKinney *et al.*, 1998). It is disturbing that after more than 100 years of research on the biochemistry and physiology of *M. tuberculosis*, the disease processes are still poorly understood and we still have no clear indications of what differentiates this organism from the lesser virulent and avirulent mycobacterial species (Ehlers, 1993).

1.2.3. Mycobacterium tuberculosis and the genus Mycobacterium

Mycobacteria are aerobic, rod-shaped, nonmotile bacteria characterized by being acid-alcohol fast, having complex lipid-rich cell wall structures containing mycolic acids with 60-90 carbon atoms which are cleaved to C22 to C26 fatty acid methyl esters, and by having a G+C (guanine + cytosine) DNA content of 61 to 71% (Shinnick and Good, 1994). *Mycobacterium leprae*, originally named *Bacillus leprae*, was the first mycobacterial species that was identified, its discovery by Armauer Hansen being the first convincing association of a microorganism with a human disease (Hansen, 1874, Hansen, 1880). There are currently 71 recognized species, which are divided into two groups, the slow-growing (having generation times of 1 - 14 days) and the fast-growing species (Shinnik and Good, 1994, Springer *et al.*, 1996). Most of the slow-growing mycobacteria are pathogenic species, and include the species causing tuberculosis, leprosy, paratuberculosis and other diseases, while the fast growing mycobacteria are predominantly non-pathogenic (Shinnick and Good, 1994). Closely-related genera to the genus *Mycobacterium* include *Corynebacterium*, *Nocardia*, and *Rhodococcus* and other actinomycete genera for example *Streptomyces* (Wayne and Kubica, 1986).

The *Mycobacterium tuberculosis* complex includes *M. tuberculosis*, *M. bovis* (including *bovis* BCG), *M. africanum* and *M. microti*, the first three of which are the causitive agents of tuberculosis in humans and animals, and the last of which are pathogenic only in rodents (Brosch *et al.*, 2000a). Other species investigated in the present study includes *Mycobacterium leprae*, the causative agent of the chronic neurological disease leprosy, or Hansen disease (Gelber, 1994); *Mycobacterium avium*, which causes pulmonary and nonpulmonary infections primarily in immunocompromised individuals (Havlir, 1994); *Mycobacterium avium* subsp. *paratuberculosis* or *Mycobacterium paratuberculosis*, which causes paratuberculosis or Johne's disease (which is a chronic granulomatous enteritis of ruminants, Chiodini *et al.*, 1984); and *Mycobacterium smegmatis*, which is a saprophytic, fast growing, non-pathogenic mycobacterium originally isolated from human smegma (the natural lubricant produced underneath the foreskin of the penis) in 1885 (Alvarez and Tavel, 1885).

1.2.4. Mycobacterial genomics

A major breakthrough in the research of tuberculosis came with the complete sequencing of the whole genome sequence of the widely used *Mycobacterium tuberculosis* reference laboratory strain H37Rv in 1998 (Cole *et al.*, 1998). This data revealed that the genome is around 4.4 Mb, has a high G+C content of 67% and is made up of around 4000 genes distributed evenly on both strands. Around 40% of the genes were identified to be of known function, 20% of vague function, and 40 % were totally unknown. Thus, the sequencing revealed a large number of previously unknown genes with the potential to be involved in the pathogenesis of the organism (Cole, 1998). Interestingly, approximately 51% of the genome originated due to gene duplication or domain shuffling (Tekaia *et al.*, 1999). Another very surprising observation was the fact that two unknown, novel gene families (representing areas on the genome with an exceptionally high G+C content of more than 80%), make up around 10% of the genome (Cole, 1999). These families were named the PE (99 members) and PPE (67 members) gene families (see below). The sequencing of the genome of *M. tuberculosis* presents an unprecedented opportunity to discover important genes through the newly evolved science of comparative genomics (Cole, 1998). Since the completion of the genome sequencing of *M. tuberculosis* H37Rv in 1998, the whole genome sequences of *M. tuberculosis* strain CDC1551 (Fleischmann *et al.*, manuscript in preparation), as well as *M. leprae* (Cole *et al.*, 2001), were also completed. Furthermore, the genomes of another nine mycobacterial species and strains are in the process of being sequenced (*M. tuberculosis* strain 210, *M. avium*, *M. paratuberculosis*, *M. bovis*, *M. bovis* BCG, *M. marinum*, *M. microti*, *M. ulcerans*, *M. smegmatis*). These will undoubtedly reveal novel genes that may be investigated to determine the cause of the pathogenesis of these organisms, as well as to identify novel drug targets (Brosch *et al.*, 2000a).

1.2.5. Proteases

Proteases are enzymes that cleave peptide bonds and are found in all organisms from humans to viruses where they are essential for a variety of biological processes. These processes range from nonspecific functions such as digestion and catabolism of proteins to highly specialized functions such as the specific activation of a secreted protein (Khan and James, 1998). In the human, various protease cascade systems contribute to the normal functioning of the blood coagulation and fibrinolysis pathways, regulation of sodium balance and blood pressure, and the immune response and apoptosis (Khan and James, 1998). These host-encoded proteases are tightly regulated, though, by various mechanisms such as the secretion of inhibitors, compartmentalization and rapid inactivation (Lantz, 1997). Pathogenic organisms infecting a human host have the potential to secrete proteases that not only contribute to tissue invasion and destruction, but also interfere with and upset the complex, multilevel control mechanisms of the host protease cascade systems (Goguen *et al.*, 1995). This leads to deregulation and uncontrolled activation of host derived protease zymogens, inducing inflammation and the degradation of tissue matrix, which in turn facilitates the translocation of the infecting organism (Maeda and Yamamoto, 1996). In addition to this, the bacterial proteases are also capable of uncontrolled degradation of the surrounding tissue, as their activities are not affected by the host-derived plasma protease inhibitors, which are even degraded by the foreign proteases (Travis *et al.*, 1995). In infected patients, secreted bacterial and fungal proteases may also display different pathogenic effects, including pain, edema, shock, and even septicemia (Maeda and Yamamoto, 1996).

Proteases are classified into five subclasses according to their catalytic type, namely the serine proteases, the cysteine proteases, the aspartic proteases, the metalloproteases and the unknown proteases (Barret, 1994). The serine protease subclass (making use of the amino acid serine as the catalytic residue) contains six different clans, of which the subtilases are the second largest clan (Barret and Rawlings, 1995). The functions of these proteases range from defense and nutrition, to highly specialized functions such as the processing and maturation of pro-proteins (Siezen and Leunissen, 1997). The most studied subtilases are produced by the *Bacillus* species where the secretion of subtilisin is associated with onset of sporulation, and many mutations which block sporulation at early stages affect expression levels of subtilisin. However, subtilisin is not necessary for normal sporulation. Subtilisin is also a potential virulence factor as it is able to induce the production of bradykinin (which is an endogenous peptide that causes pain, extravasation, vasodilation, hypotension, shock etc.) by the activation of Hageman factor and prekallikrein and are able to generate kinin from both low molecular weight and high molecular weight kininogen (Maeda and Yamamoto, 1996). It is thus clear that microbial proteases may exert pathological functions by not only directly destroying tissues, but also by uncontrollably activating normal host expressed cascades leading to inflammatory processes.

1.2.6. Mycobacterial proteases

M. tuberculosis has several strategies to subvert killing within the macrophage. It prevents acidification of the phagosome by exclusion of the proton ATPase (Sturgill-Koszycki *et al.*, 1994), and prevents fusion of the phagosome with lysosomes (Goren *et al.*, 1976). Furthermore, *M. tuberculosis* induces deactivation of macrophages, inefficient antigen presentation to T-cells and the secretion of suppressive cytokines, particularly TGF-β (Fenton and Vermeulen, 1996, Schluger and Rom, 1998). The mechanisms by which the organism induces these changes in macrophages and T lymphocytes are not clear, but viability of the organism is required for most of these effects. In addition to this, the exact intracellular nutrient source(s) utilized by the organisms are also not clear and it may be possible that *M. tuberculosis* has the potential to utilize host vacuolar proteins as a nutrient source by secreting host protein-degrading enzymes. All the abovementioned mechanisms and effects may be mediated by secreted proteins and/or glycolipids (Fenton and Vermeulen, 1996, Schluger and Rom, 1998). In agreement with this, *M. tuberculosis* culture filtrates have been shown to contain a number of secreted proteins, including proteases (Reich, 1981, Kannan, 1987). The effects of these proteases on the modification of the host zymogen cascades (coagulation, complement, fibrinolysis), and tissue necrosis (lung pathology observed during tuberculosis) has not been studied. Although the presence of protease activity in mycobacterial culture filtrates has been known since the early 1980's (Reich, 1981), only two specific proteases have so far been identified in the *M. tuberculosis* culture filtrates, both belonging to the chymotrypsin clan of serine proteases (Skeiky *et al.*, 1999). The genome of *M. tuberculosis* contains around 70 potential proteases, the majority of which have not been described and many of which may contribute to the pathogenic mechanisms observed during *M. tuberculosis* infection.

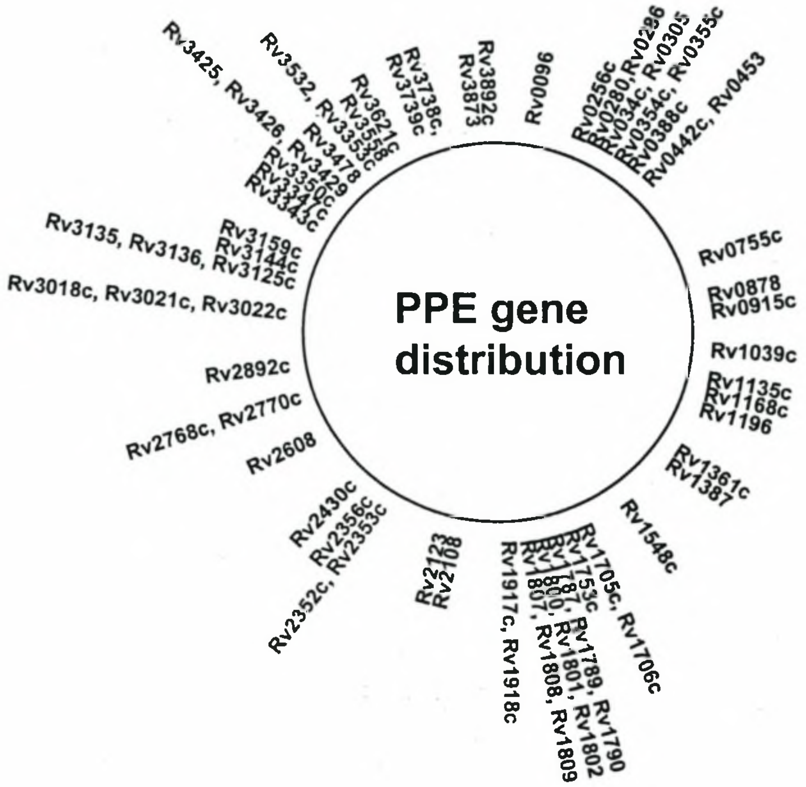
1.2.7. Secreted antigens - the ESAT-6 gene family

Proteins released from live, actively dividing bacteria have attracted considerable interest from researchers interested in tuberculosis vaccine development, as it has been shown that only live and replicating, as opposed to killed, mycobacterial preparations have the ability to generate and recall protective immunity to *M. tuberculosis* (Orme, 1988a, Orme, 1988b, Andersen *et al.*, 1991a, Orme *et al.*, 1993). The importance of the secreted antigens is highlighted by the fact that complex mixtures of secreted *M. tuberculosis* proteins have been shown to induce high levels of protection in animal models of tuberculosis (Hubbard *et al.*, 1992, Pal and Horwitz, 1992, Andersen, 1994b, Roberts *et al.*, 1995). The low molecular weight culture filtrate proteins (3 to 9 kDa) have been shown to induce the highest levels of T-cell proliferation and gamma interferon production in mice (Andersen and Heron, 1993). Interestingly, the dissection of the low molecular mass culture filtrate fraction has led to the identification of a number of small proteins belonging to the ESAT-6 gene family (Sørensen *et al.*, 1995, Berthet *et al.*, 1998, Alderson *et al.*, 2000, Skjøt *et al.*, 2000, Rosenkrands *et al.*, 2000a). This family encodes small proteins of unknown function, which are secreted by an unknown mechanism, as they do not contain the ordinary secretion signals in their protein sequences, and all share a remarkable level of immunodominance (Skjøt *et al.*, 2001). Several studies have linked the deletion of these genes with a decrease in virulence, indicating that these genes may play an important role in intracellular survival (Mahairas *et al.*, 1996, Wards *et al.*, 2000).

1.2.8. PE and PPE

Several areas with an exceptionally high G+C content of more than 80 % were identified on the genome sequence of *M. tuberculosis* (Cole, 1998). These regions were found to contain multiple copies of the PGRS (polymorphic G+C rich sequence) sequences. The PGRS along with the MPTR (major polymorphic tandem repeat) sequences were originally described as non-coding repetitive sequences (CGGCGGCAA and GCCGGTGTTG, respectively) in the genome of *M. tuberculosis* (Hermans *et al.*, 1992, Ross *et al.*, 1992, Poulet and Cole, 1995). After the completion of the genome sequence of *M. tuberculosis*, it was shown that these two groups of repetitive sequences actually encode for genes belonging to subgroups of the acidic, glycine-rich PE and PPE gene families (Cole *et al.*, 1998). The PE and PPE gene families are two large multigene families of unknown function, comprising around 10 % of the genome of *M. tuberculosis* and containing 99 and 68 members, respectively (see Figure 1.1 for distribution of the PPE gene family). The PE family is characterized by the presence of a proline-glutamic acid (PE) motif at positions 8 and 9 in a very conserved N-terminal domain of around 110 amino acids (Gordon *et al.*, 1999b) and is divided into two subgroups of which the polymorphic GC-rich sequence (PGRS) subgroup is the largest and contains proteins with multiple tandem repeats of a glycine-glycine-alanine or a glycine-glycine-asparagine motif in the C-terminal domain. The other subgroup consists of proteins with C-terminal domains of low homology.

Figure 1.1. Distribution of the PPE gene family on the genome sequence of *M. tuberculosis* H37Rv. Reproduced with kind permission of S. Sampson (Department of Medical Biochemistry, Faculty of Health Sciences, University of Stellenbosch, South Africa).



Similarly, the PPE family also contains a highly conserved N-terminal domain of around 180 amino acids, with a proline-proline-glutamic acid (PPE) motif at positions 7-9 (Cole *et al.*, 1998) and can be divided into three subgroups (Gordon *et al.*, 1999b) of which the major polymorphic tandem repeat (MPTR) subgroup is the largest. The proteins of this subgroup contain multiple repeats of the motif AsnXGlyXGlyXAsnXGly encoded by a consensus repeat sequence GCCGGTGTTG, seperated by 5 bp spacers (Cole and Barrell, 1998). The PPE-SVP subgroup is characterized by the motif

GlyXXSerValProXXTrp at position 350 in the amino acid sequence and the last subgroup consists of proteins with a low percentage of homology at the C-terminus (Gordon *et al.*, 1999b). The C-terminal domains of both these protein families are of variable size and sequence (the MPTR subgroup contains proteins consisting of more than 3000 amino acid residues and the PGRS subgroup proteins may contain up to 1400 amino acids)(Cole, 1998). These domains also contain repeat sequences of different copy numbers in a number of cases (Gordon *et al.*, 1999b), thus showing extensive polymorphisms in the different *M. tuberculosis* complex strains, which is probably due to strand slippage during replication (Cole, 1998).

The 167 members of the PE and PPE gene families are of unknown function, but it has been suggested that the proteins encoded by these gene families may inhibit antigen processing or may be involved in antigenic variation as the size and sequence variation would hold relevance in the evasion of the host immune response (Cole *et al.*, 1998, Cole, 1998, Cole, 1999, Gordon *et al.*, 1999b). In agreement with this, sequence variation has been observed between the orthologues of the PE and PPE protein families in an *in silico* analysis of the genomes of *M. tuberculosis* H37Rv and *M. bovis* (Cole *et al.*, 1998, Gordon *et al.*, 2001). Extensive variation of a subset of PPE genes in clinical isolates of *M. tuberculosis* have also been observed recently (S. Sampson, submitted for publication). These are all theories, though, and no conclusive experimental evidence has been provided to date which would indicate the function(s) of these polymorphic proteins. Several studies have highlighted different aspects of selected members of the two families. For example, Rodriguez and colleagues (1999) have found that the PPE gene Rv2123 is upregulated under low iron conditions, leading to the hypothesis that it may encode a siderophore involved in iron uptake. Recent data have suggested that the members of the PPE gene family may be involved in disease pathogenesis, as a transposon mutant of the gene Rv3018c was attenuated for growth in macrophages (Camacho *et al.*, 1999). Others have shown that the PE-PGRS subgroup may be a novel family of fibronectin binding proteins (Abou-Zeid *et al.* 1991, Cole *et al.*, 1998, Espitia *et al.*, 1999). Furthermore, it was recently shown that two members of the PGRS subfamily from *M. marinum* are essential for replication in macrophages as well as persistence in granulomas (Ramakrishnan *et al.*, 2000). The fact that these genes encode for about 4% of the total protein species in the organism (if all genes are expressed), indicates that they most probably fulfill an important function or functions in the organism.

1.3. Study Aims and Design

This dissertation is divided into eight chapters and six addenda and consists of an introductory review chapter (Chapter 1), one chapter and three addenda dealing with the mycosin proteases (Chapter 2, Addenda 2A, B and C), three chapters and three addenda dealing with the ESAT-6 gene clusters (Chapter 3, 4, and 5, Addenda 3A, B and 5A) and one chapter dealing with the PE and PPE gene families (Chapter 6). This is followed by a discussion of the results and future directions (Chapter 7), and a conclusions chapter (Chapter 8). The last section contains the list of references.

The aims of this study can be divided in three major parts, summarized as follows:

The mycosin proteases were investigated with the aim of:

(i) cloning, expressing and characterizing the family of mycobacterial subtilases (Chapter 2),

(ii) determining where these proteins are located (Addendum 2A),

(iii) determining the antigenicity of the secreted proteases (Addendum 2B),

(iv) determining the substrate and optimal activity conditions of the proteases (Addendum 2C),

(v) identifying the genetic milieu in which they are situated in the *M. tuberculosis* genome (Chapter 3).

The ESAT-6 gene clusters were investigated with the aim of:

(i) identifying the genes other than the mycosin proteases which are present within the clusters (Chapter 3),

(ii) determining the distribution of the clusters within the genomes of different mycobacterial species, as well as in the genomes of other bacterial species (Chapter 3),

(iii) identifying the most ancient progenitor gene cluster by systematic phylogenetic analyses (Chapter 3),

(iv) establishing whether the gene clusters are expressed as one or more operons (Chapter 4),

(v) identifying a promoter responsible for the expression of the clusters (Chapter 4),

(vi) examining the putative ESAT-6 multi-component secretion system function of the cluster-encoded proteins (Chapter 5),

(vii) establishing a putative function for the secretion system from the literature (Addendum 5A).

The PE/PPE gene families were investigated with the aim of:

(i) determining whether the presence of PE and PPE copies within the ESAT-6 gene clusters have any significance (Chapter 6),

(ii) determining whether there is an association between the expansion of the PE and PPE gene families and the duplication of the ESAT-6 gene clusters (Chapter 6),

(iii) discovering their evolutionary history (Chapter 6).

Chapters 2 to 6 are presented as published papers (Chapters 2 and 3) or as manuscripts in preparation (Chapters 4, 5 and 6 and Addenda 3A, 3B and 5A). It is therefore likely that repetition may occur in the introductions of many of these chapters. All cited literature has been included in a single list of references in the last section of this dissertation.

CHAPTER TWO

THE MYCOSINS

"It was no accident that man mastered the plague more easily than tuberculosis. The one comes in terrible waves of death that shake humanity to the foundations, the other slowly and stealthily; the one leads to terrible fear, the other to gradual indifference. The consequence is that man opposed the one with all the ruthlessness of his energy, while he tries to control consumption with feeble means. Thus he mastered the plague, while tuberculosis masters him."
Mein Kampf – Adolf Hitler

NOTE: The results presented in the following chapter were published as: "**The mycosins of *M. tuberculosis* H37Rv: A family of Subtilisin-Like Serine Proteases.** Brown, G.D., Dave, J.A., Gey van Pittius, N.C., Stevens, L., Ehlers, M.R.W., and Beyers, A.D., *Gene*, 2000, Aug 22, 254 (1-2): 147-155.

(The style of the text and numbering of sections has been altered to conform to the style of this dissertation. All cited literature is compiled into a single list at the end of the dissertation for ease of reference)

2.1. Introduction

Mycobacterium tuberculosis is the leading cause of bacterial related deaths, killing more than two million people per year (Murray and Salomon, 1998). The bacterium is an intracellular pathogen of macrophages and is able to modify the maturation of the phagosome within which it resides, allowing the bacterium to bypass the microbicidal effector functions of the host cell (Clemens and Horwitz, 1995). The mechanisms responsible for these modifications and the other factors contributing to the pathogenesis of the organism are largely unknown. With the control of tuberculosis hampered by the duration required for successful treatment, drug resistance and the increased susceptibility of patients with HIV/AIDS, it is hoped that the elucidation of these virulence factors as well as other key processes essential for the survival of the organism will provide new strategies to deal with this pathogen.

Proteolysis represents one essential function which has been largely ignored in the study of *M. tuberculosis* and in other mycobacteria. Proteases serve many roles in bacteria, ranging from the turnover and modification of cellular proteins to virulence factors in pathogens, and the targeting of proteolytic enzymes is a strategy showing promise in the control of other bacterial pathogens (Miyagawa *et al.*, 1991; Travis *et al.*, 1995). Despite this, only a few proteases have been examined in mycobacteria, including the 20S proteasome and the Lon protease from *M. smegmatis* (Knipfer and Shrader, 1997; Roudiak et al., 1998), the *M. leprae* Clp protease ATPase subunit, ClpC, (Misra et al., 1996) and HtrA in *M. avium* subsp. *paratuberculosis* (Cameron et al., 1994). The genome of *M. tuberculosis* H37Rv encodes over 30 proteases (Cole et al., 1998) and it is therefore surprising that so little is known about the biology of these enzymes in this organism. Only two secreted serine proteases (MTB32A and MTB32B), FtsH and an unidentified elastolytic metalloprotease have been studied in *M. tuberculosis* (Rowland et al., 1997; Anilkumar et al., 1998; Skeiky et al., 1999).

The paucity of information regarding the role of proteolysis in the biology of *M. tuberculosis* and mycobacteria in general, the possibility that proteases are involved in pathogenesis, and the likelihood that these enzymes may provide alternative drug targets, motivated the investigation described here. In this study we characterised a group of proteins, the mycosins, whose primary sequence features strongly suggest that they are transmembrane serine proteases of the subtilisin family. The expression, cellular localisation and presence of the mycosins in *M. tuberculosis* and in other mycobacteria were examined.

2.2. Materials and Methods

2.2.1. Bacterial strains and growth conditions

Escherichia coli strains were grown in LB broth with 0.2% glucose and the appropriate antibiotics at 37°C, unless otherwise indicated. *M. tuberculosis* H37Rv (ATCC 25618) and *M. bovis* BCG Tokyo (State Vaccine Institute, Cape Town), were grown at 37°C in 7H9 broth (Difco) with OADC (0.5% BSA, 0.2% glucose, 0.006% oleic acid, 140mM NaCl) and 0.05% Tween 80 (Sigma) with stirring in 1-liter screw-cap bottles. All work on *M. tuberculosis* H37Rv was performed in a Biosafety Level III facility. *M. smegmatis* mc^{2}155 (Snapper et al., 1990) cultures were grown in 7H9 broth, as described, or in Sauton's medium with 0.05% Tween 80 (Connell, 1994) at 37°C and were agitated by shaking (200 rpm). Hygromycin (0.1 mg/ml; Boehringer Mannheim) was added to cultures of *M. smegmatis* transformed with the various p19Kpro constructs (see below).

2.2.2. DNA constructs, plasmids and methods

All standard molecular techniques were performed essentially as described by Sambrook et al. (1989). Transformation *of M. smegmatis* was performed using electroporation, as described (Jacobs et al., 1991). Details of the cosmids, plasmids, plasmid constructs, genes and probes with the sequence positions relevant to their construction are presented in Table 2.1. Constructs were generated by appropriate ligations of the relevant cosmid fragments into the vectors except for the construction of p19K-P2RBS, where PCR was utilised. All essential constructs were verified by DNA sequencing. Construction of p19K-P2RBS involved the addition of a RBS 10 bp upstream from the predicted start codon. Two primers were used to amplify the 5' portion of the *mycP2* gene containing the start codon (shown in bold): 5'-aatagatctgca**ATG**gcttcgccactaaac-3' and 5'-aataagcttgtactgccacgccttgc-3'. To add the RBS, the *Bgl*II / *Hind*III digested PCR product was ligated into the *BamHI* / *Hind*III sites of the pRBS1 vector, which contained the RBS. The *mycP2* gene was then reconstituted and ligated back into the p19Kpro vector. pRBS1 was generated by ligating an adaptor fragment into the *Eco*RI / *Bam*HI site of pUC18. The oligonucleotides used to generate the adaptor were: 5'-aattcagatctAAAGGAGGag-3' and 5'-gatcctcctttagatctg-3'. The RBS sequence chosen (AAAGGAGG) was based on the 16s RNA sequence of *M. tuberculosis* (Genbank accession number X58890). Primers used to generate the PCR probes for *mycP4* and *mycP5* from H37Rv

Table 2.I. Sequence positions and accession numbers of the *mycP* genes

Plasmid / Gene (accession number)	Description	Relevant positions on genomic sequence	Source or reference
Cloning vectors			
pGex2T	*E. coli* expression system (GST fusion, Amp^r)		Pharmacia
pMalC	*E. coli* expression system (MBP fusion, Amp^r)		New England Biolabs
p19Kpro	Mycobacterial expression system derived from p16R1 with the constitutive 19kD antigen promoter ($Hygromycin^r$)		Gift from K. de Smet, Imperial College, London (Garbe *et al.*, 1994).
pRBS1	pUC18 carrying a mycobacterial ribosome binding site (Amp^r)		This study
Mycosin-1 (Z94121)			
mycP1	mycosin-1 ORF	4363414 to 4364754^C	(Cole *et al.*, 1998)
pGex-P1	*mycP1* cloned without sequence encoding signal peptide in pGex2T	4363414 to 4364703^C	(J.A. Dave *et al.*, submitted for publication)
p19K-P1	*mycP1* cloned into p19Kpro	4363414 to 4364822^C	(J.A. Dave *et al.*, submitted for publication)
mycP1 probe	DNA probe for Southern blotting	4363638 to 4364082	This study
Mycosin-2 (Z94121)			
mycP2	mycosin-2 ORF	4368515 to 4370167^C	(Cole *et al.*, 1998)
pMalC-P2	*mycP2* cloned without sequence encoding signal peptide in pMalC	4367943 to 4370037^C	This study
p19K-P2	*mycP2* cloned into p19Kpro	4367943 to 4370277^C	This study
p19K-P2RBS	*mycP2* with a RBS cloned into p19Kpro	4367943 to 4370167^C	This study
mycP2 probe	DNA probe for Southern blotting	4368712 to 4369515	This study

Mycosin-3 (AL021930)			
mycP3	mycosin-3 ORF	354496 to 355881	(Cole *et al.*, 1998)
pGex-P3	*mycP3* cloned without sequence encoding signal peptide in pGex2T	354573 to 356056	This study
p19K-P3	*mycP3* cloned into p19Kpro	354320 to 356320	This study
mycP3 probe	DNA probe for Southern blotting	355279 to 355743	This study
Mycosin-4 (Z95390)			
mycP4	mycosin-4 ORF	3869748 to 3871115	(Cole *et al.*, 1998)
mycP4 probe	PCR generated probe for Southern blotting	3870432 to 3870898	This study
Mycosin-5 (AL022021)			
mycP5	mycosin-5 ORF	2033727 to 2035484	(Cole *et al.*, 1998)
mycP5 probe	PCR generated probe for Southern blotting	2034733 to 2035373	This study

[c]complementary sequence

genomic DNA (a gift from J.T. Belisle, Colorado State University) were: 5'-aagaacgccgtcatcgtg-3' and 5'-gaatcgagtcgctgctga-3'; 5'-gtgctcgtaatgtcatcg-3' and 5'-catatcggcaccatatcg-3', respectively.

2.2.3. *Protein methods*

Expression of the protease-fusion proteins in *E. coli* was induced by the addition of isopropyl-ß-D-thiogalactoside (0.3mM final concentration) to mid-logarithmic phase cultures. The cultures were then incubated for a further 2 to 3 hours at 25°C following which the cells were harvested and disrupted by sonication. Depending on the fusion partner, the fusion proteins were purified by affinity chromatography using amylose resin (New England Biolabs) or glutathione-agarose (Sigma), as described by the manufacturers. Purified protease-fusion proteins were dialysed against PBS (Sambrook et al., 1989) before use. To obtain mycobacterial protein extracts for use in Western blotting experiments, cells were harvested and washed twice in 0.05% Tween 80. The cells were resuspended in H_20 and SDS (Sigma) was added to a final concentration of 2%. The cells were then sonicated for 5 min and boiled for 5 min. SDS-PAGE sample buffer (Sambrook et al., 1989) was added and the samples boiled for 15 min. All samples were stored at -20°C. Cellular fractions of *M. smegmatis* transformed with the various p19Kpro constructs were obtained following a modified protocol of Raynaud et al., (1998). *M. smegmatis* cultures were grown to early stationary phase (OD_{600nm}~1) in 100 ml of Sauton's medium. The cells were harvested (3000 *x g* for 10 min), washed twice in PBS, resuspended in 5 ml of PBS and disrupted by sonication. After clearing the unlysed cells (3000 *x g* for 15 min, twice) the whole cell lysate (wcl) fraction was subjected to high speed centrifugation (100 000 *x g* for 2 hr) to separate out the cytoplasmic fraction (supernatant) and membrane/cell wall fraction (pellet). The membrane/cell wall fraction was resuspended in PBS containing 0.33% NP40 (Sigma). The extracellular medium was concentrated to approximately 5 ml, using an Amicon PM10 ultrafiltration system, and dialysed against PBS. NaN_3 (5 mM) was added to all the fractions to inhibit bacterial growth. The purity of the fractions was verified using the cytoplasmic marker, isocitrate dehydrogenase, as described (Raynaud et al., 1998).

2.2.4. *Generation of antiserum*

Polyclonal antibodies to three purified fusion proteins (mycosin-1, -2 and -3) were obtained from immunized New Zealand white rabbits. Briefly, the rabbits were immunized subcutaneously with the *E. coli*-derived protease fusions in Freund's incomplete adjuvant. Similar booster immunizations were given after four weeks, and then every two to three weeks thereafter until acceptable titres were reached. Antiserum was stored in 50% glycerol at -20°C. Antisera against mycosin-1 and mycosin-2 were depleted of antibodies cross reactive to other mycobacterial antigens by the addition of sonicated *M. smegmatis* cellular lysates before use in Western blotting experiments. Antiserum to mycosin-3 was not depleted because of the possibility of a mycosin-3 homologue in *M. smegmatis*. The polyclonal antisera was not cross reactive between the mycosins.

2.3. Results and Discussion

2.3.1. Identification of the mycosins

We initially identified the first mycosin by analysis of a bank of partial sequences generated using a PhoA fusion system, which selects for genes encoding secreted proteins (data not shown; Mdluli et al., 1995). The genes encoding the four other members of this family were subsequently identified in the *M. tuberculosis* H37Rv genomic sequence through similarity (BLAST) searches (Altschul et al., 1997). The identity of these proteins ranged from 36% to 47%, suggesting that they probably arose through gene duplication. The proteins were designated mycosin-1 to -5, based on their sequential identification, and are annotated as Rv3883c, Rv3886c, Rv0291, Rv3449 and Rv1796, respectively, on the *M. tuberculosis* H37Rv genomic sequence (Cole *et al.*, 1998). A dendrogram of all annotated Mycobacterial proteases, including the mycosins and a related serine protease, is shown in Figure 2.1. The genes encoding all five mycosins (*mycP*) reside in high-density protein-encoding regions on the genome, and *mycP1* and *mycP2* are separated by only 3.7kbp containing two ORFs. Although the functions of these two ORFs and of the other genes surrounding *mycP1 - 5* are unknown (Cole et al., 1998), each *mycP* gene is located close to or next to a putative transmembrane transporter (Rv3877, Rv3887c, Rv0290, Rv3448 and Rv1795), but the significance of this association, if any, is unclear.

2.3.2. Primary sequence characteristics and gene distribution in mycobacteria

The mycosins have a number of conserved features in the primary amino acid sequence including the catalytic residues, the hydrophobic N termini, and the hydrophobic regions near the C termini (Figure 2.2). The catalytic triad (Asp90, His121 and Ser332; mycosin-1 numbering), within conserved sequences, is typical of bacterial serine proteases of the subtilisin family (peptidase family S8 (Rawlings and Barrett, 1993, pyrolysin subfamily - Siezen and Leunissen, 1997). Furthermore, the *mycP* genes possess all three active site signatures described in the Prosite database, giving them a 100% probability of encoding a serine protease from the subtilase family (www.expasy.ch/cgi-bin/prosite-search-ac?PDOC00125). Thus based on the overall similarity of the mycosins to other proteases, the highly conserved nature of the proposed catalytic residues and similarities in surrounding sequences, the *mycP* genes are likely to encode subtilisin proteases.

Figure 2.1. Dendrogram of all annotated *Mycobacterium tuberculosis* H37Rv proteases. Mycobacterial proteases were identified by Entrez protein database searches (http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Protein), searching for the terms protease, proteinase or peptidase. The proteins are named as designated on the genomic sequence (Cole *et al.*, 1998) and the gene name, where known, is shown in italics. The proteins were aligned using ClustalW (Higgins and Sharp, 1988) and the dendrogram generated using Treeview (Page, 1996). Also shown, for comparison, is Bacillus sp. NKS-21 subtilisin ALP1 (underlined; Genbank accession number D29736). The bar represents 10% sequence divergence. Vertical distances are arbitrary.

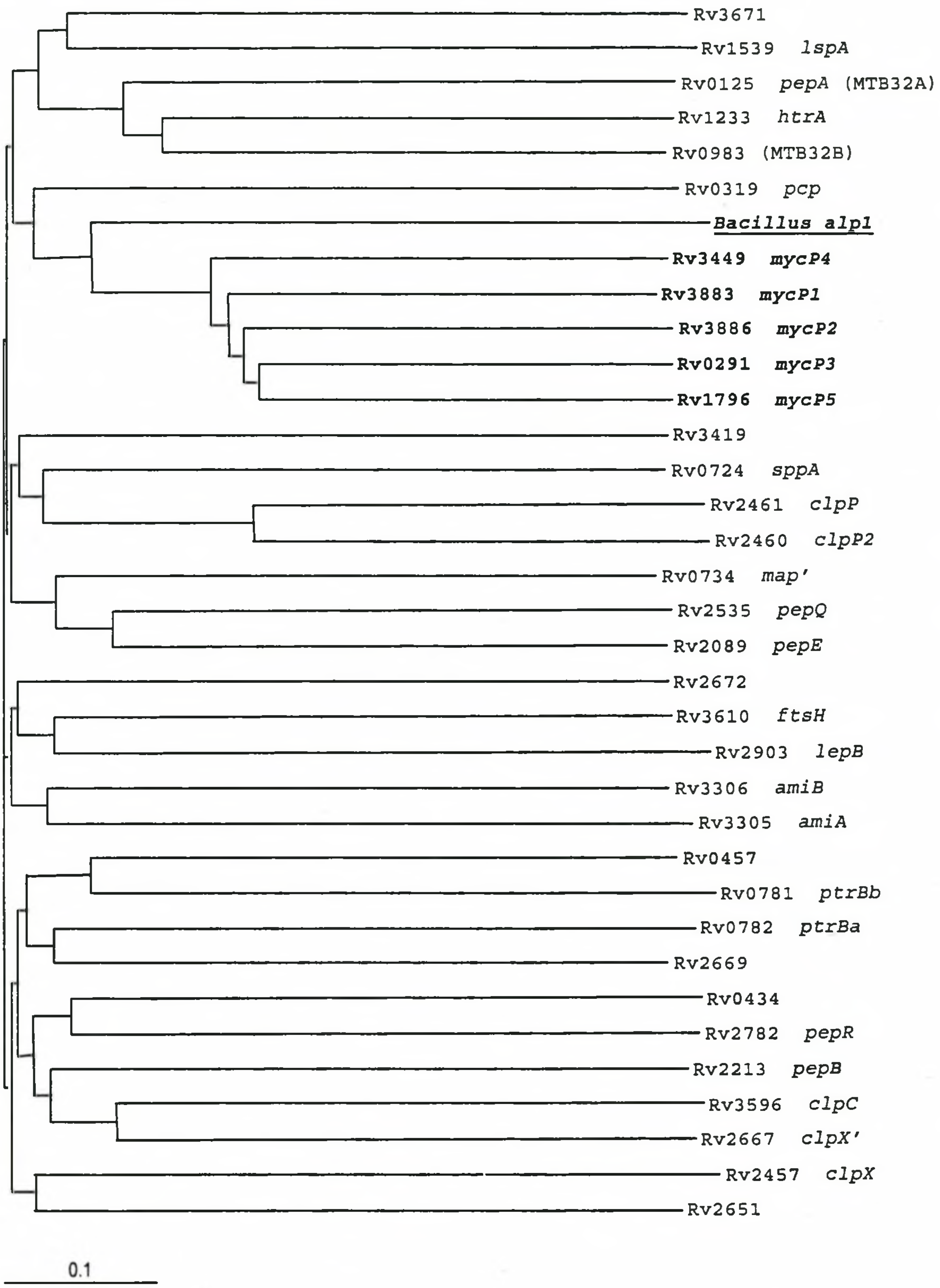
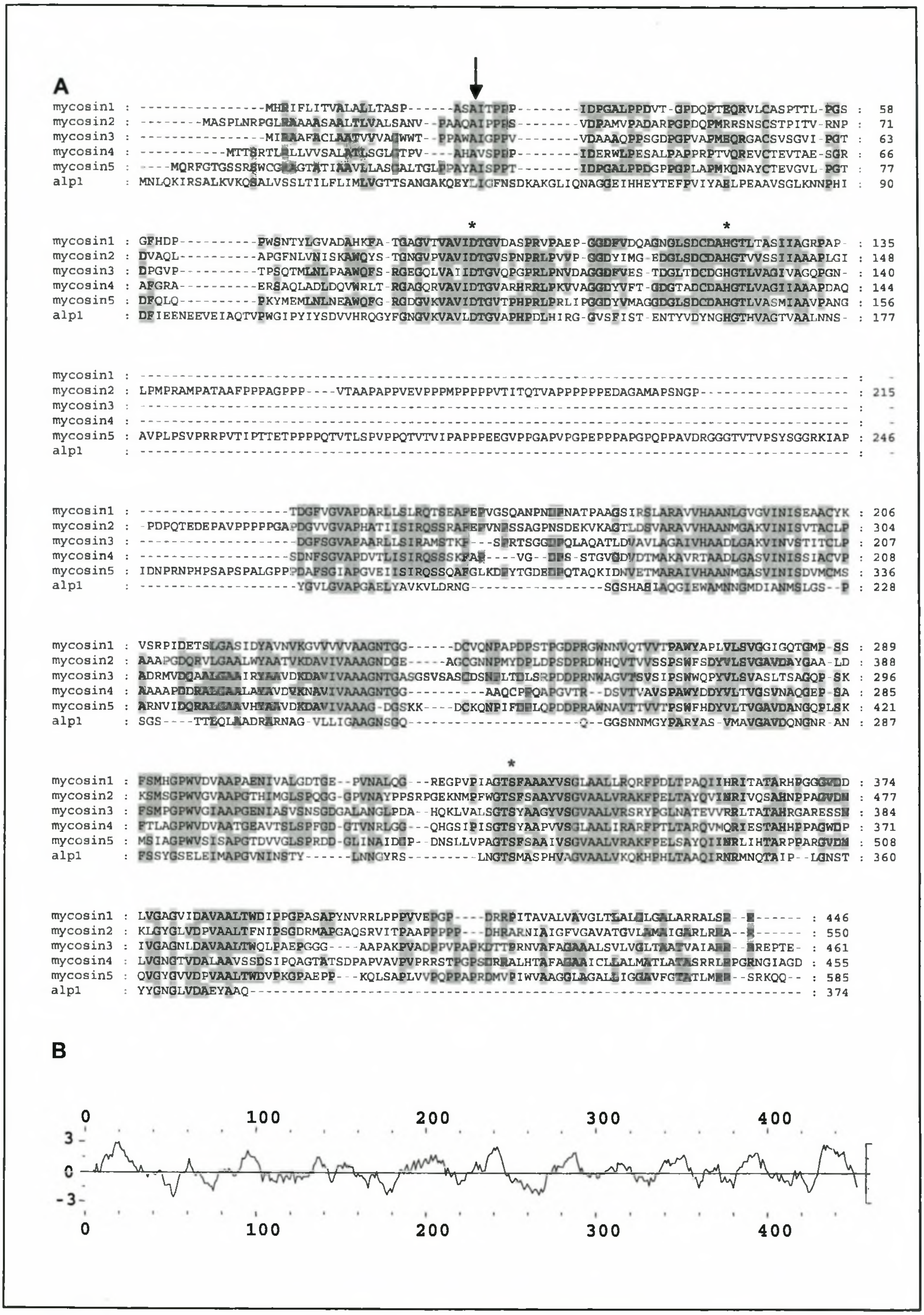

Rv3671
Rv1539 lspA
Rv0125 pepA (MTB32A)
Rv1233 htrA
Rv0983 (MTB32B)
Rv0319 pcp
Bacillus alp1
Rv3449 mycP4
Rv3883 mycP1
Rv3886 mycP2
Rv0291 mycP3
Rv1796 mycP5
Rv3419
Rv0724 sppA
Rv2461 clpP
Rv2460 clpP2
Rv0734 map'
Rv2535 pepQ
Rv2089 pepE
Rv2672
Rv3610 ftsH
Rv2903 lepB
Rv3306 amiB
Rv3305 amiA
Rv0457
Rv0781 ptrBb
Rv0782 ptrBa
Rv2669
Rv0434
Rv2782 pepR
Rv2213 pepB
Rv3596 clpC
Rv2667 clpX'
Rv2457 clpX
Rv2651
0.1

Figure 2.2. Conserved features of the serine protease family. (A) ClustalW alignment (Higgins and Sharp, 1988) of the five serine proteases and *Bacillus* sp. NKS-21 subtilisin ALP1 (Genbank accession number D29736) showing the residues thought to comprise the catalytic triad (D90, H121 and S332, mycosin-1 numbering; indicated by asterisks). The putative signal peptide cleavage sites are indicated by an arrow. (B) Hydropathy plot of mycosin-1 showing the hydrophobic N terminus (comprising the signal sequence) and the hydrophobic C terminus (the transmembrane region), followed by charged residues. The hydropathy plots for the other proteases are similar.


A
mycosin1
mycosin2
mycosin3
mycosin4
mycosin5
alp1
*
B
0
100
200
300
400
3
0
-3

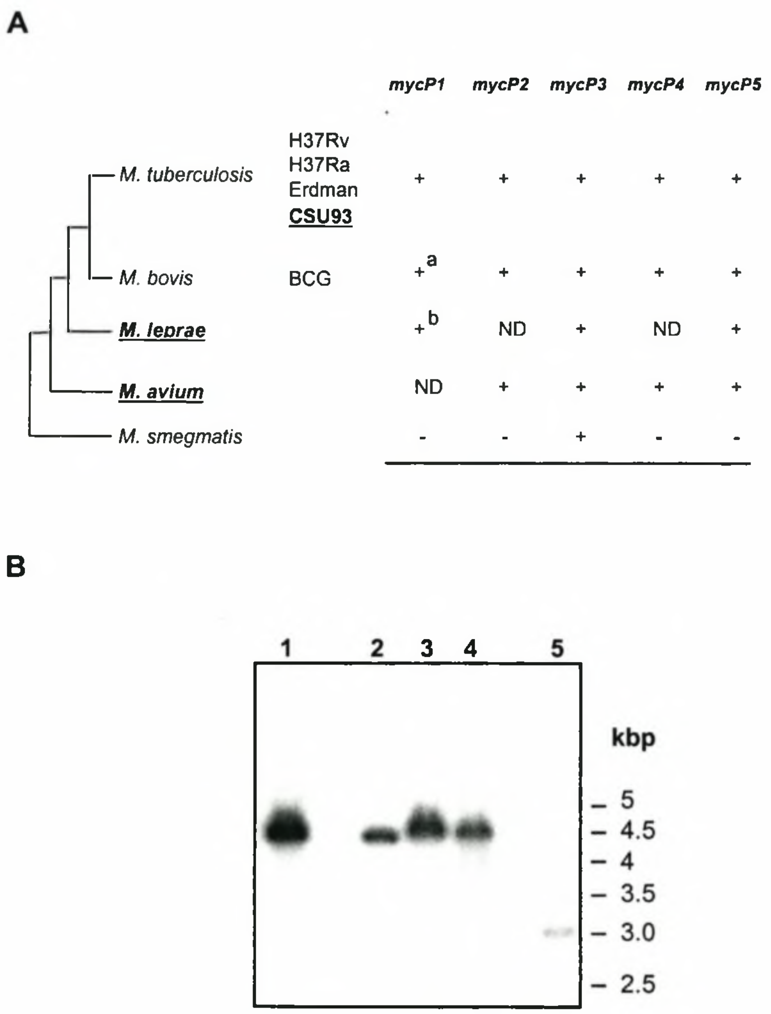
The hydrophobic N termini are probably signal peptides, which are predicted to be cleaved within a conserved sequence (AXA^I or AXA^V; http://www.cbs.dtu.dk/services/SignalP/). The cleavage sites for all the proteins were predicted as indicated, except mycosin-3 which was predicted to be cleaved at AAA^QP. As the cleavage site appears to be highly conserved, we assume that mycosin-3 will also be cleaved at the indicated position. These proteins also contain hydrophobic stretches followed by charged residues at the C termini, suggestive of transmembrane domains. Furthermore, this region was predicted to be a transmembrane domain using TMpred (data not shown; http://www.ch.embnet.org/software/TMPRED_form.html). A proline-rich linker connects the enzymatic domain to the transmembrane sequence and mycosin-2 and mycosin-5 each contains an additional, but dissimilar, highly proline-rich segment which has not been observed before in that position in any subtilase.

The multiplicity of the mycosins in *M. tuberculosis* prompted an examination of their distribution in other mycobacterial species (Figure 2.3). All the *mycP* genes were present in *M. bovis* as well as *M. bovis* BCG. *mycP1*, *mycP3* and *mycP5* homologues were detected in the incomplete genome sequence of *M. leprae* and *mycP2-5* in the incomplete sequence of *M. avium*. Only *mycP3* was detected in the avirulent *M. smegmatis* and although it appears that the multiplicity of the *mycP* genes may occur only in virulent mycobacteria, it was possible that other, more divergent, *mycP* genes were present which were not detected with the methods used.

2.3.3. Expression and localisation of the mycosins in M. smegmatis

We initially attempted to express the mycosins in *E. coli* without their signal sequences and with N-terminal fusion partners, glutathione-S-transferase or the maltose binding protein, for ease of purification. Although full-length protein fusions were observed the majority of the protein was expressed in truncated and/ or degraded form (see positive controls in Figure 2.5). Nevertheless, antisera raised against these *E. coli*-derived fusion proteins were specific for the mycosins, as verified by the detection of these proteins when heterologously expressed in *M. smegmatis* (see below, Figure 2.4).

Figure 2.3. Distribution of the protease genes in various mycobacterial species. (A) The presence of the protease genes was determined by Southern blotting except for those that are underlined, which were identified through similarity (BLAST; (Altschul *et al.*, 1997)) searches. The CSU93, *M. bovis*, *M. leprae* and *M. avium* sequence data were obtained through early release of data from The Institute for Genomic Research (http://www.tigr.org) and from the Sanger Centre (http://www.sanger.ac.uk). The phylogenetic relationship between the various species was based on 16s RNA, as described (Wang *et al.*, 1995). The Genbank accession numbers of the genes indicated a and b are: U34848 and Y14967, respectively. ND, not determined. (B) Southern blot of DNA from various mycobacterial species probed with *mycP3* and showing the presence of a mycosin 3 homologue in *M. smegmatis.* Lanes (1) *M. bovis* BCG, (2) *M. tuberculosis* Erdman, (3) *M. tuberculosis* H37Ra, (4) *M. tuberculosis* H37Rv, and (5) *M. smegmatis*.



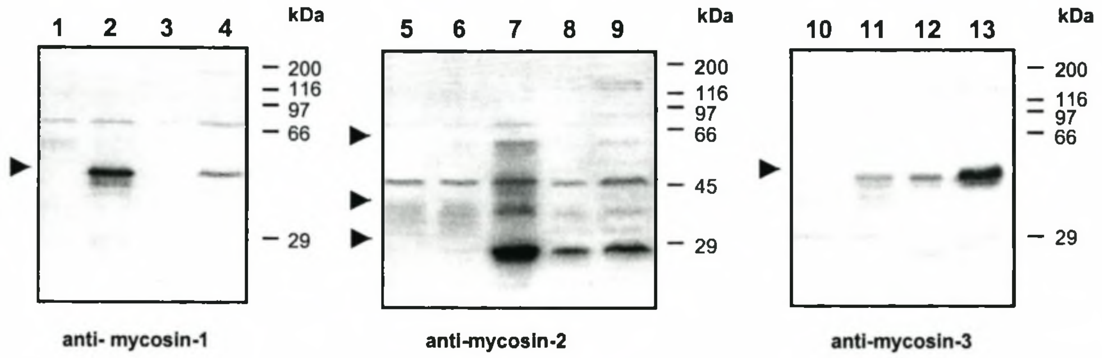
		mycP1	*mycP2*	*mycP3*	*mycP4*	*mycP5*
M. tuberculosis	H37Rv H37Ra Erdman **CSU93**	+	+	+	+	+
M. bovis	BCG	+[a]	+	+	+	+
M. leprae		+[b]	ND	+	ND	+
M. avium		ND	+	+	+	+
M. smegmatis		-	-	+	-	-

We then expressed three of the mycosins in *M. smegmatis*, by cloning the *mycP* genes in a shuttle plasmid downstream of the constitutive 19kDa antigen promoter. Expression of mycosin-1 (J.A. Dave *et al.*, submitted for publication) and mycosin-3 in *M. smegmatis* was achieved by cloning the entire ORF and more than 50 bp of the 5' non-coding region, containing any putative translation signals, downstream of the 19kDa promoter. The mycosin-1 and mycosin-3 products were of the predicted size (~50 kDa; Figure 2.4). Although *M. smegmatis* possesses a *mycP3* homologue, mycosin-3 was not detected in wild type *M. smegmatis* by Western blotting (Figure 2.4).

Expression of *mycP2* was not obtained in this vector, despite the inclusion of 110 bp of the 5' non-coding region, and was only achieved after the inclusion of a RBS 10 bp upstream of the proposed start codon. The full-length protein (~65 kDa) was cleaved into a fragment of ~36 kDa and a more predominant fragment of ~29 kDa. Although the full length protein was predicted to be ~55 kDa, mycosin-2 contains a highly proline-rich segment and it is known that proline-rich proteins can migrate anomalously on SDS-PAGE (See and Jackowski, 1989).

As the primary sequence of these proteases suggested that the enzymes were extracellularly located and anchored in the membrane, we determined the location of the heterologously expressed proteases in cellular fractions of *M. smegmatis* by Western blotting (Figure 2.5). Mycosin-1, -2 and -3 were all localised to the cell wall/membrane fraction, supporting the proposed functions of the signal peptide and transmembrane C-terminal regions. The mycosins did not appear to be released into the growth medium. Although only one fragment is likely to possess the transmembrane anchor, both mycosin-2 fragments were present in the cell wall/ membrane fraction. The localisation of both fragments in the membrane of *M. smegmatis* indicates that the two fragments were associating in some way but, despite the presence of a number of cysteine residues throughout the protein sequence, disulphide linkages between the two fragments could not be detected by analysis on non-reducing SDS-PAGE (data not shown).

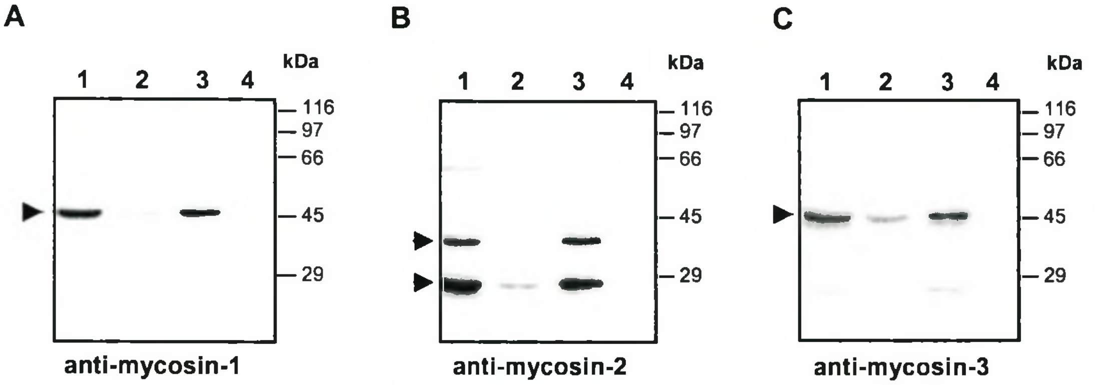
Figure 2.4. Expression of mycosin-1, mycosin-2 and mycosin-3 in various mycobacterial species. Western blots were probed with anti-protease antisera, as indicated. The protease bands are indicated by arrowheads. The *M. tuberculosis* and *M. bovis* BCG samples were taken from log-phase cells (OD_{600nm} vs H_2O 0.6-0.7). Heterologous expression of all the proteases in *M. smegmatis* was achieved using the p19Kpro expression vector except for mycosin-2 whose expression required the addition of a RBS, as described. *M. smegmatis* transformed with the vector alone was used as a negative control. Lanes (1) *M. smegmatis* (p19kpro), *M. smegmatis* (p19k-P1), (3) *M. bovis* BCG, (4) *M. tuberculosis* H37Rv, (5) *M. smegmatis* (p19kpro), (6) *M. smegmatis* (p19k-P2), (7) *M. smegmatis* (p19k-P2RBS), (8) *M. bovis* BCG, (9) *M. tuberculosis* H37Rv, (10) *M. smegmatis* (p19kpro), (11) *M. smegmatis* (p19k-P3), (12) *M. bovis* BCG, (13) *M. tuberculosis* H37Rv.



2.3.4. Mycosin expression in M. tuberculosis and M. bovis BCG

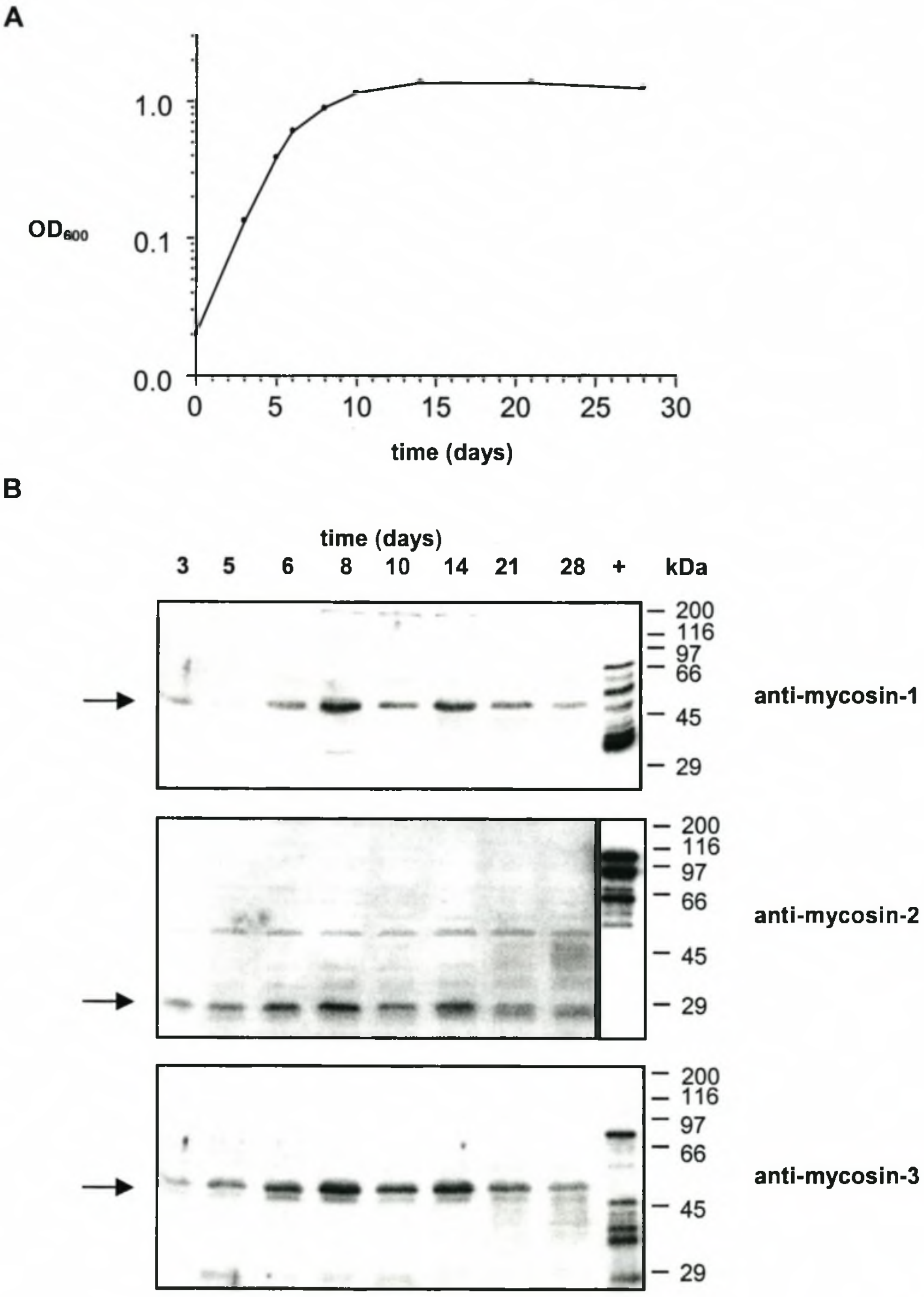
The expression of the mycosins in broth grown *M. tuberculosis* was analysed by Western blotting (Figure 2.4). Both mycosin-1 and mycosin-3 were detected and their apparent molecular weights were in agreement to those predicted (~50 kDa). Although full-length mycosin-2 could not be detected in *M. tuberculosis* lysates, two proteins were detected (~36 kDa and ~29 kDa) which corresponded to the cleavage products observed when mycosin-2 was expressed in *M. smegmatis*. The processing of mycosin-2 into these fragments may be autocatalytic as cleavage also took place in *M. smegmatis*. It is possible that cleavage occurs in the proline-rich region between the active site residues His133 and Ser435 (mycosin-2 numbering), to generate fragments of the observed size (~36 kDa and ~29 kDa).

Figure 2.5. Localisation of the heterologously expressed proteases to the membrane in *M. smegmatis*. Western blotted cellular fractions of transformed *M. smegmatis* were probed with the various anti-protease antisera, as indicated. The amount of sample loaded was adjusted to represent the proportions of the original cellular volumes. *M. smegmatis* was transformed with (A) p19k-P1, (B) p19k-P2RBS, (C) p19k-P3 and cellular fractions loaded were (1) cell free extract, (2) cytoplasm, (3) cell wall / membrane, (4) medium.



Expression of the mycosins in *M. tuberculosis* during growth in broth was examined by Western blotting (Figure 2.6). All three proteins appeared to be constitutively expressed, as they were detectable throughout the growth cycle of the organism. Although antibodies to mycosin-4 and mycosin-5 were not generated, dot blot analysis of RNA samples taken at the same times as the protein samples indicated that these genes were also constitutively expressed (data not shown). The constitutive expression of all of the mycosins in *M. tuberculosis* suggests that they play a role in normal cellular processes during the growth of the organism. As expression of mycosin-1 also occurs in intracellular bacteria and appears to be up-regulated during growth in macrophages (J.A. Dave *et al.*, submitted for publication), it is also possible that these proteases are involved in intracellular survival. As we had detected all five *mycP* genes in *M. bovis* BCG, we examined the expression of mycosin-1, -2 and -3 in this organism by Western blotting (Figure 2.4). Bands of similar molecular weights to mycosin-3 and mycosin-2 of *M. tuberculosis* were observed in lysates of *M. bovis* BCG, suggesting that these proteins were being expressed in this organism.

Figure 2.6. Expression of mycosin-1, -2 and -3 during the growth cycle of *M. tuberculosis*. (A) Growth curve of *M. tuberculosis*. Samples of *M. tuberculosis* were taken at the times indicated and measured against water at OD_{600nm}. Total protein was also isolated during sampling. (B) Western blots of *M. tuberculosis* protein samples probed with anti-protease antisera, as indicated. Protein samples, taken at the various time points indicated, were quantitatively equilibrated using coomassie blue staining. The positive control (+) in each case is 100 ng of the respective *E. coli*-derived protease-fusion protein.



Similarly to *M. tuberculosis*, mycosin-2 was cleaved into smaller molecular weight products and mycosin-2 and 3 were constitutively expressed during growth in broth (data not shown). Despite possessing a gene identical to *mycP1* (Genbank accession number U34848), no band of similar molecular weight to mycosin-1 could be detected and expression of mycosin-1 was not detectable throughout the growth cycle (data not shown). The lack of mycosin-1 expression in *M. bovis* BCG indicates that this protein is not essential for normal growth in broth, although it is possible that the other mycosins are compensating for the function of this protein.

As there are no mycosin homologs in other bacteria, it is difficult to speculate on the function of these proteases. It is unlikely that these enzymes are involved in nutrition; they are membrane bound and *M. tuberculosis* lysine auxotrophs are unable to replicate in macrophages (Hondalus *et al.*, 2000). They may function in the processing of secreted and/ or extracellular proteins, such as the 19 kDa lipoprotein antigen whose deglycosylation-dependant release from the cell surface is mediated by proteolytic cleavage (Herrmann *et al.*, 1996). These proteases may contribute to virulence, as extracellular protease activity has only been detected in the pathogenic mycobacteria (Kannan et al., 1987).

In conclusion, *Mycobacterium tuberculosis* possesses over 30 protease genes including a closely related family of five genes (the *mycP* genes), which encode transmembrane serine proteases that we have termed the mycosins. Multiple *mycP* genes are present in other virulent mycobacteria but only one gene was detected in the non-pathogenic *M. smegmatis*. All the mycosins were constitutively expressed in *M. tuberculosis* and at least one mycosin (mycosin-2) was modified by cleavage. One mycosin (mycosin-1) was not expressed in *M. bovis* BCG. The multiplicity and constitutive expression of the mycosins indicates that these enzymes play an important role in *M. tuberculosis*. We are currently investigating the substrate specificities of the mycosins and their role in the biology of *M. tuberculosis*.

ADDENDA TO CHAPTER TWO

ADDENDUM 2A

SUBCELLULAR LOCALIZATION

" ....plenty of work is waiting for microbial geneticists in the field of tuberculosis. I have never understood why the only infection in which resistance acquired in vivo is so frequent has not attracted their attention. Is it because everything takes place at such a slow pace in tuberculosis? .....with such a slow-motion picture of the process, what an opportunity for closer observation!"

Georges Canetti (1965)

2A.1. Introduction

Protein sequence analysis of the five mycosin proteases of *M. tuberculosis* indicated that each member of the family encodes a typical N-terminal signal sequence, as well as a strongly hydrophobic sequence in the C-terminal half of the protein. This suggests that the mycosins are likely to be secreted and anchored in the cytoplasmic membrane or in the lipid-rich regions of the mycobacterial cell envelope. In agreement with the evidence of a signal peptide, recent studies have established that *M. tuberculosis* contains a general *sec*-dependent protein export pathway in common with other eubacteria (Chubb,A.J., Woodman,Z.L. *et al.*, 1998). The signal peptide in the mycosin sequences is followed by a ± 42-residue segment preceding the first catalytic residue. This segment is likely to be a propeptide, but it is shorter and shows only weak homology to typical subtilisin propeptides (usually 69 - 84 residues in length). All subtilisins and most bacterial secreted proteases contain propeptides, which can be highly variable in length and sequence (Wandersman, 1989; Braun and Tommassen, 1998). Propeptides may assist with protein folding during export, and usually maintain the protease in an inactive state until the propeptide is cleaved (Braun and Tommassen, 1998).

Two of the three catalytic residues are located within hydrophobic regions. In general, some of the flanking sequences around catalytic residues in bacterial subtilisins are rich in hydrophobic amino acids, but the mycosins are even more hydrophobic than the normal subtilisins (data not shown). The significance of this is uncertain, but there are two possibilities; first, the mycosins act on hydrophobic protein substrates, or second, they reside predominantly within the lipid-rich mycobacterial cell wall, where the active sites are stabilized by hydrophobic interactions.

As a first step to elucidating the cellular location of the mycosins, we constructed a hypothetical model for the 3-dimensional structure of these enzymes, based on the consensus, empirically determined 3-dimensional structure of *Bacillus subtilis* subtilisin as a template, using the program SWISS-MODEL (http://www.expasy.ch/swissmod/SWISS-MODEL.html).

The output of this program reveals only the 3-dimensional core structure of the molecule, therefore the remaining part of the molecule (linker region and hydrophobic transmembrane anchor), was added manually (Figure 2A.1). The results of the analysis predicted the size of the mycosins to be between 4 x 4 nm and 5 x 5 nm. Although this is only an approximate prediction of the size of the proteins, it is sufficiently accurate to predict that the mycosins would not extend more than a few nanometres beyond the peptidoglycan layer of the cell wall, and would definitely not be able to reach through the wall (Figure 2A.2). Thus it seems as if these mycobacterial subtilisins must function inside the cell wall (although it is possible that it may be cleaved from the wall and secreted into the medium).

Figure 2A.1. Schematic representation of the predicted 3-dimensional structure and localization of the mycosins. The 3-dimensional structure of the core molecule was predicted using the program SWISS-MODEL, and the linker and transmembrane anchor was added manually.

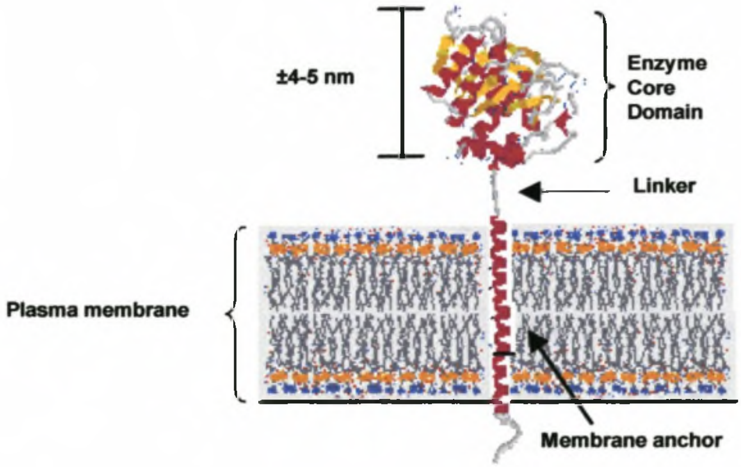
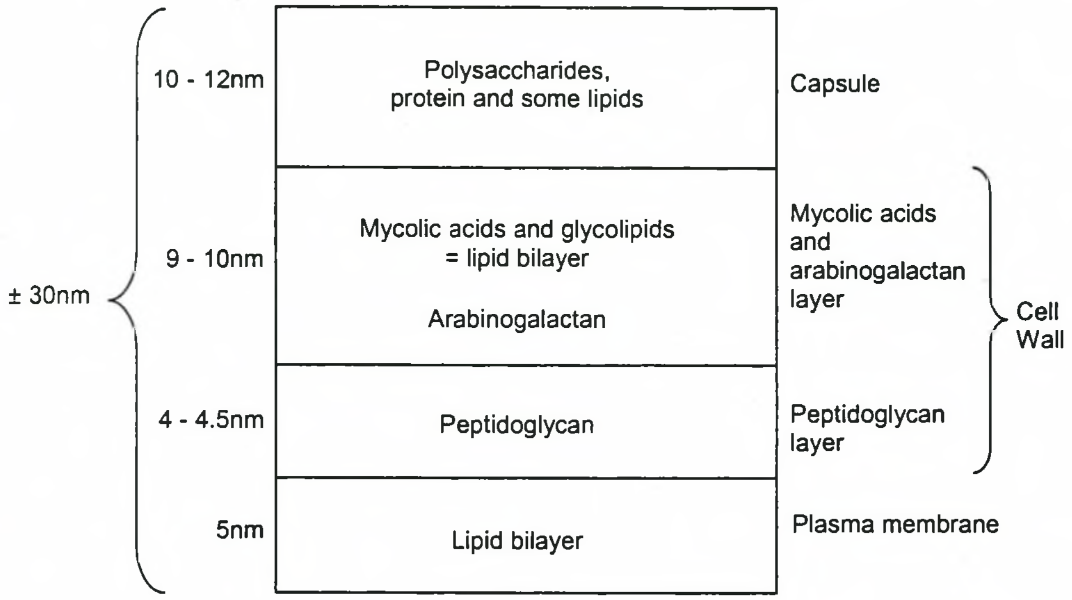


Figure 2A.2. Schematic representation of the mycobacterial cell envelope showing main constituents and predicted thickness of each layer.



The results obtained from the primary sequence analyses of these proteases led to experiments to determined the location of the heterologously-expressed proteases in cellular fractions of *M. smegmatis* by Western blotting (Brown *et al.*, 2000; Chapter 2, Figure 2.5). These results showed that mycosin-1, -2 and -3 are all partitioned in the cell wall/membrane fraction, which supported the hypothesis that the enzymes were predominantly located in the cell wall and anchored in the membrane, confirming the proposed functions of the signal peptide and transmembrane C-terminal regions. The heterologously-expressed mycosins did not appear to be released into the growth medium.

In a follow-up study on mycosin-1 (Dave *et al.*, submitted for publication), we examined the expression and localization of mycosin-1 in *M. tuberculosis* (clinical strain GSH-3052) by Western blotting. The localization of native mycosin-1 was found to be limited to the membrane and cell wall fractions with a complete absence in the cytoplasmic fraction. Interestingly, in *M. tuberculosis*, mycosin-1 protein was equally divided between cell membrane and cell wall fractions, whereas in

Figure 2A.3. Subcellular localization of mycosin-1 in *M. smegmatis*-P1 (expressing recombinant mycosin-1) and *M. tuberculosis*. Bacterial cell lysates were subfractionated into cell wall (w), cell membrane (m), and intracellular (i) fractions. Fractions were resolved by SDS-PAGE and analyzed by Western blotting using anti-mycosin-1 antiserum. Molecular weights (in kDa) are indicated on the left. (Reproduced with kind permission from Dr. J.A. Dave, Department of Medicine, Groote Schuur Hospital, Cape Town, South Africa).

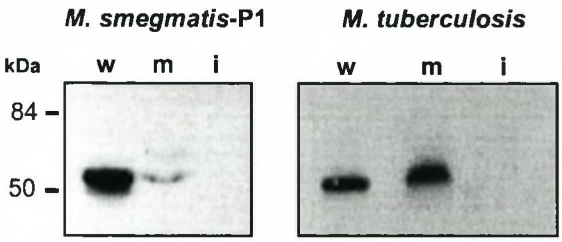
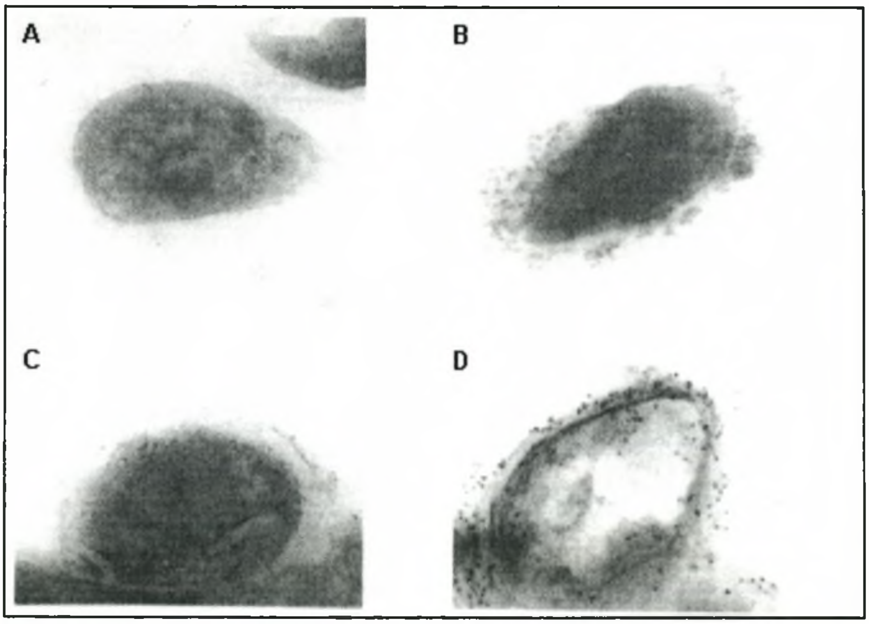
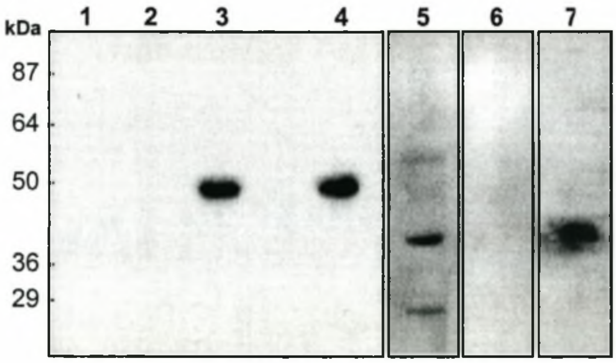


Figure 2A.4. Electron micrographs of *M. smegmatis*-P1 and *M. tuberculosis* following immunogold labeling with anti-mycosin-1 antiserum. Bacterial pellets were first incubated with anti-mycosin-1 antiserum, fixed, cryosectioned, and labeled with 5 nm gold particles. (A) *M. smegmatis* transformed with vector only (p19Kpro); (B) *M. smegmatis*-P1; (C) *M. tuberculosis* incubated with pre-immune serum as primary antibody; and (D) *M. tuberculosis* incubated with anti-mycosin-1 antiserum. (Reproduced with kind permission from Dr. J.A. Dave, Department of Medicine, Groote Schuur Hospital, Cape Town, South Africa).



M. smegmatis-P1 (expressing recombinant mycosin-1) most of the protein was located in the cell wall. The reason for this is unclear, but may reflect a greater tendency for the recombinant protein to be shed (cleaved) from the cell membrane in *M. smegmatis*. The Western blotting result was confirmed using immunogold transmission electron microscopy (Figure 2A.4), which revealed that mycosin-1 was clearly localized to the cell envelope in both *M. tuberculosis* (Figure 2A.4D) and *M. smegmatis*-P1 (expressing recombinant mycosin-1, Figure 2A.4B). In spite of this, mycosin-1 could also be identified in *M. tuberculosis* culture filtrates using Western blotting (Figure 2A.5), indicating that during growth the protein was shed from the cell envelope into the medium. Only full-length, 50 kDa mycosin-1 was observed in lysates of broth-grown *M. tuberculosis* and *M. smegmatis*-P1, whereas a 40 kDa species was detected in the 6-week *M. tuberculosis* culture filtrates. As mycosin-1 without the propeptide has a calculated mass of ~39 kDa, this could further suggest that the proenzyme is processed after shedding. A similar 40 kDa immunoreactive band was also observed in lysates of macrophages infected with the clinical strain *M. tuberculosis* GSH-3052, suggesting that mycosin-1 is also expressed and processed in a similar manner during intracellular residence (Figure 2A.5).

Figure 2A.5. Expression of mycosin-1 in mycobacteria in culture and during infection of macrophages. Samples were resolved by SDS-PAGE and analyzed by Western blotting using anti-mycosin-1 antiserum. Lanes: 1, *M. smegmatis*; 2, *M. smegmatis* transformed with p19Kpro vector; 3, *M. smegmatis* transformed with p19K-P1 (expressing recombinant mycosin-1); 4, *M. tuberculosis* clinical isolate GSH-3052 cell lysate; 5, *M. tuberculosis* clinical isolate GSH-3052 culture filtrate after growth in Kirchner's broth for 6 weeks; 6, lysate of uninfected P388D$_1$ macrophages; 7, lysate of *M. tuberculosis* clinical isolate GSH-3052-infected P388D$_1$ macrophages. Molecular weights (in kDa) are indicated on the left. (Reproduced with kind permission from Dr. J.A. Dave, Department of Medicine, Groote Schuur Hospital, Cape Town, South Africa).



2A.2. Localization of mycosin-3 in *M. tuberculosis* H37Rv

To confirm the subcellular localization results obtained with the heterologously expressed mycosin-3 protein (see Chapter 2, Figure 2.5) as well as to verify whether the same results could be obtained with mycosin-3 as was observed during the mycosin-1 analysis, we performed Western blotting experiments on subcellular fractions of *M. tuberculosis* H37Rv.

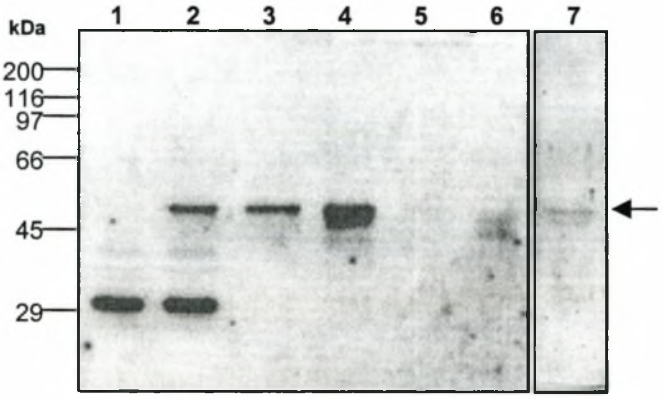
2A.2.1. Materials and Methods

Subcellular fractions (whole cell lysate, membrane, cytosol, cell wall and culture filtrate proteins) of *M. tuberculosis* H37Rv were obtained from Dr. J.T. Belisle (Department of Microbiology, College of Veterinary Medicine and Biomedical Science, Colorado State University). Samples were produced and provided through funds from the National Institutes of Health, National Institute of Allergy and Infectious Diseases, Contract No1-AI-75320, entitled "Tuberculosis Research Materials and Vaccine Testing." Five micrograms of each sample was resolved by reducing SDS-PAGE, electroblotted onto a nitrocellulose membrane and probed with 1/5000 of positively- and negatively-selected polyclonal anti-mycosin-3 antibodies (see Chapter 2 for details of antibody preparation).

2A.2.1. Results

The results of the Western blot analysis of subcellular fractions of *M. tuberculosis* H37Rv are presented in Figure 2A.6. These results indicate the presence of full-length (~50 kDa) mycosin-3 in the cell wall fraction, with very little to no protein detected in the membrane and cytosolic fractions, and a small amount in the culture filtrate fraction. The size of the protein in the culture filtrate fraction was the same as in the cell wall, which is in contrast with what was observed with mycosin-1 (Figure 2A.5). It thus seems as if some shedding of mycosin-3 does occur, as is observed in various microorganisms (Doyle *et al.*, 1988), but no processing of the protein could be detected. Although *M. smegmatis* possesses a *mycP3* orthologue, mycosin-3 was not detected in wild type *M. smegmatis* by Western blotting during previous experiments (Chapter 2, Fig. 2.4). The results from this analysis, though, did reveal an immunoreactive band at around 30 kDa in both wild-type *M. smegmatis* and *M. smegmatis* expressing mycosin-3. Although this may be non-specific antibody binding, it is possible that the anti-mycosin-3 antibodies are detecting a processed form of the *M. smegmatis* mycosin-3 orthologue, similar to what was observed with mycosin-1.

Figure 2A.6. Localization of mycosin-3 in *M. tuberculosis* H37Rv subcellular fractions. Samples were resolved by SDS-PAGE and analyzed by Western blotting using positively- and negatively-selected polyclonal anti-mycosin-3 antibodies (1/5000). Lanes: 1, *M. smegmatis* transformed with p19Kpro vector membrane fraction; 2, *M. smegmatis* transformed with p19K-P3 (expressing recombinant mycosin-3) membrane fraction; 3, *M. tuberculosis* H37Rv whole cell lysate; 4, *M. tuberculosis* H37Rv cell wall fraction; 5, *M. tuberculosis* H37Rv membrane fraction; 6, *M. tuberculosis* H37Rv cytoplasmic fraction; 7, *M. tuberculosis* H37Rv culture filtrate. Molecular weights (in kDa) are indicated on the left.



2A.2.2. Discussion

We previously determined the location of the heterologously expressed proteases in cellular fractions of *M. smegmatis* by Western blotting (Chapter 2, Figure 2.5) and showed that mycosin-1, -2 and -3 were all localised to the cell wall/membrane fraction, supporting the proposed functions of the signal peptide and transmembrane C-terminal regions. The mycosins did not appear to be released into the growth medium during heterologous expression. More complete subfractionation experiments (of cell lysates into wall, membrane, and intracellular fractions) were performed on heterologously expressed mycosin-1, as well as mycosin-1 expressed natively by *M. tuberculosis* (Dave *et al*., submitted for publication). The results obtained from these experiments indicated that mycosin-1 was present in the cell wall and membrane fractions as observed previously (Figure 2A.3 and 2A.4), but could also be detected in the culture filtrate in an apparently processed form (Figure 2A.5). The subcellular localization analysis done on mycosin-3 showed that this protein predominantly resides in the cell wall fraction of *M. tuberculosis* H37Rv, but some protein could also be detected in the culture filtrate, although apparently in an unprocessed form. It has been shown previously that mycosin-5 (Rv1796) was only able to elicit delayed-type hypersensitivity reactions in guinea pigs immunized with live mycobacteria (Romain *et al.*, 1993), providing supporting evidence that these proteins may be able to be shed from the cell wall surface during active growth of the mycobacteria as part of normal cell wall turnover (Doyle *et al.*, 1988).

In conclusion, we have used a number of approaches to determine the subcellular localization of selected members of the mycosins. The results obtained from these analyses showed that the mycosins are secreted, membrane bound, cell wall-associated subtilisin-like serine proteases that are shed by an unknown mechanism (actively or passively) from the cell wall during growth of *M. tuberculosis* under *in vitro* and *in vivo* conditions.

ADDENDUM 2B

T-CELL RESPONSES

"Gladly the humble shepherdess, responded to that gentle call; and following Mary, swift to bless, she came to Carmel's lofty wall"

Poems - St. Theresa of Lisieux

2B.1. Introduction

There has been increased interest in secreted antigens of the mycobacteria as candidates for a subunit-based vaccine (Andersen, 1997). The reason for this is the fact that it was observed that only live vaccines, and not killed preparations, can provide long-lived specific immunity towards these pathogens. In addition to this, it has been shown that in animal models of tuberculosis, culture filtrate proteins of *M. tuberculosis* cultures offers some degree of protection (Hubbard *et al.*, 1992, Pal and Horwitz, 1992, Andersen, 1994a, Andersen, 1994b, Horwitz *et al.*, 1995, Roberts *et al.*, 1995). Thus, there is a number of studies presently being done to determine novel antigens of *M. tuberculosis* that are being secreted or shed extracellulary, as these antigens in the culture filtrate are an important potential source of candidates for a subunit vaccine against tuberculosis. From the analysis of the subcellular localization of the mycosin proteases of *M. tuberculosis*, we have shown that the mycosins are located in the cell wall of the bacterium, and may be shed from the wall into the extracellular milieu during active growth of the organism (see Addendum 2A). Furthermore, we have detected a proteolytic activity in the culture filtrates of *M. tuberculosis*, which is inhibited by mixed serine/cysteine protease inhibitors and activated by Ca^{2+}, features typical of the subtilisins (see Addendum 2C). In accordance with this, Romain and coworkers (1993) have shown that mycosin-5 (Rv1796) was able to elicit delayed-type hypersensitivity (DTH) reactions only in guinea pigs immunized with live mycobacteria, indicating the release of protein only during active growth of the organism. As DTH is known to be a strictly T-cell dependant immune reaction, it is possible that the mycosin-5 protein may also be recognized by the T-cell population mediating the cell mediated immune response (CMI) reaction. This reaction is key to the effective activation of macrophages to control *M. tuberculosis* infection (Andersen, 1997, Flynn and Chan, 2001).

There is a constant search for antigens able to stimulate the CMI cellular population for use as possible vaccine candidates. We decided to perform preliminary investigations into the T-cell response profile of the recombinant mycosin fusion proteins (mycosin-1-GST, mycosin-2-MBP and mycosin-3-GST) using whole blood assays (Kirchner *et al.*, 1982, Rothel *et al.*, 1990, Rothel *et al.*, 1992), during which we determined T-cell proliferation as well as gamma interferon (IFN-γ) production. IFN-γ production was assayed because it is widely recognized that the immune response to *M.*

tuberculosis is highly dependent upon the production of this cytokine by antigen-specific T-cells (Flynn and Chan, 2001).

2B.2. Materials and Methods

2B.2.1. Methodology

The methodology utilized to investigate the T-cell responses towards the recombinant mycosins *in vitro* was the whole blood culture technique (Kirchner *et al.*, 1982, Rothel *et al.*, 1990, Rothel *et al.*, 1992), during which T-cell proliferation and IFN-γ production were assayed. This technique requires less incubation time, is technically simpler, inexpensive, more sensitive and provides a milieu closer to *in vivo* conditions than when using peripheral blood mononuclear cells (PBMC's)(Doldi *et al.*, 1985, Rothel *et al.*, 1992).

2B.2.2. Participants and skin testing

Two preliminary experiments were done. In the primary experiment, three volunteers were recruited of which one was Mantoux positive, one was a recovered tuberculosis patient and the third was a patient presenting with disease. In the second larger experiment, eight healthy volunteers with no previous history of tuberculosis were recruited from a scientific setting. The ages of the participants ranged from 22 to 47, with the median age being 27. Prior to enrollment, all subjects were skin tested with 0.1 ml of 5 TU (tuberculin units) PPD placed intradermally according to the standard technique (Mantoux skin testing technique). All PPD skin tests were placed and read by a certified nurse who performs these duties regularly. All readings were done with the palpation and ballpoint methods along two axes of the forearm (Sokal, 1975). A positive reading was defined as an induration greater than 10 mm in diameter. Five of the subjects were Mantoux positive (13, 18, 18, 19, and 22 mm induration respectively), while the rest (three subjects) were negative (0 mm induration each). All individuals had negative chest X-rays.

2B.2.3. Whole blood culture

Venous blood was collected from the subjects in sodium heparinized tubes. Whole blood culture was done by diluting blood 1:10 with sterile RPMI 1640 medium (supplemented with 5x10^{-5} M 2-mercaptoethanol, 100 IU of penicillin per ml, 50 μg of streptomycin per ml, 2 g/l $NaHCO_3$, 1mM of L-glutamine and 10% Fetal Calf Serum) and aliquoting 180 μl of diluted blood in wells of a 96-well tissue culture plate. Above procedure was done within 30 minutes after blood collection. The whole blood was stimulated with 20 µl of either sterile RPMI medium (as unstimulated control), the mitogen phytohemagglutinin (PHA, Sigma C/N: L-8754, used as a positive control for T-cell activation at a final concentration of 10 µg/ml), PPD (purified protein derivative - a gift from Stan Ress, Department of Medicine, Clinical Immunology, Groote Schuur Hospital, used as a positive control for mycobacterial-specific T-cell activation at a final concentration of 3.3 µg/ml), purified glutathione-S-transferase protein (GST, fusion partner control, final concentration of 1 µg/ml), purified maltose binding protein (MBP, fusion partner control, final concentration of 1 µg/ml), purified mycosin-1-GST fusion protein (final concentration of 1 µg/ml), purified mycosin-2-MBP fusion protein (final concentration of 1 µg/ml) and purified mycosin-3-GST fusion protein (final concentration of 1 µg/ml). Fusion proteins and fusion partners were prepared as described in Brown *et al.*, 2000 (Chapter 2) and were 0.22 µM filter sterilized. Six replicates of all tests were carried out. The tissue culture plates were incubated for 7 days at 37°C (in an atmosphere of 5% CO_2 in humidified air).

2B.2.4. T-cell proliferation assays

At 20 h prior to harvesting, 20 µl of Methyl-[^{3}H]thymidine (Amersham C/N: TRK120) was added to each well to a final concentration of 2.5 µCi. PHA samples were harvested at day 4 and all other samples at day 7. The cells were harvested onto fiberglass paper, added to a vial containing 4.5 ml of scintillation fluid (Insta-Gel Plus - Biotecknik C/N: 6013399) and the incorporated radioactivity was measured in a liquid scintillation counter. The proliferative responses were expressed in counts per minute.

2B.2.5. Interferon-γ ELISA assays

Plasma in culture supernatants was harvested from parallel cultures at day 5 and stored at -20°C for later quantification of IFN-γ using a commercially available sandwich enzyme-linked immunosorbent assay (ELISA) kit (Human IFN-γ DuoSeT, Genzyme Diagnostics C/N: 80393200). Interferon-γ ELISA's were done according to the manufacturer's specifications. Briefly, 96-well plates were coated with 100 µl of anti-human IFN-γ capture antibody in coating buffer (0.1 M Carbonate, pH 9.4 - 9.8) per well and incubated overnight at 2-8°C. Plates were washed for 5 cycles with wash buffer (PBS containing 0.05% Tween 20) and subsequently blocked with 250 µl of blocking buffer

(PBS containing 4% BSA) per well at 37°C for 2 hours. Blocking buffer was decanted, plate blotted dry and 100 μl of standards (dilutions of recombinant human IFN-γ covering an assay range from 15.6 to 1000 pg/ml) and samples (1/20 dilutions of PHA and PPD and 1/5 dilutions for other samples) added to each well. Plates were incubated for 2 hours at 37°C. Thereafter, the plates were washed for 5 cycles with wash buffer and 100 μl of diluted anti-human IFN-γ HRPO secondary antibody was added to each well. The plates were once more incubated at 37°C for 30 minutes. . After incubation, the plates were washed for 5 cycles with wash buffer, 100 μl of TMB substrate reagent was added to each well, and the plates were incubated for 30 minutes at room temperature, at which time 100 μl of stop solution (2N H_2SO_4) was added to each well to halt the reaction. Absorbance was read at 450nm within 60 minutes.

2B.2.6. Statistical methods

All data are expressed as median values of results from six wells per stimulant. Standard deviations were calculated for all T-cell proliferation values. For IFN-γ results, mean absorbance of standards and samples were calculated and background (mean OD of zero pg/ml standard) were subtracted to calculate corrected absorbance. A standard curve was constructed by plotting the mean OD_{450} values of the seven standards versus their corresponding concentrations in pg/ml.

2B.3. Results

In the primary study, the recombinant mycosin proteases were evaluated *in vitro* for their ability to induce T-cell proliferation and IFN-γ production with whole blood cultures obtained from three subjects of different disease status. The cellular proliferation results of the primary experiment are presented in Figure 2B.1. Although proliferations were low in all three subjects, the results indicate a higher level of responses in the Mantoux positive subject, in comparison to both the tuberculosis patient presenting with disease as well as the cured patient. When comparing purified antigens to complex mixtures like PPD, the IFN-γ production should always be monitored as even a very low T-cell proliferation is associated with a pronounced interferon-gamma production (Peter Andersen, Statens Serum Institut, Copenhagen, Denmark, personal communication). The results of the interferon-gamma production assays are presented in Figure 2B.2. These results clearly indicate that the Mantoux-positive subject has much higher levels of IFN-γ production when stimulated with the individual mycobacterial proteases than either of the tuberculosis patient or cured subject. This is despite the fact that in all three cases extremely high responses (both proliferative and IFN-γ secreting) were obtained against mycobacterial purified protein derivative (PPD, Figure 2B.3). The responses to the fusion partner GST (glutathione-S-transferase protein, fused to mycosin-1 and -3) are nominal, but the MBP fusion partner (maltose binding protein, fused to mycosin-2) gives substantial production of IFN-γ on its own, although it is still less than what is observed with the mycosin-2-MBP fusion protein. The results from this primary experiment underlined the necessity for a follow-up experiment using a larger sample base, to examine the differences between responses from Mantoux-positive and -negative subjects.

Figure 2B.1. Cellular reactivity to mycobacterial mycosin proteases. Proliferative responses of whole blood cultures from a tuberculosis patient presenting with disease, a recovered tuberculosis patient as well as a Mantoux-positive control subject. Values shown are median values of results from six wells ± standard deviations. Values were compensated against unstimulated controls.

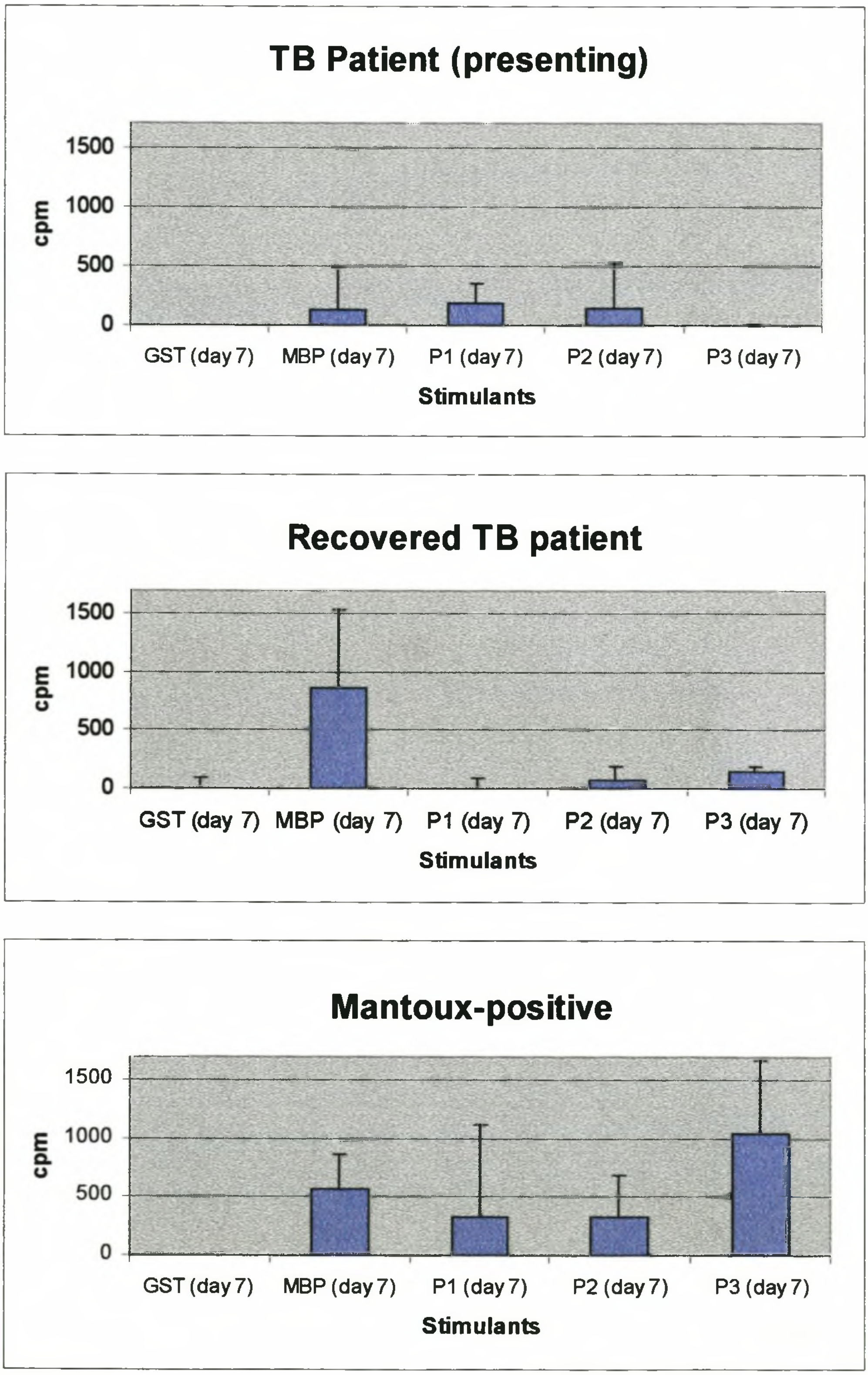


Figure 2B.2. Interferon-gamma production in response to mycobacterial mycosin proteases. Production of IFN-γ in whole blood cultures from a tuberculosis patient presenting with disease, a recovered tuberculosis patient as well as a Mantoux-positive control subject. Values shown are picograms of IFN-γ produced per milliliter and are mean values of results from six wells.

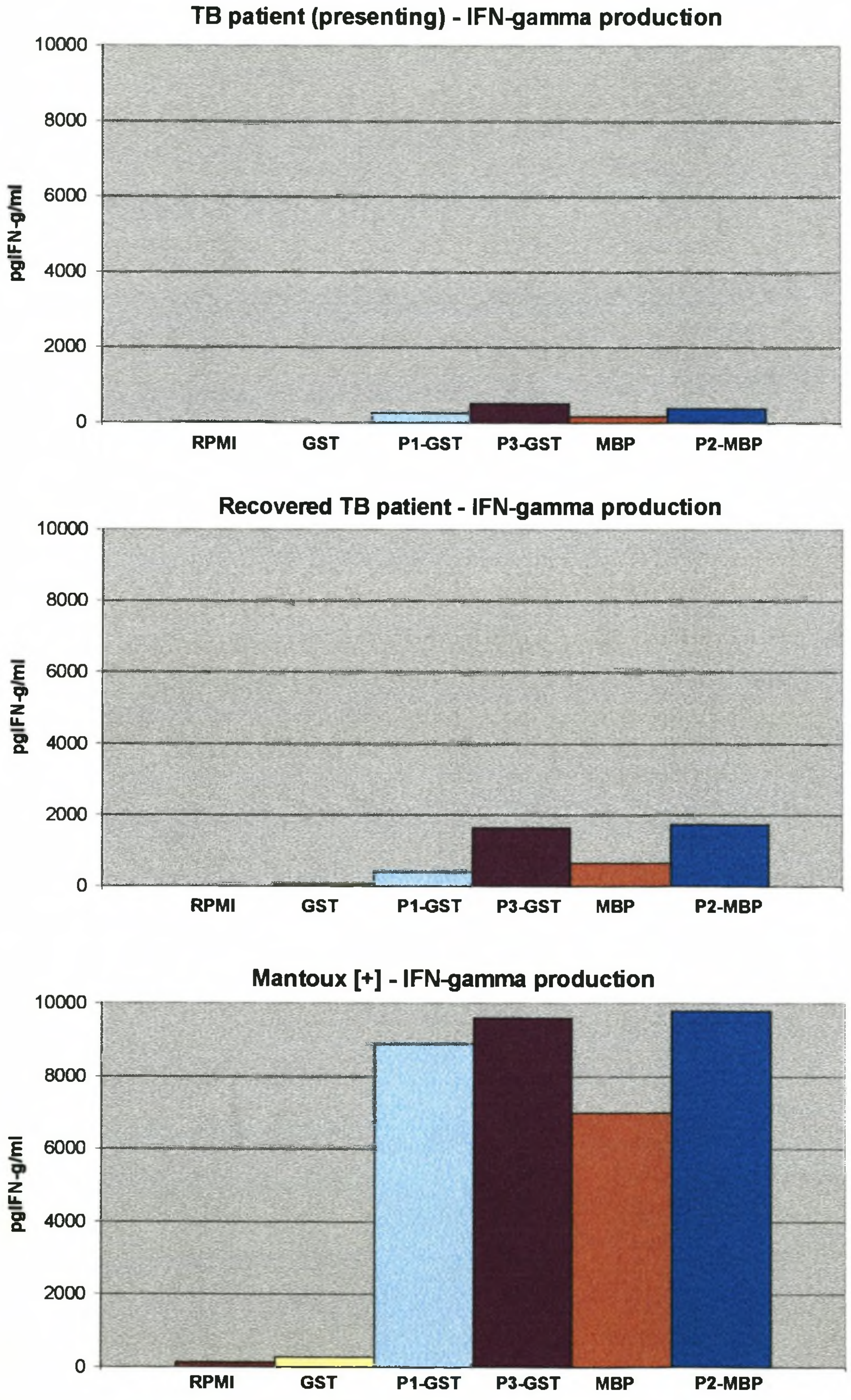
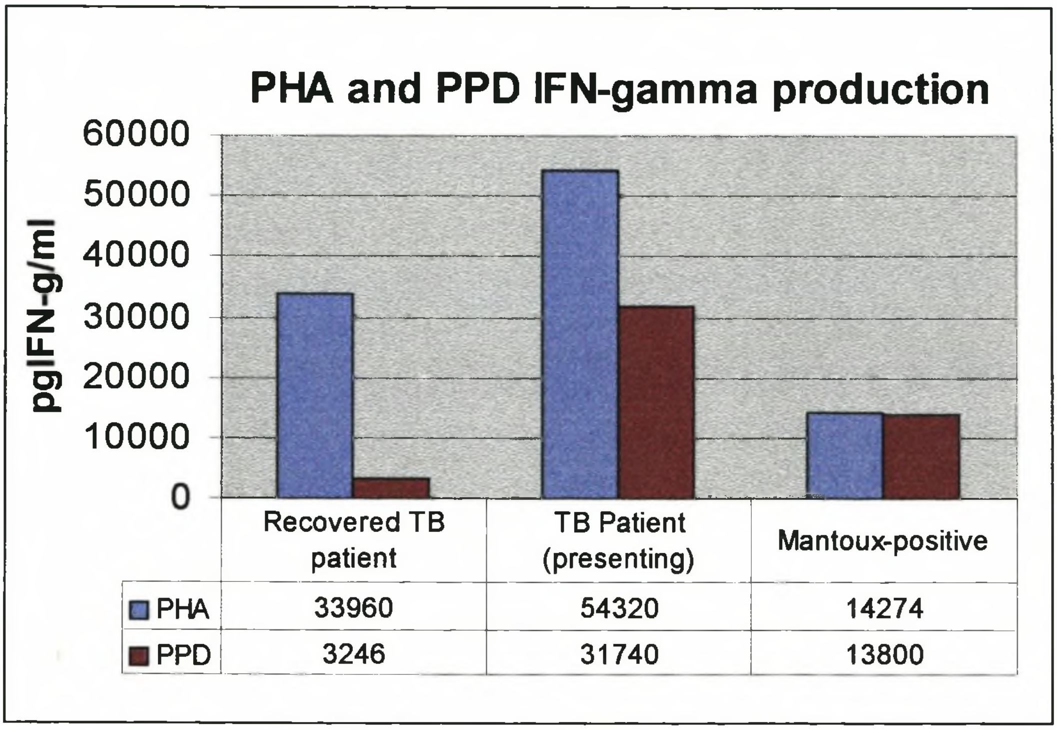


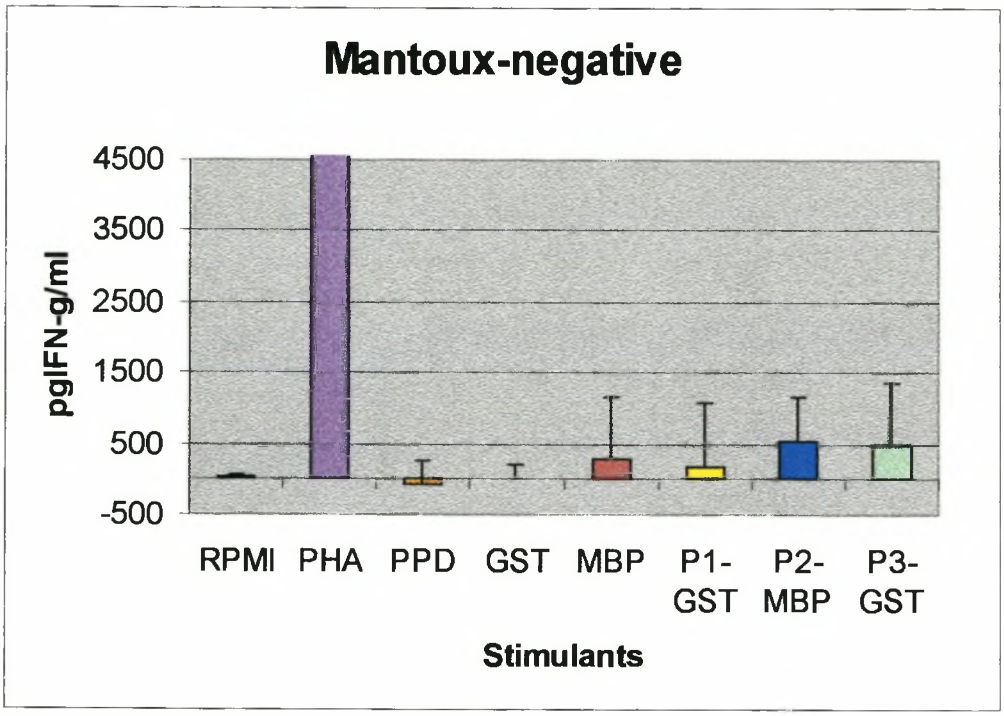
Figure 2B.3. Interferon gamma production in response to controls for cellular activation, the mitogen PHA and mycobacterial purified protein derivative (PPD). Control production of IFN-γ in whole blood cultures from a tuberculosis patient presenting with disease, a recovered tuberculosis patient as well as a Mantoux-positive control subject. Values shown are picograms of IFN-γ produced per milliliter and are mean values of results from six wells.

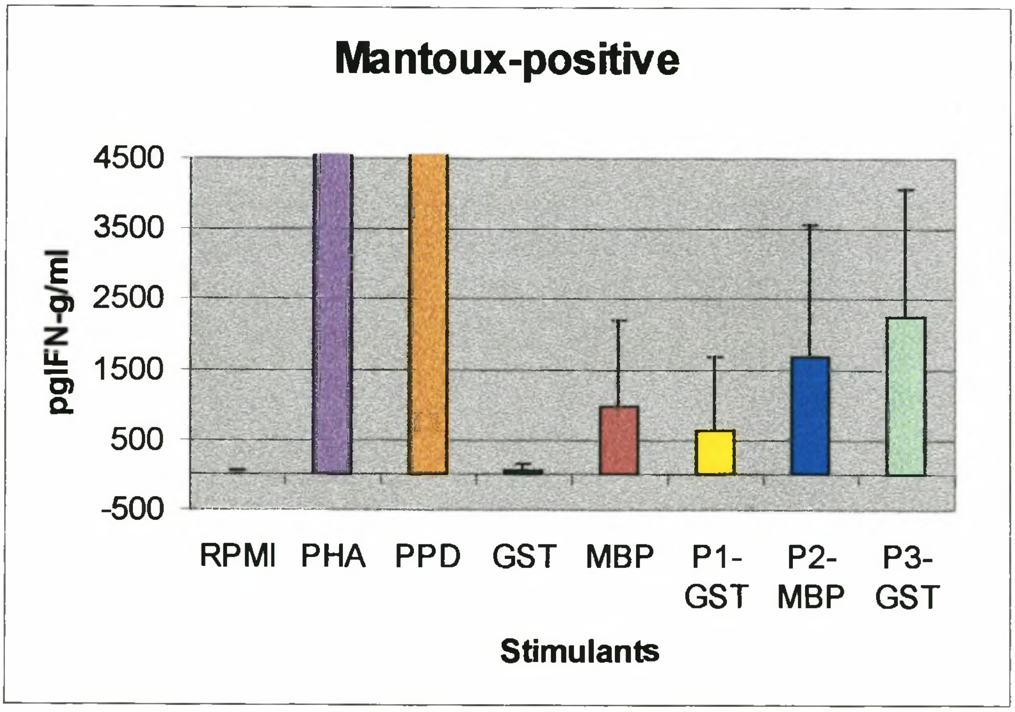


In the second experiment, the recombinant mycosin proteases were evaluated *in vitro* for their ability to induce T-cell proliferation and IFN-γ production with the whole blood cultures obtained from healthy Mantoux-positive (PPD$^+$, induration of more than 10 mm) and healthy Mantoux-negative (PPD$^-$, induration of 0 mm) subjects. The results of this analysis showed no differences in the T-cell proliferations between the two groups (data not shown). As described previously, the proliferation of T-cells towards singular antigens are mostly nominal and the IFN-γ responses were thus examined. These results are presented in Figure 2B.4, and shows that, although the standard deviations are quite large, it seems clear that the responses towards the recombinant mycosin proteases (especially mycosin-2 and -3) are higher in the PPD$^+$ individuals, when compared to the PPD$^-$ subjects. No production of IFN-γ in response to stimulation with PPD (mean value 35 pgIFN-γ/ml, less than the unstimulated control with a mean value of 48 pgIFN-γ/ml) was observed in any of the Mantoux-negative subjects, as would be expected (Katial *et al.*, 2001). All Mantoux-positive subjects had high levels of IFN-γ production in response to PPD stimulation (mean value of 14894 pgIFN-γ/ml). Again, as in the primary experiment, no response was obtained towards the fusion partner GST, while definite responses were obtained against the MBP fusion partner. It thus seems as if the maltose binding protein fusion partner has an inherent antigenicity recognized by all the individuals in this study. The fact that the MBP fusion partner on its own seem to evoke the production of IFN-γ did not influence the observed results of mycosin-2-MBP, as responses towards this fusion protein were consistently higher than that observed towards the fusion partner. As in the primary experiment, the viability of all donor whole blood culture cells was confirmed by proliferation and the secretion of IFN-γ in response to the mitogen phytohemagglutinin (PHA).

The IFN-γ response profile suggested a difference in the responses towards stimulation with the mycobacterial proteases between Mantoux-positive and -negative subjects, although, because of the low sample base, the standard deviations were large, especially in the Mantoux-positive group.

Figure 2B.4. Interferon-gamma production in response to mycobacterial mycosin proteases. Production of IFN-γ in whole blood cultures from Mantoux-negative (*n* = 3) and Mantoux-positive (*n* = 5) subjects. Values shown are picograms of IFN-γ produced per milliliter and are mean values of results from six wells.





2B.4. Discussion

The protective response against *M. tuberculosis* infection relies heavily on the cell mediated immune response (CMI) (Andersen, 1997). The reason for this is the fact that the organisms are usually found situated within the phagosome inside macrophages during infection, requiring T-cell effector mechanisms rather than antibodies to control or eliminate the bacteria (Flynn and Chan, 2001). The primary effector function of T-cells is the production of IFN-γ in response to a wide variety of mycobacterial antigens, which are, together with some other cytokines, sufficient to activate the macrophages for the control or elimination of the intracellular organisms (Flynn and Chan, 2001). IFN-γ is thus a key cytokine in the control of *M. tuberculosis* infection. The whole blood culture technique was developed in the early part of the previous decade specifically to determine the production of different types of human interferons. Interferon production was studied in mixtures of whole blood of healthy adults with tissue culture medium (Kirchner *et al.*, 1982). Since there is no need to supplement the system with additional foreign serum and since the cell populations in the whole blood assay are present in their natural distribution, this test system reflects the *in vivo* situation better than an assay using isolated peripheral blood mononuclear cells (PBMC's)(Doldi *et al.*, 1985). It took a few years before this system was applied to the study of mycobacterial infections, and it was first used in a sandwich enzyme immunoassay for bovine IFN-γ to detect bovine tuberculosis in Australian cattle (Rothel *et al.*, 1990). This assay detected IFN-γ released in response to specific antigens in a whole blood culture system and was found to be a simple, rapid and sensitive *in vitro* assay for specific cell mediated immune responsiveness to *M. bovis* infection in cattle (Rothel *et al.*, 1992). The whole blood assay technique has been successfully applied in our laboratory for a number of years to study T-cell proliferation and cytokine release in whole blood cultures from tuberculosis patients and Mantoux positive and negative controls in response to mycobacterial antigens.

In this study we aimed to obtain some insight into the T-cell antigenicity of the mycobacterial mycosin proteases. It has been shown previously that these proteases are potentially shed into the extracellular milieu of the bacterium during active growth (Addendum 2A). As the extracellular milieu of the organism is also the intracellular milieu of the phagosome of the macrophage, these proteins may be subjected to MHC class II presentation to CD4$^+$T-cells. In support of this, a DTH response (a

T-cell-dependent reaction) was previously observed to one of the members of this family (mycosin-5) only during immunization with live bacteria (Romain *et al.*, 1993). It is thus of more than academic importance to evaluate the T-cell antigenicity of these proteins, as they may have a potential to be utilized as possible vaccine candidates.

As T-cells recognize peptide antigens in association with MHC molecules, the different types of MHC molecules found in different individuals have a major influence on the antigenic epitopes presented to the T-cells in these individuals (Andersen, 1997). Each type of MHC protein is encoded by several genes, and there are several alleles that represent different MHC proteins for each of these genes. Thus, a tremendous variety of MHC proteins exists, with the effect that it is unlikely for two individuals to possess the same group of MHC molecules (Zubay, 1993). Looking for T-cell responses to single antigens is difficult, owing to low responder frequency. In other words, different individuals may respond to the limited set of T-cell epitopes of a single antigen differently because of the large MHC variety. As PPD contains hundreds of antigens, the responder frequencies to this complex mix of antigens would be relatively high with the subsequently higher observed responses. But, although T-cell proliferation to individual mycobacterial antigens such as ESAT-6 is often marginal or even absent, dominant antigens can still induce relatively prominent IFN-γ responses, that can be measured to determine levels of antigenicity (Peter Andersen, Statens Serum Institut, Copenhagen, Denmark, personal communication).

In the primary experiment, IFN-γ production was assayed in whole blood cultures from a Mantoux-positive subject, a TB patient presenting with disease as well as a cured patient. The results have shown that high levels of IFN-γ were produced after stimulation with the three mycosin protease fusion proteins only in the Mantoux-positive subject, and not in the cured tuberculosis patient or the patient presenting with disease. Interestingly, all three subjects had high responses towards PPD, indicating an inability of the subjects infected by and cured of tuberculosis to specifically respond to the proteases. A number of studies have shown that patients with active tuberculosis are characterized by deficient IFN-γ production and a depressed cell mediated immunity (Onwubalili *et al.*, 1985, Flynn and Chan, 2001). This could be due to several factors, including antigen specific T-cell depletion due to sequestration at the site of infection (Ravn *et al.*, 1999). The observation that the cured patient displayed a similar absence of specific IFN-γ release as the tuberculosis patient, as well as the observation that these two subjects each did have a prominent response to both the mitogen PHA and the mycobacterial PPD, suggests that depression of the CMI response is not a valid explanation for this observation. It is much more likely to be due to the phenomenon called the "hole in the T-cell repertoire" (Andersen, 1997). This phenomenon is linked to the MHC restriction of responses to single epitopes, leading to the observation that tuberculosis patients, in contrast to healthy subjects, do not recognize some epitopes on mycobacterial antigens. It has been suggested that these "holes in the T-cell repertoire" be involved in the immunopathogenesis of tuberculosis, and that the antigens, which are being responded to by the healthy subjects but not by the T-cells from the tuberculosis patients, could be implicated in protective immunity.

The primary experiment was extended to an investigation involving a larger sample base of eight subjects, three Mantoux-negative and five Mantoux-positive. The results from this experiment once again suggested a trend of an increased level of IFN-γ responses towards the proteases in the Mantoux-positive group, when compared to the Mantoux-negative subjects. The standard deviations were large, though, especially in the Mantoux-positive group, because of the low sample base. Thus, clearer results will only be obtained by using a larger subject group in future studies. The levels of IFN-γ production observed in response to stimulation with the mycosin-1 protease in the Mantoux positive group, was much lower than the levels produced in response to the mycosin-2 and -3 proteases. A possible explanation for this may be the fact that mycosin-1 is not expressed in the *M. bovis* BCG vaccine strain (Chapter 2). As the Mantoux positivity of these individuals may be due to BCG vaccination, it may explain the general absence of a response to this protein in this group. The question of whether these mycosins are actively released into the culture filtrate or whether the observed responses were due to shedding and normal cell death could not be answered by the results of this study.

In conclusion, we have shown that the mycosin proteases may be able to elicit low levels of T-cell dependent cellular proliferation, with concomitant production of relatively high levels of IFN-γ. There was an interesting difference observed between the three subjects of the primary study, suggesting that the mycosins may be specifically recognized only in healthy individuals, and pointing to the possibility that these proteins may be involved in protective immunity. These results are preliminary, and will have to be followed by studies involving a larger sample base. The results of the secondary study indicated that the mycobacterial mycosins are recognized by Mantoux positive individuals, thereby being able to stimulate the CMI response. These mycosins are thus potentially interesting for use as components for future subunit vaccines against tuberculosis. As the sample base for both the primary and secondary experiments was not optimal, the subject number used in these studies will have to be increased in future investigations to determine the full extent of antigenicity of the members of the mycosin family of proteases.

ADDENDUM 2C

PROTEASE ACTIVITY

" ....a number of useful contributions to our present knowledge must be discussed too briefly or not at all. Particularly ironic is the fact that the action of proteolytic enzymes on proteins is omitted from review, since here our ignorance is deepest."

Proteolytic enzymes – B. S. Hartley (1960)

2C.1 Introduction

Proteases (also termed proteinases, peptidases or peptide hydrolases) are proteolytic enzymes that catalyze the hydrolysis of proteins or peptides (Wandersman, 1989). They play an important role in normal physiological processes where they function as biological regulators in zymogen activation, release of hormones, nutrient acquisition, cell growth and protein turnover (Neurath, 1989, Wolf, 1992). Many saprophytic bacteria utilise proteases with broad substrate specificities to obtain nutrients from the surrounding environment (Travis *et al.*, 1995). Pathogenic organisms may also use proteases to obtain nutrients from their hosts and, in addition, frequently use proteases to perform more specific functions related to virulence (Lantz, 1997).

There are six different molecular mechanisms by which proteases may contribute to pathogenesis: (1) the enhancement of vascular permeability and edema formation, (2) degradation of defense orientated proteins, (3) inactivation of the complement system, (4) degradation of regulatory plasma protease inhibitors (serpins), (5) destruction of intracellular integrity and cell killing, and (6) elevated lethality by increase of viral yield during co-infection (Maeda and Molla, 1989). Proteases also contribute to tissue invasion and destruction, evasion of host defenses and the subversion of the host immune system (Kilian *et al.*, 1988, Travis *et al.*, 1995).

Pain and edema at the site of infection may be caused by the dysregulation of the kallikrein-kinin pathway by foreign bacterial proteases (Molla *et al.*, 1989). At least 16 microbial proteases can activate the host kallikrein-kinin cascade, leading to the generation of bradykinin and a local inflammatory reaction (Travis *et al.*, 1995). The local increase in vascular permeability and the recruitment of phagocytic cells, which in turn release host-derived proteases, may benefit the pathogen by increasing the supply of nutrients. Proteases from pathogens can also activate other zymogen cascades, including the coagulation, fibrinolysis and complement pathways (Wingrove *et al.*, 1992, Travis *et al.*, 1995). Many bacterial proteases cause tissue damage by degrading collagen, elastin and fibronectin: the secreted proteases of *Pseudomonas aeruginosa*, *Serratia marcesens*, *Porphyromonas gingivalis* and *Clostridium perfringens* contribute to tissue damage in cystic fibrosis, keratitis, periodontitis and gas gangrene respectively (Travis *et al.*, 1995). Proteases secreted by *P.*

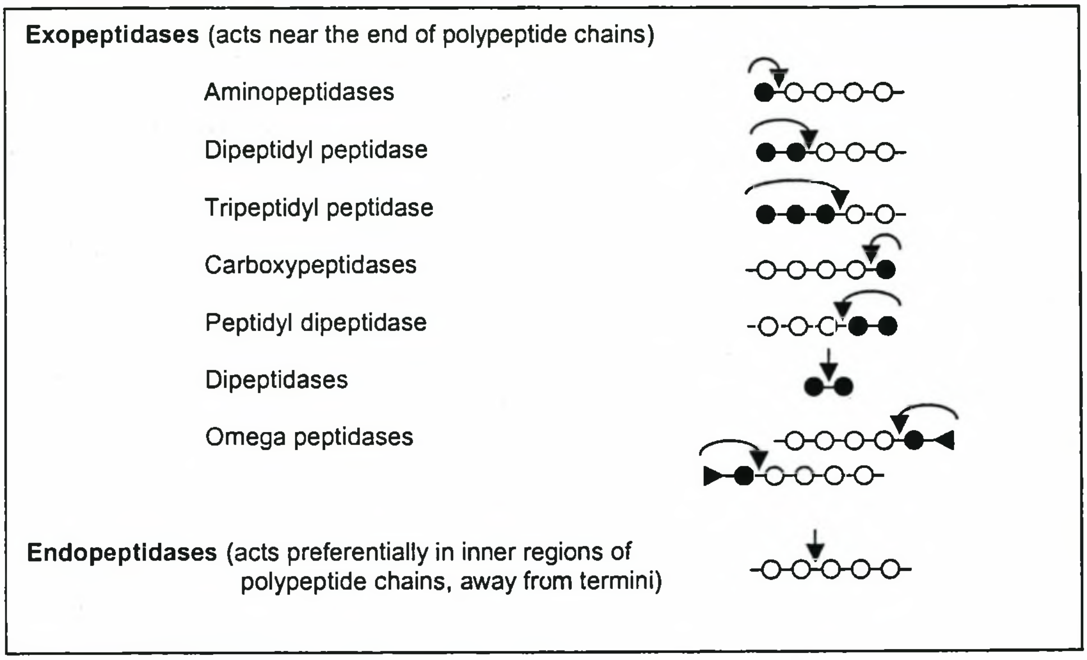
gingivalis not only activate the kallikrein-kinin system, inducing a local inflammatory reaction, but also degrade complement and immunoglobulins (Wingrove *et al.*, 1992, Cutler *et al.*, 1993, Travis *et al.*, 1995). Other examples include the HtrA serine protease of *Salmonella typhimurium*, which has been associated with virulence during intracellular infection (Johnson *et al.*, 1991) as well as a secreted serine protease identified in the culture filtrates of virulent strains of *Dichelobacter nodosus* associated with virulent foot-rot disease (Kortt *et al.*, 1994). Thus, many proteases are involved in various human diseases and therefore these enzymes are targets for the development of inhibitors as new therapeutic agents (Powers *et al.*, 1993).

Three major criteria are used to classify proteases: (1) the reaction catalyzed, (2) the chemical nature of the catalytic site, and (3) the structural and evolutionary relationship of the protein (Barrett, 1994). The first classification system depends on the type of reaction that the protease uses to hydrolyse the substrate as well as the position of the cleavage site. A list of the major possibilities is included in Figure 2C.1. Proteases either cleave a peptide from the termini (exopeptidases) or within the peptide chain (endopeptidases). The exopeptidases are further divided into subfamilies depending on the recognition of the cleavage site as well as the amount of amino acid residues that are cleaved from the terminus.

The second, more widely used classification system is based on the chemical nature of the catalytic site. Using this form of classification, all proteases have been divided into five groups, depending on the residue essential for enzyme activity. These five groups are:

1) Serine peptidases (dependent on a serine residue for catalytic activity)
2) Cysteine peptidases (dependent on a cysteine residue for catalytic activity)
3) Aspartic peptidases (dependent on two aspartic acid residues for catalytic activity)
4) Metallopeptidases (dependent on a metal ion, commonly zinc, for catalytic activity) and
5) Unknown peptidases (of unknown catalytic type)

Figure 2C.1. Classification of peptidases by type of reaction catalyzed. Open circles represent amino acid residues, and filled circles represent cleaved blocks of residues. Triangles indicate blocked termini providing substrates for some of the omega peptidases. Adapted from Barrett, 1994.



Exopeptidases (acts near the end of polypeptide chains)

Aminopeptidases

Dipeptidyl peptidase

Tripeptidyl peptidase

Carboxypeptidases

Peptidyl dipeptidase

Dipeptidases

Omega peptidases

Endopeptidases (acts preferentially in inner regions of polypeptide chains, away from termini)

The third classification system is also the most recently introduced (Rawlings and Barrett, 1993) and was a direct consequence of the enormous amount of gene and genome sequence information that have become available since the early 1990's. This system makes use of the structural and evolutionary relationships between different proteases within a specific catalytic type, as described above, to classify them into clans and families. A family is defined as a group of enzymes in which each member shows evolutionary relationship to at least one other, while a clan comprises a group of families for which there are indications of evolutionary relationship, primarily from the linear order of catalytic site residues and from the tertiary structure (Rawlings and Barrett, 1993). Thus, each catalytic type was assigned a code (S, C, A, M, or U, for serine, cysteine, aspartic, metallo-, and unknown), each clan a letter starting from A, B, C etc and each family a number e.g. 1, 2, 3, etc.

Using this powerful classification system, the serine proteases were subdivided into clans SA, SB, SC, SE, SF and SG (Barrett and Rawlings, 1995), where SA is the chymotrypsins, SB is the subtilases, SC the carboxypeptidases, etc. Thus, subtilisin belongs to the peptidase clan SB, family S8; also known as the subtilisin-like serine proteases. The first three clans of the serine proteases (chymotrypsin, subtilisin and carboxypeptidase) share a common reaction mechanism by all making use of a "catalytic triad" of three amino acids: serine (the nucleophile), aspartate (electrophile), and histidine (base) (Rawlings and Barrett, 1994). The divisions into the different clans are reflected in the arrangement of the catalytic residues in the primary sequence of the proteins, as follows:

SA – H, D, S = Chymotrypsin-like

SB – D, H, S = Subtilisin-like

SC – S, D, H = Carboxypeptidase-like

were H denotes histidine, D denotes aspartate and S denotes serine.

The subtilisin clan (also known as the subtilases or the superfamily of subtilisin-like serine proteases) is the second-largest serine protease clan, after chymotrypsin. Over 200 subtilases were already known in 1997 (Siezen and Leunissen, 1997), and more are discovered every year (Siezen *et al.*, 1991, Siezen and Leunissen, 1997, Brown *et al.*, 2000, Gey van Pittius *et al.*, 2001). This clan is subdivided into six separate families, namely the subtilisin family, the thermitase family, the proteinase K family, the lantibiotic peptidase family, the kexin family and the pyrolysin family. All the members of the subtilases are endopeptidases (or sometimes tripeptidylpeptidases) and are mostly found extracellular (in the case of the bacteria) where they nonspecifically cleave proteins and are required for either defense or growth on proteinaceous substrates (Siezen and Leunissen, 1997). In certain cases, however, members of this clan have developed into highly specialized enzymes of biosynthetic pathways where they are involved in processing and maturation of pro-proteins. These are the lantibiotic peptidases in the bacteria and the kexin pro-protein convertase family of the higher eukaryotes (Siezen and Leunissen, 1997).

To date over 1627 proteases, which can be divided into 42 clans and 220 families, have been identified (Merops peptidase database release 5.7, updated 17/12/2001, http://www.merops.co.uk). Of these, 131 proteases have been identified in the genomes of members of the genus Mycobacterium (*M. avium* -1, *M. fortuitum* - 1, *M. paratuberculosis* - 1, *M. smegmatis* - 6, *M. leprae* -

52, *M. tuberculosis* - 70). However, only a small number of these mycobacterial proteases have been described experimentally. These include the proteases from *M. tuberculosis* (unknown culture filtrate proteases - Reich, 1981, extracellular proteases - Kannan, 1987, elastolytic protease - Rowland, 1997, FtsH - Anilkumar, 1998, pepA and Rv0983 - Skeiky, 1999, mycosins - Brown, 2000, ESAS-7 - Nair *et al.*, 2001), *M. smegmatis* (DD-carboxypeptidase - Eun *et al.*, 1978, 20S proteasome - Knipfer, 1997, lon - Roudiak, 1998, Roudiak and Shrader, 1998), *M. avium* subsp. *paratuberculosis* (pepA - Cameron, 1994), and *M. leprae* (clpC - Misra, 1996). At present, little is known about the diversity of mycobacterial proteases, patterns of expression, substrate specificities, specific functions, and possible links to pathogenesis.

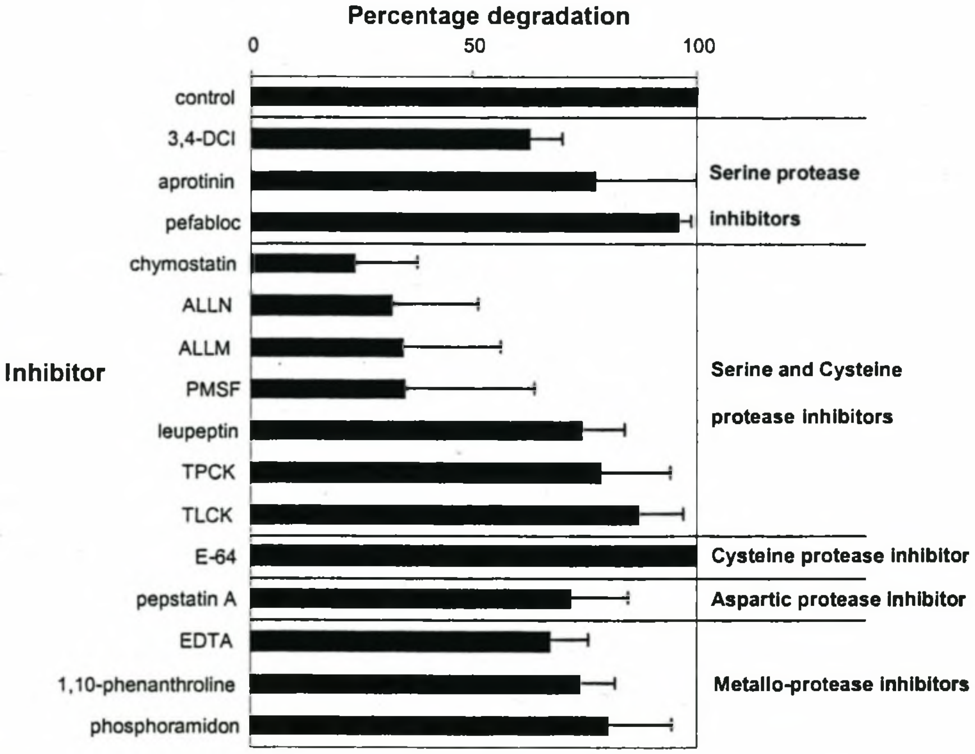
Mycobacterial proteases may contribute to many of the following mechanisms that are essential for the infection process: (1) intracellular nutrition (there is an absence of readily-available nutrients in the macrophage for the infecting organism), (2) modification of the host vacuole by cleavage of vacuolar proteins (to prevent normal lysosomal fusion with the phagosome), (3) modification of other host proteins (e.g. modification of serpins, cytokines, or cytokine receptors), (4) modification of mycobacterial proteins (activation of surface or secreted proteins of *M. tuberculosis* by cleavage), (5) modification of host zymogen cascades (coagulation, complement, fibrinolysis), and (6) tissue necrosis (lung pathology observed during tuberculosis). As described above, a number of studies have linked the secretion or expression of serine proteases on the surface of bacterial cell walls to various diseases caused by pathogens. One appropriate example with regard to lung disease is the injury and degradation by hydrolysis of the major extracellular matrix components of the lungs of patients infected by *Aspergillus fumigatus*, caused by a serine protease secreted by this organism (Iadarola *et al.*, 1998). Although it is widely accepted that host-mediated destruction of lung tissue is the major factor contributing to the lung pathology associated with advanced tuberculosis (Flynn and Chan, 2001), the contribution of extracellularly shedded or secreted proteases of *M. tuberculosis* can not be excluded.

We have previously identified a family of five secreted, cell-wall associated, membrane-anchored subtilisin-like serine proteases in the genome of *M. tuberculosis* H37Rv (Brown *et al.*, 2000). These subtilases were named the mycosins, and are the only subtilisin-like serine proteases in the mycobacteria (Gey van Pittius *et al.*, 2001). Previously, we have examined whether mycosin-1 possessed proteolytic activity, using ^{125}I-fibrinogen as a substrate (J.A. Dave *et al.*, submitted for publication). Assays of fractions of *E. coli* or *M. smegmatis* expressing mycosin-1, as well as purified GST-mycosin-1 fusion protein, revealed no protease activity. Propeptide removal is usually required for activation of secreted bacterial proteases (Wandersman, 1989, Eder *et al.*, 1993, Eder and Fersht, 1995), but pre-treatment of mycosin-1-expressing *M. smegmatis* lysates with acidic buffers or limited tryptic digestion was not successful (J.A. Dave *et al.*, submitted for publication).

We also examined whether protease activity could be detected in culture supernatants of *M. tuberculosis*, as lower molecular weight isoforms of mycosin-1 were identified in culture filtrates and during infection of macrophages (see Addendum 2A). Although highly variable and requiring extended (16 - 18 h) incubations with substrate, protease activity could be detected in *M. tuberculosis* culture filtrates after growth in Kirchner's medium for 2 weeks, reaching a maximum after 4 weeks (J.A. Dave *et al.*, submitted for publication).

To examine this proteolytic activity further, ^{125}I-fibrinogen was incubated with *M. tuberculosis* GSH-3052 culture filtrate in the presence of a diverse array of class-specific protease inhibitors. The effects of different inhibitors provide reliable information for the classification of the catalytic type of a peptidase (Barrett, 1994). This analysis revealed that the proteolytic activity was significantly inhibited by serine and cysteine protease inhibitors (Figure 2C.2), with strongest inhibition obtained with chymostatin (77 ± 14%), ALLN (69 ± 20%), ALLM (66 ± 22%) and PMSF (66 ± 30%)(J.A. Dave *et al.*, submitted for publication). Some subtilisins, including mycosin-1, contain a cysteine residue near the active site histidine, which renders these enzymes susceptible to some cysteine protease inhibitors (Barrett and Rawlings, 1991, Rawlings and Barrett, 1994). Moderate inhibition was also noted with 3,4-dichloroisocoumarin (38 ± 8%) and EDTA (33 ± 9%), whereas inhibition by other inhibitors was variable and generally less than 20%. Inhibition of a protease by a nonspecific chelating agent such as EDTA cannot be taken as a reliable indication that the enzyme is a metalloprotease, though, because the activity of many peptidases of other types are enhanced and activated by cations, notably Ca^{2+}. Such enzymes include members of the subtilisin family (Barret, 1994, Braxton and Wells, 1992).

Figure 2C.2. Degradation of ^{125}I-fibrinogen by *M. tuberculosis* culture filtrates and inhibition by class-specific protease inhibitors. *M. tuberculosis* GSH-3052 culture supernatants were assayed for proteolytic activity by incubation for 16-18 h with ^{125}I-labeled fibrinogen as described (Shephard *et al.*, 1989). Fibrinogen degradation was determined quantitatively by trichloroacetic acid (TCA) precipitation. The inhibition of proteolytic activity of *M. tuberculosis* culture filtrates by class-specific protease inhibitors was investigated by addition of each of the following inhibitors to the reaction: (A) Serine protease specific inhibitors: 3, 4-DCI (3, 4-dichloroisocoumarin) (1 mM), aprotinin (2 µg/ml) and pefabloc (1 mM). (B) Mixed serine and cysteine protease inhibitors: chymostatin (0.2 mM), ALLN (N-acetyl-Leu-Leu-norleucinal or Calpain Inhibitor I) (200 µg/ml), ALLM (N-acetyl-Leu-Leu-methional or Calpain Inhibitor II) (100 µg/ml), PMSF (phenylmethanesulfonyl fluoride) (1 mM), leupeptin (0.4 mM), TPCK (L-1-chloro-3-[4-tosylamido]-4-phenyl-2-butanone) (100 µg/ml) and TLCK (L-1-chloro-3-[4-tosylamido]-7-amino-2-heptanone HCl) (100 µg/ml). (C) Cysteine protease specific inhibitor: E-64 (L-trans-epoxysuccinyl-leucylamido-[4-guanidino]-butane) (0.05 mM). (D) Aspartic protease specific inhibitor: pepstatin A (50 µg/ml). (E) Metallo-protease inhibitors: EDTA (ethylenediaminotetraacetic acid) (10 mM), 1, 10-phenanthroline (2 mg/ml) and phosphoramidon (0.09 mM). Results are expressed as percent degradation relative to control (no inhibitor) and are the means ± standard deviations of four independent experiments. (Reproduced with kind permission from Dr. J.A. Dave, Department of Medicine, Groote Schuur Hospital, Cape Town, South Africa).



Proteolytic activity was enhanced 38% by the addition of 2.5 mM Ca^{2+} but was unaffected by 0.1 mM Zn^{2+}. These results suggested that the predominant protease activity in *M. tuberculosis* culture filtrates comprise one or more serine and/or cysteine proteases that are partially Ca^{2+} dependent, but all attempts to purify this activity failed (J.A. Dave *et al.*, submitted for publication).

It has been observed since early in the 1980's that culture filtrates of *M. tuberculosis* contains proteolytic enzyme activity which are able to hydrolyze extracellular tuberculoproteins (Reich *et al.*, 1981). Serine protease activity has also been observed and two trypsin-like serine proteases (pepA or Rv0125 and Rv0983) have been identified recently (Skeiky *et al.*, 1999). To determine the number of proteases that could potentially contribute to the proteolytic activity observed in the culture filtrate, all possible protease sequences were identified in the *M. tuberculosis* gene database (http://genolist.pasteur.fr/TubercuList/) and are listed in Table 2C.1. This revealed a number of proteases with the potential to be secreted into the extracellular milieu of the organism. It is thus highly likely that there are many different protease activities present in the culture filtrate, of which the majority would belong to the serine and cysteine protease families.

In the present study we aimed to further investigate the protease activity of the mycosins, in order to confirm that these proteins are proteases, to identify the optimal conditions for activity and to identify possible substrates, and thereby to obtain clues to their potential function.

Table 2C.1. Potential proteases present in the genome of *M. tuberculosis* H37Rv*

	Rv Number	Gene Name	Description and Function	Predicted transmembrane region	Predicted signal peptide	Predicted localization	Possible protease activity in cell wall fraction? (Yes/No)	Reference
1	Rv0125	pepA	probable trypsin-like serine protease	1 X (N - terminal)	Yes	Secreted	?	Downing *et al.*, 1999 and Skeiky *et al.*, 1999
2	Rv0185		unknown (neutral zinc metallopeptidases zinc-binding region signature)	None	None	Cytoplasm		
3	Rv0198c		probable zinc metalloprotease	None	None	Cytoplasm		
4	Rv0291	mycP3	subtilisin-like serine protease	2 X (1 X N - terminal, 1 X C - terminal)	Yes	Secreted, C - terminally membrane anchored, enzyme domain in cell wall	Yes	Brown *et al.*, 2000
5	Rv0319	pcp	pyrrolidone-carboxylate peptidase	None	None	Cytoplasm		
6	Rv0359		unknown (neutral zinc metallopeptidases zinc-binding region signature)	6 X (Whole length of protein)	Yes	Secreted, integral membrane protein		
7	Rv0399c	lpqK	possible penicillin binding protein (eg. d-alanyl-d-alanine carboxypeptidase protein)	None	None	Cytoplasm		
8	Rv0418	ipqL	probable aminopeptidase Y	1 X (N - terminal)	Yes	Secreted	?	
9	Rv0419	lpqM	unknown (neutral zinc metallopeptidases zinc-binding region signature)	1 X (N - terminal)	Yes	Secreted	?	
10	Rv0434		unknown (similar to atp-dependent protease la 2)	None	None	Cytoplasm		
11	Rv0457c		probable peptidase	None	None	Cytoplasm		
12	Rv0724	sppA	endopeptidase IV, signal peptide peptidase	None	None	Cytoplasm		Bolhuis *et al.*, 1999, Suzuki *et al.*, 1987 and Ichihara *et al.*, 1986
13	Rv0734	map'	probable methionine aminopeptidase	None	None	Cytoplasm		

14	Rv0773c	ggtA	putative g-glutamyl transpeptidase	None	None	Cytoplasm		
15	Rv0781	ptrBb	protease II, b subunit	None	None	Cytoplasm		Kanatani *et al.*, 1992
16	Rv0782	ptrBa	protease II, a subunit	None	None	Cytoplasm		Kanatani *et al.*, 1992
17	Rv0800	pepC	aminopeptidase I	None	None	Cytoplasm		Frederick *et al.*, 1993
18	Rv0838	lpqR	unknown (similarity to d-alanyl-d-alanine dipeptidase)	None	Yes	Secreted	?	
19	Rv0840c		probable proline iminopeptidase	None	None	Cytoplasm		
20	Rv0983		probable trypsin-like serine protease	1 X (N - terminal)	Yes	Secreted	?	Skeiky *et al.*, 1999
21	Rv1191		unknown (some similarity to proline iminopeptidase)	None	None	Cytoplasm		
22	Rv1223	htrA	DegP protease serine protease homologue	1 X (internal)	None	Membrane anchored, enzyme domain in cytoplasm		Kim *et al.*, 1999 and Poquet *et al.*, 2000
23	Rv1539	lspA	probable lipoprotein signal peptidase	4 X (whole length of protein)	None	Integral membrane protein		
24	Rv1577c		unknown (similarity to putative bacteriophage HK97 prohead protease (gp4))	None	None	Cytoplasm		
25	Rv1796	mycP5	subtilisin-like serine protease	2 X (1 X N - terminal, 1 X C - terminal)	Yes	Secreted, C - terminally membrane anchored, enzyme domain in cell wall	Yes	Brown *et al.*, 2000
26	Rv1887		unknown (eukaryotic thiol (cysteine) proteases histidine active site at N-terminus)	None	None	Cytoplasm		
27	Rv1922		possible penicillin binding protein (eg. d-alanyl-d-alanine carboxypeptidase protein)	1 X (N - terminal)	Yes	Secreted	?	
28	Rv1977		unknown (neutral zinc metallopeptidases zinc-binding region signature)	None	None	Cytoplasm		

29	Rv1983		unknown (Eukaryotic and viral aspartyl proteases active site)	None	Yes	Secreted	?	
30	Rv2008c		unknown (Signal peptidases I serine active site)	None	None	Cytoplasm		
31	Rv2089c	pepE	probable pepQ, cytoplasmic peptidase	None	None	Cytoplasm		
32	Rv2109c	prcA	proteasome a-type subunit 1	None	None	Cytoplasm		
33	Rv2110c	prcB	proteasome b-type subunit 2	None	None	Cytoplasm		
34	Rv2141c	dapE2	ArgE/DapE/Acy1/Cpg2/yscS family (similarity to carboxypeptidase S precursor)	None	None	Cytoplasm		
35	Rv2213	pepB	probable pepA, similar to aminopeptidases	None	None	Cytoplasm		
36	Rv2223c		probable exported protease	1 X (N - terminal)	Yes	Secreted	?	
37	Rv2224c		probable exported protease	1 X (N - terminal)	Yes	Secreted	?	
38	Rv2394	ggtB	probable gamma-glutamyltranspeptidase precursor	1 X (N - terminal)	Yes	Secreted	?	Shetty *et al.*, 1981
39	Rv2457c	clpX	ATP-dependent Clp protease ATP-binding subunit ClpX	None	None	Cytoplasm		
40	Rv2460c	clpP2	ATP-dependent Clp protease proteolytic subunit	None	None	Cytoplasm		
41	Rv2461c	clpP	ATP-dependent Clp protease proteolytic subunit	None	None	Cytoplasm		
42	Rv2467	pepD	probable aminopeptidase	None	None	Cytoplasm		
43	Rv2515c		unknown (neutral zinc metallopeptidases zinc-binding region signature)	None	None	Cytoplasm		
44	Rv2535c	pepQ	probable pepQ cytoplasmic peptidase	None	None	Cytoplasm		
45	Rv2575		unknown (neutral zinc metallopeptidases zinc-binding region signature)	1 X (N - terminal)	None	Membrane anchored, enzyme domain in cytoplasm		
46	Rv2625c		unknown (neutral zinc metallopeptidases zinc-	6 X (N - terminal)	None	Integral membrane protein, C-		

			binding region signature)			terminus in cytoplasm		
47	Rv2651c		unknown (similarity to putative bacteriophage HK97 prohead protease (gp4))	None	None	Cytoplasm		
48	Rv2667	clpX'	similar to ATP-dependent ClpC protease from M. leprae but shorter	None	None	Cytoplasm		
49	Rv2672		putative exported protease	1 X (N - terminal)	Yes	Secreted	?	
50	Rv2725c	hflX	hflA bacterial membrane-bound ATP-dependent protease GTP-binding protein	None	None	Cytoplasm		Noble *et al.*, 1993
51	Rv2782c	pepR	probable protease	None	None	Cytoplasm		
52	Rv2861c	map	methionine aminopeptidase	None	None	Cytoplasm		
53	Rv2870c		unknown (zinc carboxypeptidases, zinc-binding region2 signature)	None	None	Cytoplasm		
54	Rv2903	lepB	signal peptidase I	1 X (N - terminal)	None	Membrane anchored, enzyme domain in cytoplasm		
55	Rv3207c		unknown (neutral zinc metallopeptidases zinc-binding region signature)	1 X (N - terminal)	Yes	Secreted	?	
56)	Rv3305c	amiA	probable N-acyl-L-amino acid amidohydrolase or peptidase	None	None	Cytoplasm		
57	Rv3306c	amiB	probable aminohydrolase	None	None	Cytoplasm		
58	Rv3365c		unknown (neutral zinc metallopeptidases zinc-binding region signature)	2 X (N - terminal)	Yes	Secreted, N-terminally membrane anchored, enzyme domain in cytoplasm		
59	Rv3419c	gcp	glycoprotease	None	None	Cytoplasm		
60	Rv3449	mycP4	subtilisin-like serine protease	2 X (1 X N - terminal, 1 X C - terminal)	Yes	Secreted, C - terminally membrane anchored, enzyme domain in cell wall	Yes	Brown *et al.*, 2000

61	Rv3596c	clpC	probable ATP-dependent Clp protease ATP-binding subunit	None	None	Cytoplasm		
62	Rv3610	ftsH	membrane bound ATP-dependent zinc protease	2 X (N - terminal)	Yes	Secreted, N-terminally membrane anchored, enzyme domain in cytoplasm		Makino *et al*., 1999
63	Rv3626c		unknown (neutral zinc metallopeptidases zinc-binding region signature)	None	None	Cytoplasm		
64	Rv3668c		probable alkaline serine protease	1 X (N - terminal)	Yes	Secreted	?	
65	Rv3671c		probable trypsin-like serine protease	4 X (N - terminal)	None	Integral membrane protein, C-terminus in cytoplasm		
66	Rv3836		unknown (neutral zinc metallopeptidases zinc-binding region signature)	None	None	Cytoplasm		
67	Rv3883c	mycP1	subtilisin-like serine protease	2 X (1 X N - terminal, 1 X C - terminal)	Yes	Secreted, C - terminally membrane anchored, enzyme domain in cell wall	Yes	Brown *et al*., 2000
68	Rv3886c	mycP2	subtilisin-like serine protease	2 X (1 X N - terminal, 1 X C - terminal)	Yes	Secreted, C - terminally membrane anchored, enzyme domain in cell wall	Yes	Brown *et al*., 2000

* Obtained by searching Tubercullst website (http://genolist.pasteur.fr/TubercuList/) with the terms protease, proteinase and peptidase.

2C.2. Materials and Methods

2C.2.1. Bacterial strains and plasmids

Escherichia coli JM109 was used as a host for the expression of recombinant mycosin fusion proteins and *Mycobacterium smegmatis* mc^{2}155 (Snapper *et al.*, 1990) as a heterologous host for the expression of the cloned full-length *M. tuberculosis* mycosins. The construction of the plasmids containing the mycosin fusion proteins (pGex-P1, pMalC-P2 and pGex-P3) as well as the native mycosin clones for expression in *M. smegmatis* (p19K-P1, p19K-P2RBS and p19K-P3) was described by Brown *et al.* (2000)(Chapter 2). *M. tuberculosis* H37Rv (laboratory strain; ATCC 25618) and *M. tuberculosis* clinical isolate GSH-3052 (from a pleural effusion)(Cywes *et al.*, 1997) were used for the characterization of the natively expressed mycosins. The clinical isolate was a gift from Dr. J.A. Dave (Department of Medicine, Groote Schuur Hospital, Cape Town, South Africa) and was originally obtained from the Bacteriology Laboratory, Department of Medical Microbiology, Groote Schuur Hospital, Cape Town, South Africa.

2C.2.2. Media and culture conditions

E. coli was grown on solid or liquid Luria-Bertani (LB) medium as described by Sambrook *et al.* (1989). Mycobacterial strains were grown at 37°C for 2 days with shaking (200 rpm, *M. smegmatis*) or 14 days with stirring (*M. tuberculosis*) in Middlebrook 7H9 broth (Difco) supplemented with filter-sterile ADC (0.5% BSA, 0.2% glucose, 0.015% catalase) and containing 0.05% Tween 80 (Sigma). For the *M. tuberculosis* clinical isolate GSH-3052 fractionation samples, 100 ml of Kirchener's medium (3 g/l Na_2HPO_4, 4 g/l KH_2PO_4, 1.07 g $MgSO_4.7H_2O$, 2.5 g/l Tri-sodium citrate, 20% glycerol, 5 g/l asparagine) was inoculated with the *M. tuberculosis* GSH-3052 and incubated with stirring for 4 weeks at 37°C. All work on *M. tuberculosis* H37Rv and *M. tuberculosis* clinical isolate GSH-3052 was done in a Biosafety Level III facility. Hygromycin (100 μg/ml, Roche) and Ampicillin (50 μg/ml, Roche) were added to bacterial cultures when antibiotic selection was required. For plate activity assays on solid media, *M. smegmatis* were grown on Middlebrook 7H11 agar supplemented with filter-sterile OADC (0.005% oleic acid, 0.5% BSA, 0.2% glucose, 0.02% catalase, 0.085% NaCl) and containing 0.05% Tween 80 (Sigma) as well as 0.025 g/ml skim milk powder.

2C.2.3. Protease activity assays sample preparation

For protease activity assays involving purified recombinant mycosins, purified proteins were obtained as described in Brown *et al.* (2000). During the purification and preparation all protease inhibitors were omitted. For the other activity assay samples, mycobacterial cultures (10 ml) were grown to an optical density of 1. The cells were pelleted by centrifugation, resuspended in 500 µl phosphate-buffered saline (PBS) and 1% NP40 and either sonicated at 4.5 setting in a Misonix cup sonicator on ice for a total of 5 minutes (15-second bursts with 30-second intervals), freeze-thawed (15 x in liquid nitrogen) or fast-prepped (with glass-beads - 6.5 power for 45 seconds) to disrupt the cells without denaturing the proteases.

For the PepTag assays, in addition to the samples described above, another set of samples was prepared by immunoprecipitation of native mycosin-2 from *M. smegmatis* expressing mycosin-2. Polyclonal positively- and negatively-selected anti-mycosin-2 antisera (Brown *et al.*, 2000, Chapter 2) were absorbed to Protein A sepharose beads at 4°C for 1h with rotation. The beads were washed with phosphate-buffered saline (PBS), pelleted by centrifugation, and the mycosin-2 expressing *M. smegmatis* lysate was added to the sample. This was allowed to immunoprecipitate at 4°C overnight with rotation. The immunoprecipitate was washed four times with PBS and used in the PepTag protease assays. Presence of the absorbed proteins was determined by separation of the samples on SDS-PAGE and subsequent Western blot analyses. Similarly, a vector-expressing *M. smegmatis* lysate was immunoprecipitated to act as a negative control.

For the *M. tuberculosis* clinical isolate GSH-3052 fractionation samples, the culture was centrifuged at 3000 *x g* for 20 minutes, after which the supernatant was removed. The supernatant contained the culture proteins and was filtered through 1.0, 0.45 and 0.22 micron filters, dialyzed at 4ºC overnight against PBS and concentrated with Centriprep concentrator (Amicon). The pellet was resuspended in 2 ml PBS and ribolyzed (6.5 power for 45 seconds). After standing for 10 minutes on ice, the sample was centrifuged for 20 minutes at low speed to remove unbroken cell residues. 0.33% NP40 was added to the pellet, which contained the whole cell lysate fraction. The supernatant was filter sterilized through a 0.22 micron filter, subfractionationated, and centrifuged twice for 1h at 27 000 *x g*, whereafter the pellet contained the cell wall fraction. The pellet was resuspended in PBS containing 0.33% NP40. The supernatant was further centrifuged at 100 000 *x g* for 2 hours and the resulting pellet contained the membrane fraction while the supernatant contained the cytosol. The pellet was resuspended in PBS containing 0.33% NP40 and 0.33% NP40 was also added to the cytosol. Sodium azide (5 mM NaN_3) was added to all samples to prevent microbial growth. As there is no available marker to test for the purity of the mycobacterial fractions, the fractions were assayed using three enzyme assays, namely isocitrate dehydrogenase (a marker for cytoplasmic proteins), catalase (a marker for cell surface proteins) and aminopeptidase (a marker for extracellular proteins).

During all assays, between 1 and 10 ng of purified commercial subtilisin (Roche), proteinase K (Roche), trypsin (Roche) or alkaline protease (Promega) was used as a positive control, depending on the specific assay. Unless stated otherwise, samples were assayed in the presence of calcium (10 mM $CaCl_2$) at all times, while magnesium (10 mM MgCl) and zinc (0.1 mM $ZnSO_4$) was also added in certain cases to determine the effect of metal ions. To determine the effect of pH and temperature, assays were incubated at different temperatures and under different buffer conditions. In an attempt to activate the zymogens, some of the samples were either subjected to limited proteolysis (with commercial subtilisin), low pH treatment (pH 3.3), the addition of dithiothreitol (DTT, 10mM), or by heat activation (65°C).

2C.2.4. Plate protease activity assays

Plate activity assays depend on the radial diffusion of proteolytic enzymes secreted from bacterial colonies into agar containing skim milk as substrate. This results in cleared zones surrounding the colonies, which is the result of the proteolysis of the skim milk proteins. Plates were prepared by adding sterile skim milk solution in H_2O to autoclaved agar medium (LB for *E. coli* and Middlebrook 7H11 for *M. smegmatis*). *E. coli* expressing the mycosin fusion proteins, as well as *M. smegmatis* expressing wild type mycosins were assayed by spreading the bacteria onto the skim milk agar plates and incubating at 37°C for the desired time interval. Results were visually evaluated by the clearing of zones surrounding the bacterial colonies.

2C.2.5. Zymogram protease activity assays

Protease activity was assayed by zymography by applying samples to a 10% SDS-polyacrilamide gel containing co-polymerized gelatin or casein (1 mg/ml) as described by Heussen *et al.* (1980). After electrophoresis at 4°C (to prevent autodigestion), SDS was removed by successive washes in 2.5% (w/v) Triton X-100 in distilled H_2O for 20 minutes, 2.5% (w/v) Triton X-100 in 50 mM Tris-HCl (pH 7.4) for 20 minutes, and 50 mM Tris-HCl (pH 7.4) for 20 minutes. The gel was then incubated overnight at 37°C in 50 mM Tris-HCl (pH 7.4), fixed in 10% methanol / 10% acetic acid for 10 minutes, stained with Coomassie brilliant blue R-250 followed by destaining with 10% methanol / 10% acetic acid until the proteolytic band was observed.

2C.2.6. PepTag protease activity assays

This commercial assay (Promega) makes use of two fluorescent dye-linked peptides, the A1 peptide (Dye-L-R-R-A-S-L-G) and C1 peptide (Dye-P-L-S-R-T-L-S-V-A-A-K), as substrates for proteases. Proteolytic cleavage of the peptides changes the peptide's mobility in an agarose gel by altering the peptide's net charge and size. This system detects picogram quantities of proteases in solution after a 30 minute incubation, followed by a 15 - 30 minute agarose gel electrophoresis. The assay was done according to the manufacturer's conditions. Briefly, 3 μl of either peptide was added to the sample to be tested. For a negative control, sample was added to buffer only, while alkaline protease (20 - 100 ng) was added to another sample for a positive control. The samples were incubated under the required conditions, whereafter 1 μl of 80% glycerol was added and the samples were analysed by agarose gel electrophoresis.

2C.2.7. FITC-casein protease activity assays

The fluorescein isothiocyanate-labeled casein (FITC-casein) assay for proteolytic enzyme activity was done essentially as described by Twining (1984). Briefly, 20 μl of 0.5% FITC-labelled casein (Sigma) was added to 20 μl of a desired assay buffer (depending on pH being measured). The reaction mixture was completed by adding 10 μl of protease sample and were incubated at 37°C for 1 hour, after which 120 μl of 5% TCA were added. The sample was left to stand at room temperature for 1h and centrifuged for 5 minutes in a microfuge to sediment the TCA-insoluble protein. A 60 μl aliquot of the supernatant was diluted to 1 ml with 500 mM Tris-HCl (pH 8.5) and the fluorescence was measured in a flourometer at excitation 490 nm and emission 525 nm.

2C.2.8. Azo-casein protease activity assays

For the azocasein protease activity assay, 20 µl of the protease sample was added to a 460 µl 50 mM Tris-HCl buffer (pH 7.5). 20 µl of a 5% azocasein (Sigma) solution in 0.2 M Tris-HCl (pH 7.5) and containing 1 mM $CaCl_2$, was added to the mixture and incubated for 30 minutes at 37°C. After incubation, 500 µl of 10% trichloroacetic acid (TCA) was added, the mixture was incubated for 15 minutes on ice, and the sample was centrifuged for 2 minute at maximum speed in a microfuge. The supernatant was removed and 800 µl of the supernatant was added to 200 µl 1.8 N NaOH, whereafter the optical density was read at 420 nm.

2C.2.9. Gene sequence analyses

Annotations, descriptions and protein sequences of individual genes from the *M. tuberculosis* H37Rv genome were obtained from the publicly available genome sequence database for *M. tuberculosis* H37Rv (http://genolist.pasteur.fr/TubercuList/). Protein sequences of the five mycosins (Rv3883c, Rv3886c, Rv0291, Rv3449, and Rv1796) as well as the *Bacillus* sp. NKS-21 subtilisin ALP1 (Genbank accession number D29736) were aligned using ClustalW 1.5 on the ClustalW WWW server at the European Bioinformatics Institute website (http://www2.ebi.ac.uk/clustalw/; Thompson *et al.*, 1994). These protein sequences were visually analysed for common subtilase residues and motifs and primary structural features according to Siezen and Leunissen (1997).

2C.3. Results and Discussion

2C.3.1. Plate protease activity assays

Colonies of *E. coli* expressing recombinant mycosin-1, -2 and -3 fusion proteins were grown on LB agar containing skim milk powder as a protein source. No cleared zones were detected on these plates, indicating that no proteins containing protease activity were secreted. A similar assay was done using *M. smegmatis* heterologously expressing the native *M. tuberculosis* mycosin proteases and grown on Middlebrook 7H11 agar containing skim milk as protein substrate. Once again, no secreted protease activity could be detected.

2C.3.2. Zymogram protease activity assays

Samples were tested for protease activity by zymogram analysis, by separating the samples on an SDS-PAGE gel co-polymerized with either casein or gelatin. Renaturation and incubation in the desired assay buffer should result in a clearing in the gel surrounding the separated protease protein band. Although various parameters were changed during these analyses (pH, temperature, length of incubation, non-denaturing, calcium etc.), no protease activity of the right molecular size could be detected in any of the samples. This was not due to the assay not working, as the control enzyme resulted in a clear zone in each case (results not shown). In certain instances, enzyme activity of a much larger size than the expected was obtained (Figure 2C.3). As this activity was obtained in both the control lysate and the test lysate, it was clear that this was due to another protease expressed by *M. smegmatis*.

2C.3.3. PepTag protease activity assays

The protease activity of various samples was assayed using the extremely sensitive commercial PepTag protease activity assay system. No protease activity could be detected for any of the recombinant mycosin fusion proteases (Figure 2C.4A). Although protease activity could be detected in wild-type *M. smegmatis* and mycosin-expressing *M. smegmatis* whole cell lysate samples, these activities could never be attributed to the mycosins as such, as the activity was in most cases observed in both lysates (see for example Figure 2C.4B).

Figure 2C.3. Zymogram protease activity assays. Samples tested for protease activity by zymogram analysis were seperating on SDS-PAGE gel co-polymerized with either casein or gelatin. Protease activity is detected by clear bands on stained gels. Lanes (1) Wild type *M. smegmatis* whole cell lysate as control, (2) whole cell lysate of *M. smegmatis* expressing mycosin-2. A clear band showing native background *M. smegmatis* protease activity is indicated by an arrow.

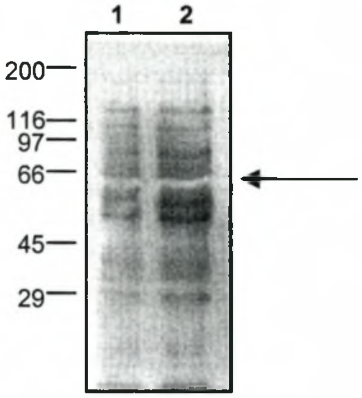
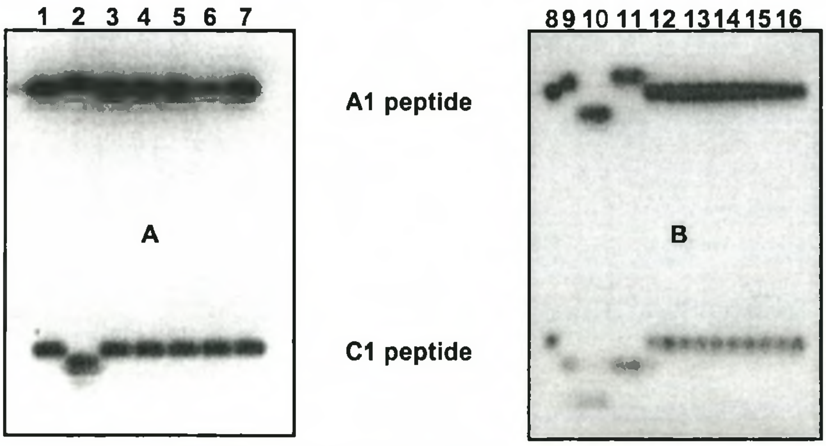


Figure 2C.4. PepTag protease activity assays. Two fluorescent dye-linked peptides (A1 and C1) were used as substrates for detecting protease activity. Proteolytic cleavage of the peptides changes the mobility in an agarose gel by altering the peptide's net charge and size. (A) Assay using purified recombinant mycosin fusion proteins. Lanes (1) control uncleaved peptide, (2) Alkaline protease positive control, (3) GST fusion partner protein, (4) MBP fusion partner protein, (5) mycosin-1-GST, (6) mycosin-2-MBP, (7) mycosin-3-GST. (B) Assay using culture supernatants of wild-type *M. smegmatis* and *M. smegmatis* expressing mycosin-2. Lanes (8) control uncleaved peptide, (9) Alkaline protease positive control, (10-16) wild-type *M. smegmatis* and mycosin-2 expressing *M. smegmatis* whole cell lysates alternately loaded in decreasing concentrations from 910 µg/ml to 1 ng/ml.



2C.3.4. FITC-casein protease activity assays

FITC-casein protease activity assays were done on different samples and results were measured using a fluorometer. The results of one of the analyses are indicated in Table 2C.2 and reveals that although high levels of activity could be detected with the control subtilisin, no activity could be detected in any of the other samples.

Table 2C.2. FITC-Casein protease activity assay

Sample	**whole cells**	**lysate**	**low pH**
subtilisin control (10ng)	883.1	849.1	848.1
p19kDpro vector control	0	4.1	0
p19K-P1 (mycosin-1)	0.1	0	0
p19K-P2 (mycosin-2)	0	2.1	0
p19K-P3 (mycosin-3)	0	0	4.1

All results are arbitrary and are calculated as actual reading - blank (31.9 in this assay).

2C.3.5. Azo-casein protease activity assays

Azocasein was used as a substrate for the final protease activity assays using different samples including the *M. tuberculosis* clinical isolate GSH-3052 fractionation samples. Once again, no protease activity could be attributed to the recombinant mycosins. Experiments were also carried out to complement the activity of *M. smegmatis* to comparable levels of *M. bovis* BCG (which contains one extra mycosin). Protease activity could be detected in all the samples tested, with higher levels of protease activity in the *M. smegmatis* cells expressing mycosin-1 and 3 (comparable to the activity of *M. bovis* BCG). However, these differences were not significant as only low levels of protease activity, comparable to wild-type *M. smegmatis* cells, were detected in *M. smegmatis* cells expressing mycosin-2 (Table 2C.3).

Table 2C.3. Azocasein protease activity assays

Sample	**OD_{420}**
M. smegmatis cells	0.529
p19K-P1 cells	1.405
p19K-P2 cells	0.060
p19K-P3 cells	1.407
M. bovis BCG cells	1.695
E. coli JM109 cells	0.224

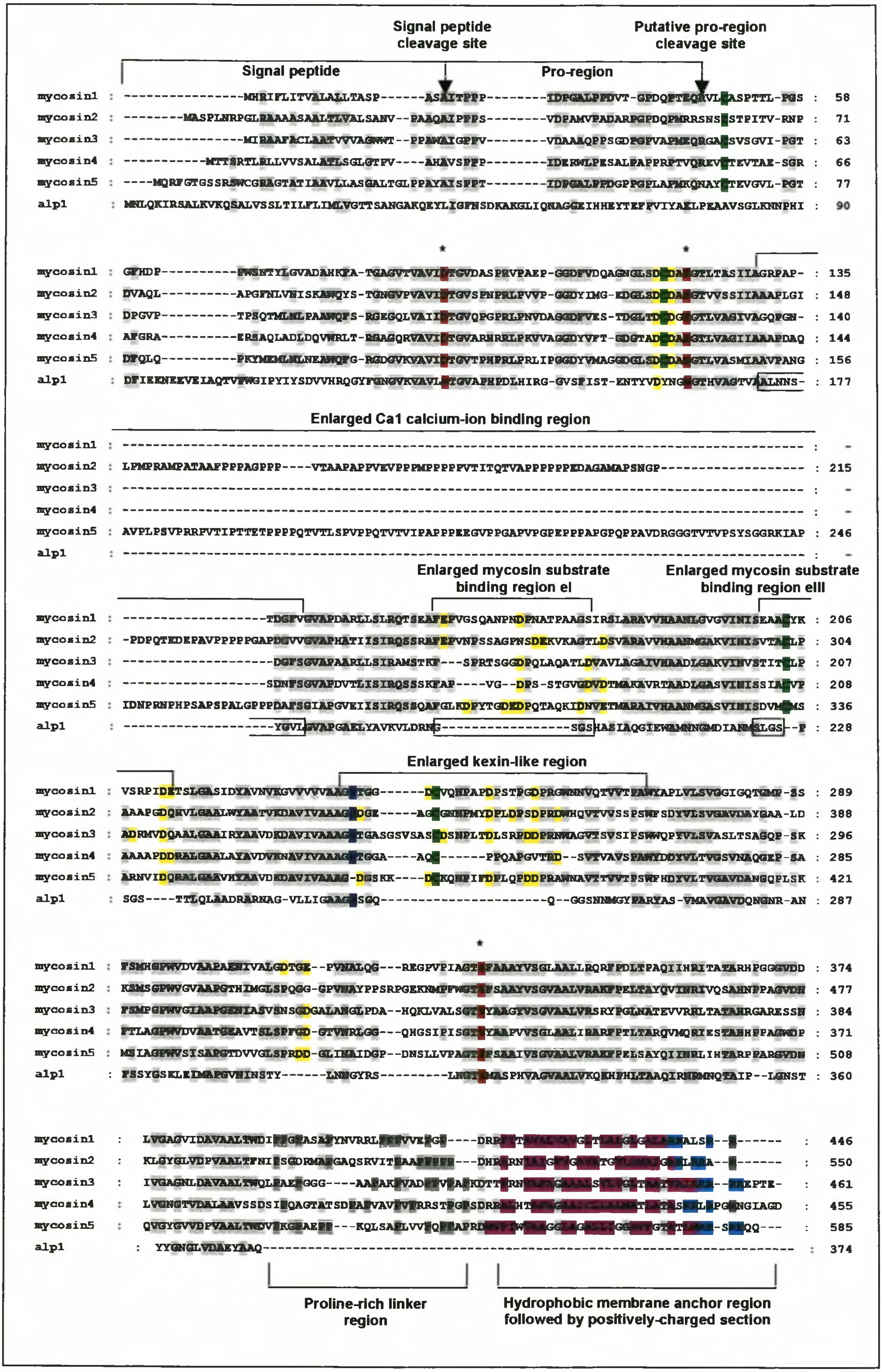
To summarize, the protease activity assays done in the present and previous studies clearly demonstrates the presence of serine protease activities in mycobacterial whole cell lysates as well as culture filtrates. However, these activities could not be attributed to the mycosins *per se* as no differences could be detected between controls and test samples. Furthermore, the recombinant mycosin fusion proteins did not show any protease activity under any of the conditions tested. It was thus concluded that these proteases must require a specific substrate or conditions for activity.

2C.3.6. Gene sequence analyses

To obtain clues for the elucidation of the function and substrate specificity of the mycosin proteases, and to look for structural reasons that could explain why no protease activity could be detected in any of the assays, we did a complete sequence analysis of the protein sequences of all five the mycosin proteases, according to the common subtilase motifs as described by Siezen and Leunissen (1997). A multiple sequence alignment of the mycosin proteases are presented in Figure 2C.5.

The five mycosins of *M. tuberculosis* are more closely related to each other than to any other known subtilase sequence, and therefore they must have evolved from each other through gene duplication. All members contain an N-terminal signal sequence with a signal peptidase I cleavage site (AxA//x) as well as a C-terminal hydrophobic domain, followed by a short positively charged segment, that could act as a transmembrane anchor (Figure 2C.5, Figure 3A.8).

Figure 2C.5. Conserved features of the mycosins. A multiple sequence alignment of the five mycosin subtilisin-like serine proteases and *Bacillus* sp. NKS-21 subtilisin ALP1 (Genbank accession number D29736) showing the catalytic triad residues (D90, H121 and S332, mycosin-1 numbering; indicated by asterisks and in red). Conserved and semi-conserved residues are indicated in light grey. The putative signal peptide and pro-region cleavage sites are indicated by arrows. Conserved cysteine residues are indicated in green, acidic residues in or near substrate binding sites are indicated in yellow and the oxyanion-hole residue (N237, mycosin-1 numbering) is indicated in dark blue. Light blue residues indicate hydrophobic residues of the membrane anchor and purple residues the hydrophilic basic residues forming the positively charged segment of the membrane anchor. The proline-rich linker region is indicated (prolines are indicated in dark grey).


Signal peptide
cleavage site
Putative pro-region
cleavage site
Signal peptide
Pro-region
mycosin1
mycosin2
mycosin3
mycosin4
mycosin5
alp1
Enlarged Ca1 calcium-ion binding region
Enlarged mycosin substrate
binding region eI
Enlarged mycosin substrate
binding region eIII
Enlarged kexin-like region
Proline-rich linker
region
Hydrophobic membrane anchor region
followed by positively-charged section

All the mycosins contain the subtilase conserved active site residues Asp-His-Ser (Figure 2C.5, indicated by asterisks and highlighted in red). It is commonly accepted that if a protein includes at least two of the three active site signatures, the probability of it being a serine protease from the subtilase family is 100% (PROSITE database, http://www.expasy.org/prosite/, Bairoch, 1991). In addition to the catalytic triad, the regions displaying the highest percentage of conservation all correspond to the 3D structurally conserved core of the subtilases, so the overall macromolecular structure should be consistent with the structures of subtilisin or thermitase (Siezen and Leunissen, 1997). Four of the five mycosins also display the conserved asparagine residue at the oxyanion hole (highlighted in dark blue in Figure 2C.2), with only the oxyanion hole residue in mycosin-5 being substituted by an aspartic acid (D367, mycosin-5 numbering). In all subtilisins, the oxyanione hole is formed by the active site residue serine and the residue asparagine-262 (preprosubtilisin BPN' numbering). The N-residue (in a conserved segment AAGN) helps to stabilize the oxyanion generated in the tetrahedral transition state (Carter and Wells, 1990). There is only one other known exception where the conserved asparagine residue is substituted, and this is in the mammalian furin known as PC2 (which form part of the PC2 subgroup of the kexin family)(Siezen and Leunissen, 1997). This protease has an aspartic acid residue (D) in the place of the Asn-262 in the oxyanione hole, which is the same substitution that is observed in mycosin-5. This substitution was shown not to influence the catalytic efficiency of the pro-protein processing protease PC2 (Zhou *et al.*, 1995).

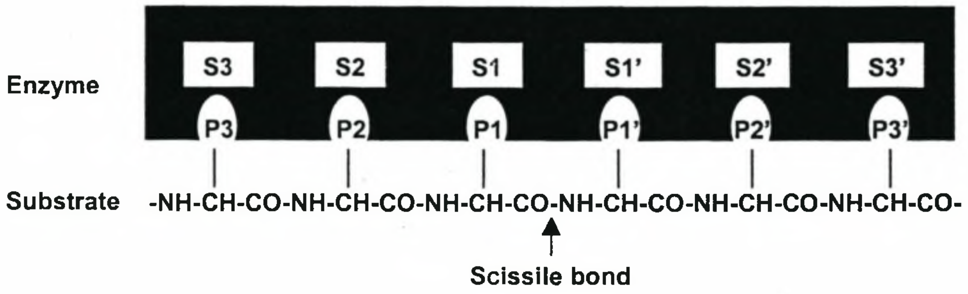
In Gram-positive bacteria extracellular proteins normally contain NH_2-terminal signal peptides that direct the protein to the membrane for secretion (Pugsley and Schwartz, 1985, Pugsley, 1993). This is also the case for all secreted or extracellularly located proteases (Wong and Doi, 1986, Power *et al.*, 1986). In addition to this, all known cellular and bacterial proteolytic enzymes are produced as inactive precursors (zymogens), containing an additional polypeptide segment (pro-sequence) located C-terminally to the signal sequence (the pre- sequence), that needs to be auto- or transproteolytically cleaved off to reveal active enzyme (Khan and James, 1998). Activation is rapid and irreversible and the cleaved activation segment is usually degraded by the activated enzyme. This propeptide has several functions including: (1) inhibiting and maintaining the protease inactive, (2) promoting correct folding of the protease and stabilizes the protein, (3) alters protease specificity, (4) act as an anchor in the membrane, and (5) potentially plays a role in secretion (Ikemura *et al.*, 1987, Baker *et al.*, 1992, Eder and Fersht, 1995). Almost all extracellular subtilases also have, in addition to the signal peptide, a propeptide that is (auto)proteolytically cleaved to obtain active enzyme (Eder *et al.*, 1993, Hu *et al.*, 1996). If another mycobacterial protease is required for this maturation step, this may explain why the recombinant mycosins expressed in *E. coli* are inactive. The mycosins have a fairly short predicted propeptide of about 30 - 40 residues (instead of the normal 70 aa), and the first part of the predicted propeptide has 5 - 6 conserved prolines, which is very unusual. This highly conserved region of the propeptide ends just before the cysteine residue at position 68 (mycosin-5 numbering, Figure 2C.5).

There are only four totally conserved cysteines in all five mycosins (something which is not normally found in bacterial subtilisins; highlighted in green in Figure 2C.2), which is highly significant and suggests two conserved disulfide bridges (S-S). As extracellular enzymes of the Gram-positive bacteria rarely contain disulfides, it is hypothesized that these proteases may be secreted (due to the signal peptide) and then subsequently situated inside the cell wall of the organism. Disulfide bridges can contribute to the overall stability of the protein, and two disulfide bonds are also found in the subtilases of the kexin/furin family of proprotein processing proteases (Van de Ven *et al.*, 1990, Siezen *et al.*, 1994). Thus, the first conserved cysteine would be part of the mature enzyme, and not part of the cleaved propeptide, which suggests that the propeptide is cleaved just before this first C-residue, probably directly after the partially conserved TEQR/MEQR/TVQR/MKQN sequence (Figure 2C.5). This first C-residue would hold the new N-terminus closely bound to the mature enzyme after propeptide cleavage, making degradation of the protease by aminopeptidase activity virtually impossible. The positions of the conserved C-residues signify a disulphide bridge between the third and fourth C-residues (residue 334 and 373, mycosin-5 numbering) because these domains normally interact to contribute to the formation of the S1 substrate binding pocket (see Figure 2C.7, residues 129 and 160, subtilisin BPN' numbering). The first C-residue must therefore link to the second residue situated close to the catalytic active site residue histidine. Reducing agents in the protease activity assay buffer may therefore inactivate the enzymes by reducing these S-S bridges. The eukaryotic proprotein processing subtilases (kexin/furin-type subtilisins) also contains a cysteine near the active site histidine (although on the other side of the histidine) that confers a requirement for thiol activation to these enzymes (Rawlings and Barrett, 1994). Although thiol activation was attempted in the mycosin activity assays, this did not result in protease activity. It was previously observed that although only one fragment is likely to possess the transmembrane anchor, both mycosin-2 fragments were present in the cell wall/membrane fraction when heterologously expressed in *M. smegmatis* (Brown *et al.*, 2000). The localisation of both fragments in the membrane of *M. smegmatis* indicates that the two fragments are associating in some way but the disulphide linkages between the two fragments could not be detected by analysis on non-reducing SDS-PAGE (data not shown).

There are several large amino acid sequence insertions into loop regions of the mycosins, probably classifying these subtilases in the pyrolysin subfamily of subtilases (a heterogeneous group of low sequence conservation and characterized by large insertions, Siezen and Leunissen, 1997). These are unusually rich in proline, and most of the prolines seem to occur in the extra loop regions, suggesting low flexibility of these loops and possibly reduced susceptibility to proteolysis. These loop regions are approximately from residues 40-60, 153-266, 370-387, 525-552 (mycosin-5 numbering). The activity of most subtilisins is stimulated by Ca^{2+} (Barrett and Rawlings, 1991, Braxton and Wells, 1992), which is bound to the molecule by up to four calcium-ion binding sites, and is essential for stability and activity (Siezen *et al.*, 1995, Siezen and Leunissen, 1997). The most common calcium binding sites in the subtilisins were predicted to be the Ca1 (strong) and Ca3 (weak) sites, with the medium strength site Ca2 being less common (Siezen *et al.*, 1991). Very exceptional extra large proline-rich inserts occur in mycosin-2 (Rv3886c) and mycosin-5 (Rv1796). These occur in the region that is supposed to contain the Ca1 major calcium-ion binding site, thus it is predicted that this site is not likely to be present or is disrupted in the members of the mycosin family, placing a question mark on the calcium requirements for these proteases. Such large inserts have never been found before in that position in any subtilase. The nisin lantibiotic processing peptidase NisP also contains a number of insertions into the catalytic domain and has also been shown to not contain the strong Ca1 binding site (Siezen *et al.*, 1995). In addition to this, the Ca3 (weak) and Ca4 (weak) binding sites were also absent, with only the medium strength site Ca2 being present. Closer inspection of the mycosin sequences revealed a potential for only the weak Ca3 region to be present, although it is possible that there may be other novel Ca^{2+} binding sites present in the large inserts in the protein sequences (data not shown).

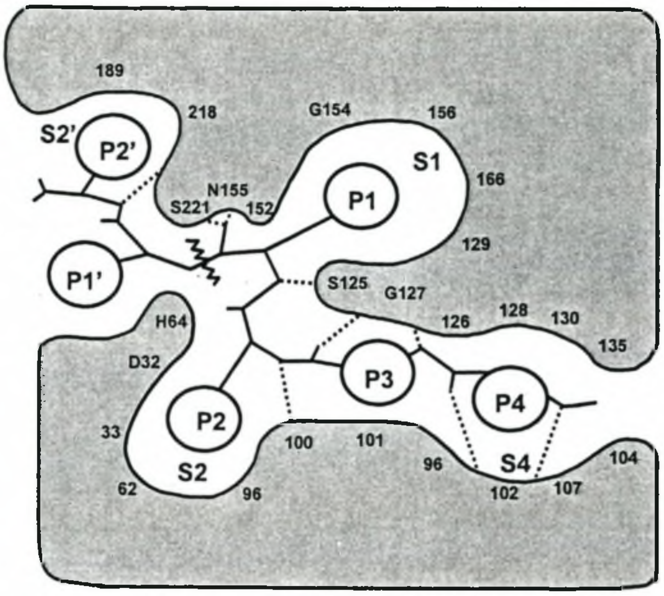
In all the mycosins there is also an insert of 10 - 17 residues (starting at about residue 290 in mycosin-5) in the substrate-binding region el, making prediction of substrate specificity or modelling of the substrate binding region impossible. The catalytic site of a protease is considered to be flanked by specificity subsites, each able to accommodate the side chain of a single amino acid residue (Barrett, 1994). The nomenclature of these subsites was proposed by Schechter and Berger in 1967 and the numbering extend from S1 tot S*n* (from catalytic site to N-terminus) and S1' to S*n*' (catalytic site to C-terminus). The residues accommodated by these sites are numbered accordingly, as indicated in Figure 2C.6.

Figure 2C.6. Substrate binding of a protease substrate into the specificity subsites of the enzyme. The arrow indicates the peptide bond cleaved (called the scissile bond).



The substrate specificity of many serine proteases is determined by the interaction between enzyme and substrate at the primary substrate binding site (S1) (Powers *et al.*, 1993, Lu *et al.*, 1997). In the subtilases, the binding region is able to accommodate at least six amino acid residues (P4-P2') of a polypeptide (see Figure 2C.7, Siezen and Leunissen, 1997), and the substrate specificity seems to be largely determined by not only the S1, but also the S4 binding sites (Gron *et al.*, 1992).

Figure 2C.7. Schematic presentation of the substrate binding to a subtilase. Nomenclature of peptide side chain residues and enzyme binding sites are according to Schechter and Berger (1967). Circles indicate peptide side chains (P4-P2') inside enzyme sites (S4-S2'). Numbering of amino acids is according to the protein sequence of subtilisin BPN'. The scissile bond is shown as a jagged line. The catalytic residues (D, H, and S) as well as the oxyanione hole residue (N) are indicated in positions 32, 64, 221 and 155 respecively. Adapted from Siezen and Leunissen, 1997.



Variations in the substrate specificity of the subtilases are due to variations in the residues situated in the substrate binding regions, specifically in the S1 and S4 sites. An example of this is the presence of an additional Asp residue in the bottom of the S1 pocket, found in all members of the kexin/furin family (eukaryotic pro-protein processing subtilisins) and the lantibiotic peptidases. This addition makes the S1 binding site very acidic and prone to the binding of and cleaving C-terminally to basic residues like Arg (Figure 2C.8, Rawlings and Barrett, 1994, Siezen *et al.*, 1994, Siezen and Leunissen, 1997). This region was found to be enlarged in the mycosin protein sequences (Figure2C.8). An alignment of the mycosin regions with the regions of the kexin and subtilisin families revealed a high number of aspartic acid residues present in the mycosin regions, similar to what is observed in the kexins. In addition to this these regions also contain a conserved cysteine residue which is another feature of kexin family (as well as the lantibiotic peptidases, Figure 2C.8). It is significant that the kexin and lantibiotic peptidase families of proteases are both highly substrate specific and are all involved in the activation of pro-proteins by cleavage at (pairs of) basic residues as the precursor protein is transported through the secretory pathway (Barr, 1991, Steiner *et al.*, 1992, Seidah and Chrétien, 1994, Siezen *et al.*, 1995). The members of the kexin family contains a number of other acidic residues in conserved positions which also form part of the substrate binding pockets on the three dimensional structure of the protein. These residues form a high density of negative charge on the surface of the substrate binding region, in particular the S1, S2 and S4 sites (Creemers *et al.*, 1993, Siezen *et al.*, 1994, Siezen and Leunissen, 1997). The acidic residues that are present in the members of the kexin family are found at positions 33, 61, 97, 104, 107, 129, 130, 131, 161, 166, 191 and 209 (subtilisin BPN' numbering) and are all situated in or near the substrate binding region (see Figure 2C.7). The protein sequences of the mycosins were also inspected for acidic residues found in the corresponding regions. Multiple acidic residues were found in the mycosins in most of these regions (acidic residues present in or near substrate binding pockets are highlighted in yellow in Figure 2C.5). The similarity between the amino acid sequences suggests that the mycosins may have a high specificity for cleaving a substrate pro-protein between paired basic residues and may thus also be examples of bacterial pro-protein-processing subtilases. This could explain why no protease activity was found with the common substrates used in the activity assays.

It is interesting to observe that when the S1 pocket regions of the mycosins are arranged according to their duplication order (Gey van Pittius *et al.*, 2001, see Chapter 3), it is clear that the substrate binding regions of these proteases seem to be evolving, as the aspartic acid residue numbers in this region are increasing (Figure 2C.8). Thus, the sequence of the most ancient mycosin (mycosin-4) only contains one residue, while the most recent duplicate (mycosin-5) contains five. Mycosin-1, -3 and -2 contain three, four and five residues respectively. This may indicate that each mycosin has evolved to cleave a specific substrate that is not cleaved by the other mycosins. Supporting evidence for this comes from the fact that the eI and eIII substrate binding regions of the mycosins (Figure 2C.2) are reasonably unconserved, indicating differences in the substrate specificity of the different mycosins.

Figure 2C.8. Structural basis for the specificity of some members of the subtilisin family for cleavage at paired basic residues. Part of the S1 specificity subsite (residues 161-172 of thermitase), was aligned with other bacterial subtilisins and the eukaryotic proprotein processing subtilisins (the kexin family), showing clearly that the kexin-like molecules contains aspartic acid residues in this part of the sequence that is not found in the bacterial subtilisins (Barrett and Rawlings, 1991). The corresponding enlarged region of the mycosins is also aligned with the other regions and indicates a high level of aspartic acid residues present in these sequences, as well as a conserved cysteine residue also only found in the kexin family. Mycosins were arranged in duplication order (Gey van Pittius *et al.*, 2001), indicating evolution of region by addition of acidic residues.

Bacterial subtilisins

Thermitase	GNAGN--TAPNY---PA
Subtilisin BPN'	GNEGTSGSSSTV-GYPG
Subtilisin Carlsberg	GNSGSSGNTNTI-GYPA
Subtilisin DY	GNSGSSGSQNTI-GYPA
Subtilisin amylosacchariticus	GNEGSSGSSSTV-GYPA

Eukaryotic subtilisins (kexin family)

Yeast kexin	GNGGTRGDNCNYDGYTN
Human PC2	GDGGSY-DDCNCDGYAS
Mouse PC1	GNGGRQGDNCDCDGYTD
Mouse PC2	GDGGSY-DDCNCDGYAS
Human furin	GNGGREHDSCNCDGYTN
Rat furin	GNGGREHDSCNCDGYTN

Mycosins

mycosin-1	GNTGG------DCVQNPAPDPSTPGDPRGWNNVQTVVTPA
mycosin-2	GNDGE-----AGCGNNPMYDPLDPSDPRDWHQVTVVSSPS
mycosin-3	GNTGASGSVSASCDSNPLTDLSRPDDPRNWAGVTSVSIPS
mycosin-4	GNTGGA----AQC------PPQAPGVTRD--SVTVAVSPA
mycosin-5	G-DGSKK----DCKQNPIFDPLQPDDPRAWNAVTTVVTPS

Mycosins (arranged according to duplication order)

mycosin-4	GNTGGA----AQC------PPQAPGVTRD--SVTVAVSPA
mycosin-1	GNTGG------DCVQNPAPDPSTPGDPRGWNNVQTVVTPA
mycosin-3	GNTGASGSVSASCDSNPLTDLSRPDDPRNWAGVTSVSIPS
mycosin-2	GNDGE-----AGCGNNPMYDPLDPSDPRDWHQVTVVSSPS
mycosin-5	G-DGSKK----DCKQNPIFDPLQPDDPRAWNAVTTVVTPS

In addition to the pro-protein processing features identified in the substrate binding sites of the mycosins, there is another feature of these proteases which indicates a function similar to that of the lantibiotic proprotein processing peptidases. The enzyme domains are attached to a membrane anchor (very typical with about 20 hydrophobic residues followed by 3 - 4 arginine residues), via a short linker of 20 - 30 residues, which is very proline-rich as mentioned earlier. As the *M. tuberculosis* cell wall is about 30 nm thick, the linker is much too short to position the enzyme outside the cell wall, so that the protein must function inside the cell wall. This feature has only been detected in one other subtilase, the NisP subtilisin-like serine protease that specifically cleaves and activates the nisin-lantibiotic (Van der Meer *et al*., 1993, Kok and De Vos, 1994, Siezen *et al*., 1995) and belongs to the family of lantibiotic peptidases. The substrate for this enzyme is pre-nisin, the precursor of the bacteriocin nisin which comes from inside the cell (*Lactococcus lactis*) and is activated by cleavage of the pro-region by the membrane-bound protease as it translocates across the cell membrane. This protease is encoded by a gene situated in a cluster of genes involved in the specific biosynthesis and active transport of the nisin lantibiotic. Subsequent to the protease activity assay analyses of the present study, several genes encoding active transport-associated proteins were identified to be situated adjacent to the mycosin genes in a cluster formation in the genome of *M. tuberculosis* (Tekaia *et al*., 1999, Gey van Pittius *et al*., 2001), similar to what is observed in the lantibiotic biosynthesis clusters of the lactococci.

It is thus possible that the mycosins are also lantibiotic/kexin-type proteases displaying a high degree of substrate specificity (especially for basic amino acid residues) and which are involved in the maturation of a pro-protein secreted by *M. tuberculosis*. Owing to the extreme specificity of these types of proteases, this would explain why no protease activity was obtained during the protease activity analyses using casein and other common substrates.

2C.3.5. Conclusions

In conclusion, although there are a number of possible reasons for not observing protease activity for the recombinant mycosins (which could include incorrect folding of the fusion protein, no or incorrect processing of the prepromycosins, substrate specificity, pH or temperature, cofactors etc.), the most likely explanation is provided by the fact that the mycosins reveal characteristics shared only by the lantibiotic peptidases and the proprotein convertases, indicating that substrate specificity may be extremely crucial. Almost all of the known subtilisins that have been examined previously were identified from an observed unknown protease activity, thereby immediately providing a usable substrate. The fact that the mycosins were identified from their gene sequences, makes the postulation of a possible substrate, as well as optimal conditions for activity extremely difficult. It is clear from this study that the substrate of the mycosins has to be identified before any further activity analyses could be done.

CHAPTER THREE

THE ESAT-6 GENE CLUSTERS

"the mere existence of such arrangements shows that they must be beneficial, conferring an evolutionary advantage on individuals and populations which exhibit them."

Complex loci in microorganisms – M. Demerec and P. Hartman (1959)

NOTE: The results presented in the following chapter were published as: **"The ESAT-6 gene cluster of *Mycobacterium tuberculosis* and other high G+C gram-positive bacteria."** Gey van Pittius, N.C., Gamieldien, J., Hide, W., Brown, G.D., Siezen, R.J., and Beyers, A.D., *Genome Biology* 2001 2(10): research0044.1-0044.18.

(The style of the text and numbering of sections has been altered to conform to the style of this dissertation. Literature cited in the text takes the form of author name and year of publication as opposed to the number format specified by Genome Biology. All cited literature is compiled into a single list at the end of the dissertation for ease of reference)

3.1. Introduction

Mycobacterium tuberculosis remains a serious threat to human health and in spite of significant investment into the research of this organism, the mechanisms of its pathogenicity are still not clearly understood. One of the strategies used to decipher these mechanisms is the comparison of the presence and absence of genes in different species (e.g. virulent and avirulent) and extrapolation of these differences to variation in phenotype. The genomes of *M. tuberculosis* H37Rv, *M. tuberculosis* H37Ra, *M. bovis*, and the attenuated *M. bovis* BCG have been compared in different combinations using a variety of methods (subtractive genomic hybridization – Mahairas *et al.*, 1996, BAC restriction profile analysis – Philipp *et al.*, 1996, Brosch *et al.*, 1998, Brosch *et al.*, 1999, Brosch *et al.*, 2000c, BAC arrays – Gordon *et al.*, 1999a, DNA microarrays – Behr *et al.*, 1999, and Southern blotting – Zumarraga *et al.*, 1999), resulting in the identification of a number of regions of difference (RD) between the various organisms.

One of these regions, designated the RD1 (region of difference 1) deletion region (Mahairas *et al.*, 1996), is a 9505 bp region absent in all *M. bovis* BCG strains. RD1 is commonly thought to be the primary deletion that occurred during the serial passage of *M. bovis* by Calmette and Guérin between 1908 and 1921, and is thus thought to possibly be responsible for the primary attenuation of *M. bovis* to *M. bovis* BCG (Behr *et al.*, 1999, Brosch *et al.*, 2000a). Consequently, the genes contained in this region have also been the subject of a number of studies focusing on diagnosis of *M. tuberculosis* infection, the search for efficient vaccine candidates and virulence (Ahmad *et al.*, 1999, Arend *et al.*, 2000a, Brandt *et al.*, 2000, Wards *et al.*, 2000). This region encompasses the genes Rv3871 to Rv3879c (annotation according to Cole *et al.*, 1998), which include the 6 kDa early-secreted antigenic target ESAT-6 (*esx or esat-6*) and L45 homologous protein CFP-10 (*lhp*) genes (Andersen *et al.*, 1995, Berthet *et al.*, 1998). The *esat-6* and *lhp* genes are situated directly adjacent to each other and encode for potent T-cell antigens that are secreted but lack detectable secretion signals (Sørensen *et al.*, 1995, Van Pinxteren *et al.*, 2000).

During the genome sequencing of *M. tuberculosis* H37Rv, Cole *et al.* (1998) identified at least eleven additional genes encoding small proteins of approximately 100 amino acids that share sequence similarities with *esat-6*, and grouped them into the *esat-6* gene family. In addition, they found several small genes that share similarity with *lhp* that are also situated directly adjacent to the *esat-6* family member genes. Sequence analyses indicated that the *lhp* family members belong to and extend this *esat-6* gene family. It was also found that the *lhp* gene is co-transcribed and thus forms part of an operon with *esat-6* (Berthet *et al.*, 1998).

The genes encoding the originally annotated CFP-10 and ESAT-6 proteins within the RD1 deletion region lie in a cluster of 12 other genes (encompassing the deletion region), which seems to have been duplicated five times in the genome of *M. tuberculosis*. The duplicated gene clusters have been previously described as the ESAT-6 loci in an analysis of the proteome of *M. tuberculosis* (Tekaia *et al.*, 1999). An examination of the sets of genes in the clusters reveals that each of the clusters also contains (in addition to a copy of ESAT-6 and CFP-10), genes encoding putative ABC transporters (integral inner-membrane proteins), ATP-binding proteins, subtilisin-like membrane-anchored cell wall-associated serine proteases (the mycosins – Brown *et al.*, 2000), and other N-terminal membrane-associated proteins (Tekaia *et al.*, 1999).

We have implemented a sequence comparison approach to establish the relationship between the multiple copies of the ESAT-6 gene cluster. Our results demonstrate that the ESAT-6 gene cluster is of ancient origin, is present in and restricted to the genomes of other members of the high G+C gram-positive bacteria such as *Corynebacterium diphtheriae* and *Streptomyces coelicolor* and is duplicated multiple times in *Mycobacterium tuberculosis* and other mycobacteria. We discuss the conservation of this gene cluster in the context of possible functional importance and diagnoses of mycobacterial infection.

3.2. Materials and Methods

3.2.1. Genome sequence data and analyses

Annotations and descriptions of individual genes as well as gene and protein sequences of individual organisms were obtained from the publicly available finished and unfinished genome sequence databases (Table 3.1). All gene and protein sequences were subjected to analyses with the following programs to confirm annotation and to look for additional information: SignalP V2.0.b2 (Nielsen *et al.*, 1997, http://www.cbs.dtu.dk/services/SignalP-2.0/#submission), ClustalW WWW server at the European Bioinformatics Institute (Thompson *et al.*, 1994, http://www2.ebi.ac.uk/clustalw/), TMHMM v0.1 transmembrane prediction server (Sonnhammer *et al.*, 1998, http://www.cbs.dtu.dk/services/TMHMM-1.0/), MOTIF (http://www.motif.genome.ad.jp/) and BLASTP (Altschul *et al.*, 1990, http://www.ncbi.nlm.nih.gov/blast/blast.cgi?Jform=0). No data, progress report, or BLAST search function is available for the genome sequencing of *Mycobacterium bovis* BCG Pasteur 1173P2 produced by the Pasteur Institute, but information concerning genome deletions was obtained from published data (Mahairas *et al.*, 1996, Philipp *et al.*, 1996, Brosch *et al.*, 1998, Behr *et al.*, 1999, Gordon *et al.*, 1999a, Brosch *et al.*, 2000c) as well as the Pasteur Institute website (http://www.pasteur.fr/recherche/unites/Lgmb/Deletion.html).

3.2.2. Analyses of similar gene clusters

BLAST similarity searches (Altschul *et al.*, 1990), using the BLAST 2.0 program with tblastn and the BLOSUM-62 weight matrix, were utilized to identify stretches of DNA containing putative ORF's homologous to the genes found in the *M. tuberculosis* ESAT-6 gene cluster regions from finished and unfinished genome sequences available at the NCBI website (http://www.ncbi.nlm.nih.gov/Microb_blast/unfinishedgenome.html). A total of 98 finished and unfinished genome sequences (35 from Gram positive species) were used in the analysis, as summarized in Table 3.1. Where applicable, BLAST servers in database search services of individual sequencing centers were also used for protein identification. The Sanger Centre and The Institute for Genomic Research (TIGR) use the program WU-BLAST version 2.0 (http://blast.wustl.edu/), while the University of Minnesota uses BLASTN with supplied defaults (http://www.cbc.umn.edu/cgi-bin/blasts/AGAC.restrict/blastn.cgi). Sequences were only admitted to analyses when found to be part of one of the five gene clusters. In other words, no single homologous genes in the mycobacteria or other organisms (for example *Bacillus subtilis*) that did not form part of a similar gene cluster were considered for analyses, to exclude any potential unassociated similarity that could lead to false positives.

Table 3.1. Publicly available finished and unfinished genome sequence databases used in this study

Acidithiobacillus ferrooxidans
Actinobacillus actinomycetemcomitans
Aquifex aeolicus
Bacillus anthracis
Bacillus halodurans
Bacillus subtilis
Bacillus stearothermophilus
Bordetella bronchiseptica
Bordetella parapertussis
Bordetella pertussis
Borrelia burgdorferi
Brucella melitensis biovar Suis
Buchnera sp. APS
Burkholderia mallei
Burkholderia pseudomallei
Campylobacter jejuni **NCTC 11168**
Carboxydothermus hydrogenoformans
Caulobacter crescentus
Chlamydia muridarum
Chlamydia pneumoniae
Chlamydia trachomatis **D/UW-3/CX**
Chlamydophila pneumoniae **AR39**
Chlamydophila psittaci
Chlorobium tepidum
Clostridium acetobutylicum
Clostridium difficile
Corynebacterium diphtheriae
Coxiella burnetii
Dehalococcoides ethenogenes
Desulfovibrio vulgaris
Deinococcus radiodurans
Escherichia coli **K-12 MG1655**
Escherichia coli **O157:H7**
Escherichia coli **O157:H7 EDL933**
Enterococcus faecalis
Geobacter sulfurreducens
Haemophilus ducreyi 35000HP
Haemophilus influenzae **Rd**
Helicobacter pylori **26695**
Helicobacter pylori **J99**
Klebsiella pneumoniae
Lactococcus lactis subsp. lactis
Legionella pneumophila
Listeria monocytogenes
Mesorhizobium loti
Methylococcus capsulatus
Mycobacterium avium
Mycobacterium avium subsp. paratuberculosis
Mycobacterium bovis
Mycobacterium leprae
Mycobacterium smegmatis
Mycobacterium tuberculosis 210
Mycobacterium tuberculosis **CDC1551**
Mycobacterium tuberculosis **H37Rv**
Mycoplasma genitalium **G37**
Mycoplasma pneumoniae **M129**
Neisseria gonorrhoeae
Neisseria meningitidis **MC58**
Neisseria meningitidis **Z2491**
Pasteurella multocida **PM70**
Porphyromonas gingivalis W83
Pseudomonas aeruginosa
Pseudomonas putida KT2440
Pseudomonas putida PRS1
Pseudomonas syringae pv. tomato
Rickettsia prowazekii
Rhodobacter sphaeroides
Salmonella dublin
Salmonella enteritidis
Salmonella paratyphi
Salmonella typhi
Salmonella typhimurium LT2
Shewanella putrefaciens
Sinorhizobium meliloti
Staphylococcus aureus COL
Staphylococcus aureus MRSA
Staphylococcus aureus MSSA
Staphylococcus aureus **Mu50**
Staphylococcus aureus **N315**
Staphylococcus aureus NCTC 8325
Staphylococcus epidermidis
Streptococcus equi
Streptococcus gordonii
Streptococcus mutans
Streptococcus pneumoniae
Streptococcus pyogenes
Streptococcus pyogenes Manfredo
Streptomyces coelicolor A3(2)
Synechocystis **PCC6803**
Thermotoga maritima
Treponema denticola
Treponema pallidum
Ureaplasma urealyticum
Vibrio cholerae
Wolbachia
Xylella fastidiosa
Yersinia enterocolitica
Yersinia pestis

Finished genome sequences are indicated in **bold**, Gram-positive species are underlined

Contig sequences corresponding to the gene clusters were obtained from their respective genome databases and used in further analyses. The Genetics Computer Group (Wisconsin Package Version 10.0, Genetics Computer Group (GCG), Madison, Wisconsin) program FRAMESEARCH was used to obtain whole sequence ORF's from the contigs. These ORF's were translated to protein sequences with the program TRANSLATE (also from GCG). All multiple sequence alignments and phylogenetic analyses were conducted on the protein level with these translated protein sequences.

3.2.3. Multiple sequence alignments

Multiple sequence alignments were performed on separate gene families belonging to the different clusters using ClustalW 1.5 (Thompson *et al.*, 1994) with the default parameters. The alignments were manually checked for errors and refined where appropriate. Multiple sequence alignments were also manually edited in some analyses during which unaligned regions (inserts) were removed (resulting in so-called edited alignments).

3.2.4. Phylogenetic trees

Bootstrapping resampling of the data sets were performed on the edited alignments, which generated 100 randomly chosen subsets of the multiple sequence alignment. Pairwise distances were determined with PROTDIST using the Dayhoff PAM matrix and neighbor-joining phylogenetic trees were calculated using NEIGHBOR (PHYLIP 3.5, Felsenstein, 1989). In the case of each family of proteins, the *C. diphtheriae* sequence was firstly used as the outgroup after which the *S. coelicolor* sequence was used. Further phylogenetic analyses were performed using the programs FITCH and KITSCH with and without the outgroups respectively. A majority rule and strict consensus tree of all bootstrapped sequences were obtained using CONSENSE. The same analyses as described above were performed on a combined protein consisting of the edited aligned sequences of all six conserved proteins in these gene clusters as well as a combined protein constructed from the edited aligned sequences of all available ESAT-6 and CFP-10 family members. Finally, to confirm the results obtained on the singular protein level, an analysis was performed with whole, unedited aligned sequences of the six most conserved proteins, using the program Paup 4.0b4a (Swofford, 1998), during which negative branches were collapsed and 1000 subsets were generated for Bootstrapping resampling of the data. The consensus trees of all of the above were drawn using the program Treeview 1.5 (Page, 1996).

3.3. Results

3.3.1. Individual gene families and genomic organization in M. tuberculosis

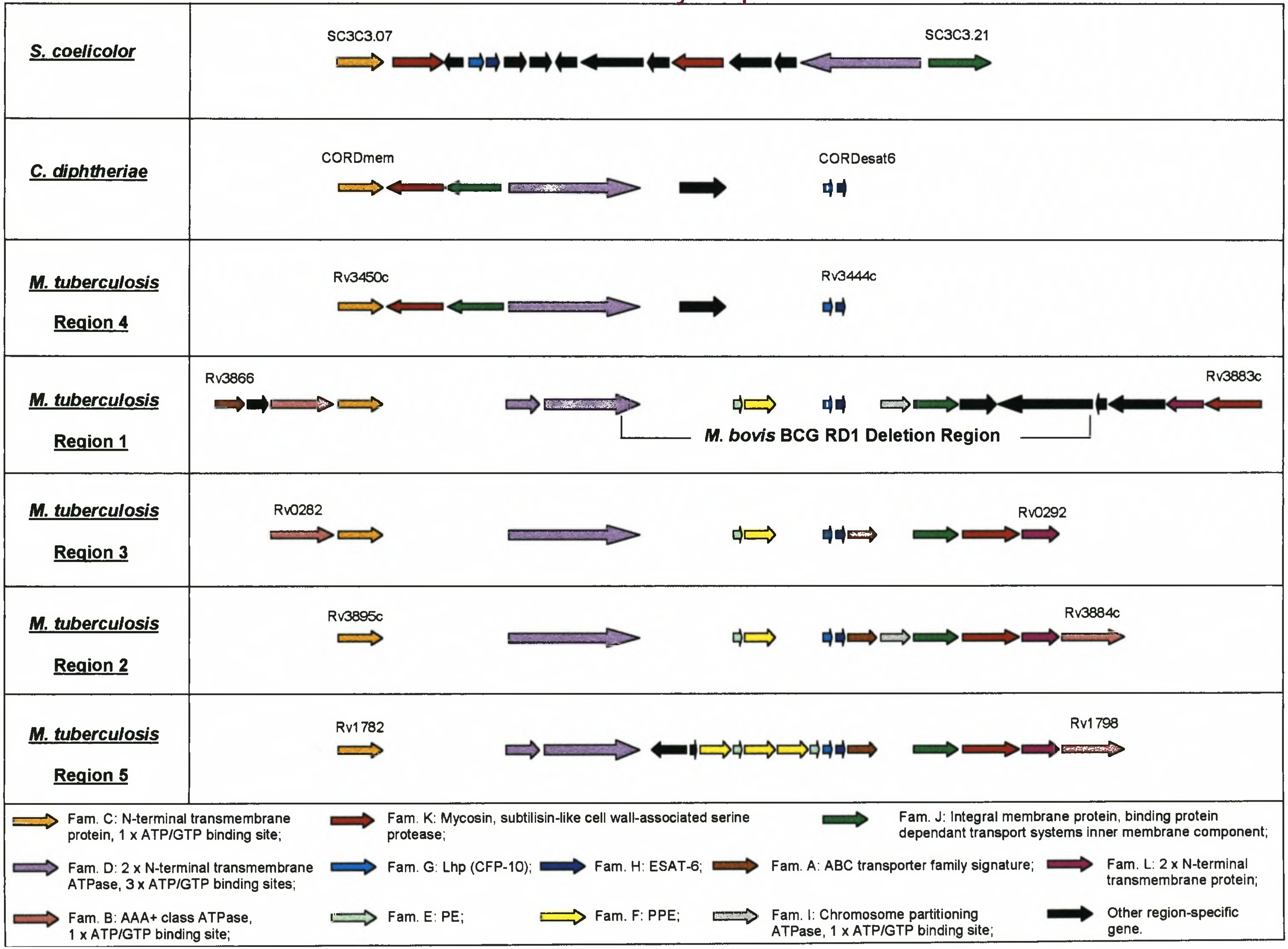
The five ESAT-6 gene cluster regions present in *Mycobacterium tuberculosis* H37Rv were named region 1 (Rv3866-Rv3883c), 2 (Rv3884c-Rv3895c), 3 (Rv0282-Rv0292), 4 (Rv3444c-Rv3450c) and 5 (Rv1782-Rv1798) consistent with the arbitrary numbering system used previously to classify the five mycosin (subtilisin-like serine protease) genes identified from these regions (Brown *et al.*, 2000). Orthologues of the ESAT-6 gene clusters of *M. tuberculosis* H37Rv could be identified in the genomes of eight other strains and species belonging to the genus Mycobacterium, as well as two species belonging to other genera (Table 3.2). Up to twelve different genes representing different gene families were identified in the five gene cluster regions and were designated family A to L according to their position in region 1 (Table 3.3).

Figure 3.1 shows a schematic representation of the genomic organization of the respective gene families present in each of the five ESAT-6 gene cluster regions of *M. tuberculosis*. Annotations and descriptions of single genes in these regions can be found at (http://genolist.pasteur.fr/TubercuList/). Region 1 and 2 are situated directly adjacent to each other on the genome and are transcribed in opposite directions. The large gene belonging to family D (encoding the ATPase protein) has been disrupted by an insertion in both regions 1 and 5 (Figure 3.1). This insertion has caused an in-frame stop codon, giving rise to two smaller genes (containing all the motifs of the larger homologue) located directly adjacent to each other. The gene positions of members of family C, D, G, and H are maintained throughout the five regions (see Figure 3.1), while most of the families that are not present in region 4, seem to be more flexible with regard to their position within the gene cluster regions (family A, B, I, and L). There are also some genes present within the ESAT-6 gene cluster regions that do not have any homologues in the other clusters, suggesting subsequent insertions or deletions from the ancestral region (indicated by black arrows in Figure 3.1, see also Table 3.3).

Table 3.2. Bacterial Genome Sequencing Projects of Species and Strains Containing ESAT-6 Gene Clusters

	Organism	Strain	Status	Last Access Date	Last Update	Sequencing Centre(s)	Website(s)	Reference
1	Mycobacterium tuberculosis	H37Rv	Completed	5-Mar-2001	11-Jun-1998	Sanger Centre/Pasteur Institute	http://www.sanger.ac.uk/Projects/M_tuberculosis/ and http://genolist.pasteur.fr/TubercuList/	Cole *et al.*, 1998
2	Mycobacterium tuberculosis	CDC1551 (Oshkosh strain or CSU#93)	Completed	5-Mar-2001	28-Jan-1999	TIGR	http://www.tigr.org/tigr-scripts/CMR2/GenomePage3.spl?database=gmt	Fleischmann *et al.*, manuscript in preparation
3	*Mycobacterium tuberculosis*	210	Partial sequencing project completed, no additional sequencing anticipated	21-May-2001	4-May-2001	TIGR	http://www.tigr.org/cgi-bin/BlastSearch/blast.cgi?	
4	*Mycobacterium bovis*	AF2122/97(spoligotype 9)	Shotgun in progress	5-Mar-2001	29-Aug-2000	Sanger Centre/Pasteur Institute	http://www.sanger.ac.uk/Projects/M_bovis/	
5	*Mycobacterium bovis BCG*	Pasteur 1173P2	Unfinished	-	-	Pasteur Institute	http://www.pasteur.fr/recherche/unites/Lgmb/mycogenomics.html	
6	*Mycobacterium leprae*	TN	Completed	7-Mar-2001	21-Feb-2001	Sanger Centre/Pasteur Institute	http://www.sanger.ac.uk/Projects/M_leprae/ and http://genolist.pasteur.fr/Leproma	Cole *et al.*, 2001
7	*Mycobacterium avium*	104	Gap closure finished	6-Mar-2001	22-Feb-2001	TIGR	http://www.tigr.org/cgi-bin/BlastSearch/blast.cgi?organism=m_avium	
8	*Mycobacterium paratuberculosis*	K10	Unfinished (6.9 x coverage)	6-Mar-2001	25-Feb-2001	University of Minnesota	http://www.cbc.umn.edu/ResearchProjects/AGAC/Mptb/Mptbhome.html	
9	*Mycobacterium smegmatis*	MC^2 155	Shotgun completed, assembly	6-Mar-2001	22-Feb-2001	TIGR	http://www.tigr.org/cgi-bin/BlastSearch/blast.cgi?organism=m_smegmatis	
10	*Corynebacterium diphteriae*	NCTC13129	Finishing/gap closure	5-Mar-2001	26-Feb-2001	Sanger Centre	http://www.sanger.ac.uk/Projects/C_diphtheriae/	
11	*Streptomyces coelicolor*	A3(2)	Cosmid sequencing	5-Mar-2001	1-Mar-2001	Sanger Centre	http://www.sanger.ac.uk/Projects/S_coelicolor/	

Figure 3.1. Schematic representation of the genomic organization of the genes present in the five ESAT-6 gene cluster regions of *Mycobacterium tuberculosis* H37Rv as well as the regions in *C. diphtheriae* and *S. coelicolor*. ORF's are represented as blocked arrows showing the direction of transcription, with the different colors reflecting the specific gene family and the length of the arrow reflecting the relative lengths of the genes. Annotations of *M. tuberculosis* H37Rv genes are according to Cole *et al.* (1998). Black arrows indicate unconserved genes present in these regions. Gaps between genes do not represent physical gaps between genes on the genome, but have been inserted to aid in indicating conservation among gene positions. Gene families were named arbitrarily according to their position in *M. tuberculosis* H37Rv region 1. The regions were named after the numbering system of Brown *et al.* (2000) used arbitrarily for the five mycosin (subtilisin-like serine protease) genes identified from these regions (family K). *M. tuberculosis* regions are shown in order of suggested duplication events (see phylogenetic results) and not by numbering. The results of the analyses of the primary features of these genes and their corresponding proteins are included in a short summary at the bottom of the figure (see also Table 3.3).


S. coelicolor
SC3C3.07
SC3C3.21
C. diphtheriae
CORDmem
CORDesat6
M. tuberculosis
Region 4
Rv3450c
Rv3444c
M. tuberculosis
Region 1
Rv3866
Rv3883c
M. bovis BCG RD1 Deletion Region
M. tuberculosis
Region 3
Rv0282
Rv0292
M. tuberculosis
Region 2
Rv3895c
Rv3884c
M. tuberculosis
Region 5
Rv1782
Rv1798
Fam. C: N-terminal transmembrane protein, 1 x ATP/GTP binding site;
Fam. K: Mycosin, subtilisin-like cell wall-associated serine protease;
Fam. J: Integral membrane protein, binding protein dependant transport systems inner membrane component;
Fam. D: 2 x N-terminal transmembrane ATPase, 3 x ATP/GTP binding sites;
Fam. G: Lhp (CFP-10);
Fam. H: ESAT-6;
Fam. A: ABC transporter family signature;
Fam. L: 2 x N-terminal transmembrane protein;
Fam. B: AAA+ class ATPase, 1 x ATP/GTP binding site;
Fam. E: PE;
Fam. F: PPE;
Fam. I: Chromosome partitioning ATPase, 1 x ATP/GTP binding site;
Other region-specific gene.

The ESAT-6/CFP-10 operon is not only found in the ESAT-6 gene cluster regions, but distributes as 6 additional copies of the gene pair in the genome of *M. tuberculosis*. Figure 3.2 gives a schematic representation of the positions of the six additional gene pairs. In four of the six cases, the ESAT-6/CFP-10 operon is flanked by PPE and PE genes, indicating possible linked-duplication between the ESAT-6/CFP-10 operon and the PE/PPE gene pair.

3.3.2. ESAT-6 gene cluster identification in other mycobacteria

Table 3.3 presents the results of the similarity searches and all available data for the twelve identified gene families present in the different regions. All the mycobacteria currently being sequenced contain multiple copies of these regions in their genomes. As these different copies are also found in the same respective genomic locations (corresponding flanking genes) in all the mycobacteria, it indicates that the duplication events took place prior to the divergence of the different species.

3.3.2.1. *Mycobacterium tuberculosis* CDC1551, *Mycobacterium tuberculosis* 210 and *Mycobacterium bovis*

The genomes of the *M. tuberculosis* CDC1551 and 210 clinical strains as well as the genome of *M. bovis* contain all five of the ESAT-6 gene cluster regions present in the genome of *M. tuberculosis* H37Rv (sharing between 99 and 100% similarity to *M. tuberculosis* H37Rv at protein level). However, it is interesting to note that two of the genes present in region 2 in CDC1551 (MT4000 and MT4001) contain frameshifts in their sequences, indicating that they and the rest of the region may no longer be functional in CDC1551. Part of region 2 (a 2405 bp fragment containing Rv3887c, Rv3888c and Rv3889c) is also deleted in only certain strains of *M. bovis*, including the strain AF2122/97 that is currently being sequenced (Rauzier *et al.*, 1999). An in-frame stop codon found in Rv1792 (family G) is also present in the orthologues in CDC1551 (MT1841) and strain 210 (MTB196G), indicating that the mutation may have taken place before divergence of the three strains. Two of the H37Rv as well as the strain 210 Family D genes (in region 1 and 5) have obtained in-frame stop codons resulting in two genes lying adjacent to each other, whereas the Family D Rv1783 and Rv1784 orthologues in CDC1551 are still one intact gene (MT1833). The orthologues of this gene in *M. bovis* (MB771.1D), *M. leprae* (ML1543) *M. avium* (MA221D), and *M. paratuberculosis* (MP1783) are also intact, implying that the mutation in the H37Rv and strain 210 orthologues must have occurred after divergence of the three *M. tuberculosis* strains.

Figure 3.2. Schematic representation of the six additional ESAT-6/CFP-10 operon duplications and the regions that surround them in the genome of *M. tuberculosis* H37Rv. ORF's are represented by blocked arrows indicating direction of transcription, with the different colors reflecting the specific gene family and the length of the arrow reflecting the relative lengths of the genes as in Figure 3.1. The ESAT-6/CFP-10 genes deleted in *M. bovis* RD 07 and RD 09 deletion regions (Behr *et al.*, 1999) are indicated.

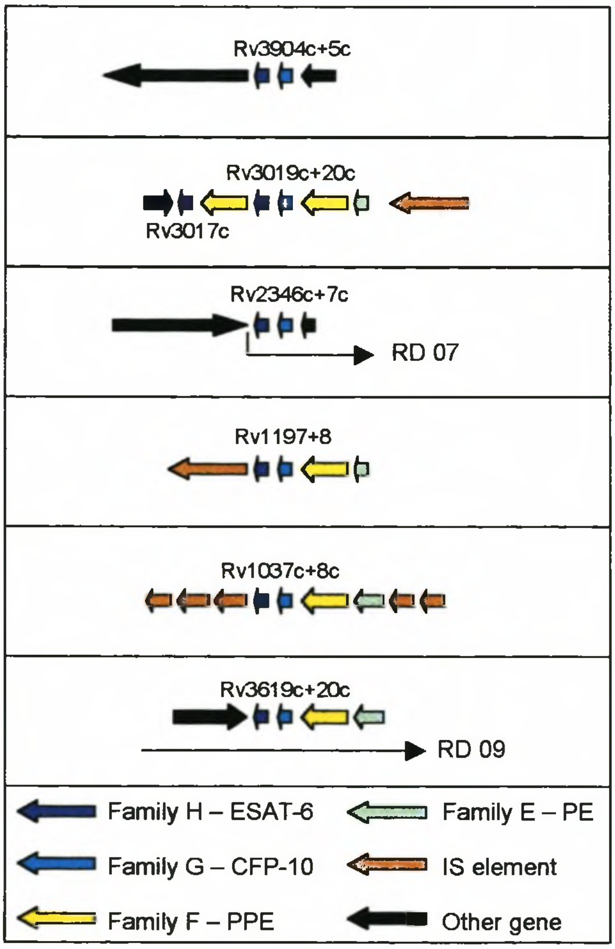


Table 3.3. Presence of Genes in Gene Clusters of All Available Finished and Unfinished Genome Sequences

Gene Family	Description	Protein size (In _M.tb_)	ESAT-6 Cluster Region	Presence and names of genes in each species				
				M. tuberculosis H37Rv	M. tuberculosis CDC1551 (CSU#93)	_M. tuberculosis_* 210	_M. bovis_* AF2122/97(spoligotype 9)	_M. bovis_* BCG Pasteur 1173P2
A	ABC transporter family signature, 19-27% homology	283	1	Rv3866	MT3980	ND	MB851A	No Sequence Data
		276	2	Rv3889c	MT4004	MTB12A	MB727.3A (Partly deleted #)	No Sequence Data
		295	3	Rv0289	MT0302	MTB203A	MB548A	No Sequence Data
		-	4	No Duplication	No Duplication	No Duplication	No Duplication	No Duplication
		300	5	Rv1794	MT1843	MTB196A	MB557A	No Sequence Data
B	AAA+ class ATPases, CBXX/CFQX family, SpoVK, 1 x ATP/GTP binding site, 29-39% homology	573	1	Rv3868	MT3981	MTB44B	MB851B	No Sequence Data
		619	2	Rv3884c	MT3999	MTB12B	MB727.1B	No Sequence Data
		631	3	Rv0282	MT0295	MTB23B	MB672B	No Sequence Data
		-	4	No Duplication	No Duplication	No Duplication	No Duplication	No Duplication
		610	5	Rv1798	MT1847	MTB196B	MB542B	No Sequence Data
C	N-terminal transmembrane protein, possible ATP/GTP binding motif, 31-41% homology	480	1	Rv3869	MT3982	MTB44C	MB851C	No Sequence Data
		495	2	Rv3895c	MT4011	MTB136C	MB780.1C	No Sequence Data
		538	3	Rv0283	MT0296	MTB23C	MB672C	No Sequence Data
		470	4	Rv3450c	MT3558	MTB45C	MB493.1C	No Sequence Data
		506	5	Rv1782	MT1832	MTB46C	MB771.1C	No Sequence Data
D	DNA segregation ATPase, ftsK chromosome partitioning protein, SpoIIIE, yukA, 3 x ATP/GTP binding sites, 2 x N-terminal transmembrane protein, 28-39% homology	747+591	1	Rv3870+71	MT3983+85	MTB44Da+Db	MB851D	MB851D (Partly deleted)
		1396	2	Rv3894c	MT4010	MTB3D	MB780.1D	No Sequence Data
		1330	3	Rv0284	MT0297	MTB23D	MB672D	No Sequence Data
		1236	4	Rv3447c	MT3553	MTB45D	MB585.1D	No Sequence Data
		435+932	5	Rv1783+84	MT1833	MTB46Da+Db	MB771.1D	No Sequence Data
E	PE, 18-90% homology	99	1	Rv3872	MT3986	MTB44E	MB851E	Deleted
		77	2	Rv3893c	MT4008	MTB3E	MB780.1E	No Sequence Data
		102	3	Rv0285	MT0298	MTB23E	MB389E	No Sequence Data
		-	4	No Duplication	No Duplication	No Duplication	No Duplication	No Duplication
		99 & 99	5	Rv1788 & 91	MT1837 & 40	MTB196Ea & Eb	MB771.0E & MB557E	No Sequence Data
F	PPE, 19-88% homology	368	1	Rv3873	MT3987	MTB44F	MB851F	Deleted
		399	2	Rv3892c	MT4007	MTB3F	MB780.1F	No Sequence Data
		513	3	Rv0286	MT0299	MTB472F	MB528F	No Sequence Data
		-	4	No Duplication	No Duplication	No Duplication	No Duplication	No Duplication
		365, 393 & 350	5	Rv1787 & 89 & 90	MT1836 & 38 & 39	MTB196Fa & Fb & Fc	MB771.0Fa & Fb & MB557F	No Sequence Data
G	lhp or CFP-10, also MTSA-10, grouped into ESAT-6 family, potent secreted T-cell antigens, 9-32% homology	100	1	Rv3874	MT3988	MTB44G	MB851G	Deleted
		107	2	Rv3891c	MT4006	MTB12G	MB727.3G	No Sequence Data
		97	3	Rv0287	MT0300	MTB472G	MB548G	No Sequence Data
		125	4	Rv3445c	MT3550	MTB45G	MB585.0G	No Sequence Data
		98	5	Rv1792 (Stop)	MT1841 (Stop)	MTB196G (Stop)	MB557G	No Sequence Data
H	ESAT-6 family, cfp7, L45 or l-esat, also Mtb9.9 family, potent secreted T-cell antigens, 15-27% homology	95	1	Rv3875	MT3989	MTB44H	MB851H[b]	Deleted
		95	2	Rv3890c	MT4005	MTB12H	MB727.3H	No Sequence Data
		96	3	Rv0288	MT0301	MTB203H	MB548H	No Sequence Data
		100	4	Rv3444c	MT3549	MTB45H	MB585.0H	No Sequence Data
		94	5	Rv1793	MT1842	MTB196H	MB557H	No Sequence Data
I	ATPases involved in chromosome partitioning, 1 x ATP/GTP binding motif, 33% homology	666	1	Rv3876	MT3990	MTB60I	MB477I	Deleted
		341	2	Rv3888c	MT4003	MTB12I	Deleted #	No Sequence Data
		-	3	No Duplication	No Duplication	No Duplication	No Duplication	No Duplication
		-	4	No Duplication	No Duplication	No Duplication	No Duplication	No Duplication
		-	5	No Duplication	No Duplication	No Duplication	No Duplication	No Duplication
J	Integral inner membrane protein, binding protein dependent transport systems inner membrane component signature, putative transporter protein, 19-27% homology)	511	1	Rv3877	MT3991	MTB369J	MB477J	Deleted
		509	2	Rv3887c	MT4002	MTB12J	MB727.3J (Partly deleted #)	No Sequence Data
		472	3	Rv0290	MT0303	MTB203J	MB548J	No Sequence Data
		467	4	Rv3448	MT3554	MTB45J	MB585.1J	No Sequence Data
		503	5	Rv1795	MT1844	MTB196J	MB506J	No Sequence Data
K	Mycosins, subtilisin-like cell-wall associated serine proteases, 43-49% homology	446	1	Rv3883c	MT3998	MTB12Ka	MB727.0K	No Sequence Data
		550	2	Rv3886c	MT4001(Frame)	MTB12Kb	MB727.2K	No Sequence Data
		461	3	Rv0291	MT0304	MTB203K	MB548K	No Sequence Data
		455	4	Rv3449	MT3555	MTB45K	MB585.1K	No Sequence Data
		585	5	Rv1796	MT1845	MTB196K	MB506K	No Sequence Data
L	2 x N-terminal transmembrane protein, 16-27% homology	462	1	Rv3882c	MT3997	MTB12La	MB727.0L	No Sequence Data
		537	2	Rv3885c	MT4000 (Frame)	MTB12Lb	MB727.2L	No Sequence Data
		331	3	Rv0292	MT0305	MTB203L	MB694.0L	No Sequence Data
		-	4	No Duplication	No Duplication	No Duplication	No Duplication	No Duplication
		406	5	Rv1797	MT1846	MTB196L	MB542L	No Sequence Data

Table 3.3. (Continued)

Gene Family	Description	Protein size (in *M.tb*)	ESAT-6 Cluster Region	M. leprae TN	Presence and names of genes in each species				
					M. avium* 104	***M. paratuberculosis**** K 10	***M. smegmatis**** MC^2 155	***C. diphtheriae**** NCTC13129	***S. coelicolor*** A3 (2)
A	ABC transporter family signature, 19-27% homology	283	1	ML0057(pseudo)	ND	ND	MS29A	ND	ND
		276	2	MLabc (pseudo)**	MA138A	MP3889c	ND	ND	ND
		295	3	ML2530	MA141A	MP0289	MS32A	ND	ND
		-	4	No Duplication	No Duplication	No Duplication	No Duplication	No Duplication	No Duplication
		300	5	ML1540	MA310A	MP1794	ND	ND	ND
B	AAA+ class ATPases, CBXX/CFQX family, SpoVK, 1 x ATP/GTP binding site, 29-39% homology	573	1	ML0055	ND	ND	MS29B	ND	ND
		619	2	ML0039(pseudo)	MA177B	MP3884c	ND	ND	ND
		631	3	ML2537	MA78B	MP0282	MS32B	ND	ND
		-	4	No Duplication	No Duplication	No Duplication	No Duplication	No Duplication	No Duplication
		610	5	ML1536	MA310B	MP1798	ND	ND	ND
C	N-terminal transmembrane protein, possible ATP/GTP binding motif, 31-41% homology	480	1	ML0054	ND	ND	MS29C	ND	ND
		495	2	Deleted	MA144C	MP3895c	ND	ND	ND
		538	3	ML2536	MA78C	MP0283	MS32C	ND	ND
		470	4	Deleted	MA94C	MP3450c	MS8C	CORDmem	SC3C3.07
		506	5	ML1544	MA221C	MP1782	ND	ND	ND
D	DNA segregation ATPase, ftsK chromosome partitioning protein, SpoIIIE, yukA, 3 x ATP/GTP binding sites, 2 x N-terminal transmembrane protein, 28-39% homology	747+591	1	ML0053+52	ND	ND	MS29D (Stop$)	ND	ND
		1396	2	Deleted	MA144D	MP3894c	ND	ND	ND
		1330	3	ML2535	MA78D	MP0284	MS32D	ND	ND
		1236	4	Deleted	MA504D	MP3447c	MS8D	CORDyuk	SC3C3.20c
		435+932	5	ML1543	MA221D	MP1783	ND	ND	ND
E	PE, 18-90% homology	99	1	Deleted	ND	ND	MS29E	ND	ND
		77	2	Deleted	MA138E	MP3893c	ND	ND	ND
		102	3	ML2534	MA78E	MP0285	MS32E	ND	ND
		-	4	No Duplication	No Duplication	No Duplication	No Duplication	No Duplication	No Duplication
		99 & 99	5	Deleted	MA310Ea & Eb	MP1788 & 91	ND	ND	ND
F	PPE, 19-88% homology	368	1	ML0051	ND	ND	MS29F	ND	ND
		399	2	Deleted	MA138F	MP3892c	ND	ND	ND
		513	3	ML2533 (pseudo)	MA78F	MP0286	MS32F	ND	ND
		-	4	No Duplication	No Duplication	No Duplication	No Duplication	No Duplication	No Duplication
		365, 393 & 350	5	Deleted	MA310Fa & Fb & Fc	MP1787 & 89 & 90	ND	ND	ND
G	lhp or CFP-10, also MTSA-10, grouped into ESAT-6 family, potent secreted T-cell antigens, 9-32% homology	100	1	ML0050	ND	ND	MS29G	ND	SC3C3.10 and SC3C3.11[c]
		107	2	Deleted	MA138G	MP3891c [a]	ND	ND	ND
		97	3	ML2532	MA141G	MP0287	MS32G	ND	ND
		125	4	Deleted	MA319G	MP3445c	MS8G	CORDcfp10	ND
		98	5	MLcfp (pseudo)**	MA310G	MP1792	ND	ND	ND
H	ESAT-6 family, cfp7, L45 or l-esat, also Mtb9.9 family, potent secreted T-cell antigens, 15-27% homology	95	1	ML0049	ND	ND	MS29H	ND	SC3C3.10 and SC3C3.11[c]
		95	2	ML0034 (pseudo)	MA138H	MP3890c [a]	ND	ND	ND
		96	3	ML2531	MA141H	MP0288	MS32H	ND	ND
		100	4	ML0363	MA319H	MP3444c	MS8H	CORDesat6	ND
		94	5	MLesat (pseudo)**	MA310H	MP1793	ND	ND	ND
I	ATPases involved in chromosome partitioning, 1 x ATP/GTP binding motif, 33% homology	666	1	ML0048	ND	ND	MS29I	ND	SC3C3.03c
		341	2	ML0035 (pseudo)	MA138I	MP3888c	ND	ND	ND
		-	3	No Duplication	No Duplication	No Duplication	No Duplication	No Duplication	No Duplication
		-	4	No Duplication	No Duplication	No Duplication	No Duplication	No Duplication	No Duplication
		-	5	No Duplication	No Duplication	No Duplication	No Duplication	No Duplication	No Duplication
J	Integral inner membrane protein, binding protein dependent transport systems inner membrane component signature, putative transporter protein, 19-27% homology)	511	1	ML0047	ND	ND	MS29J	ND	ND
		509	2	ML0036 (pseudo)	MA138J	MP3887c	ND	ND	ND
		472	3	ML2529	MA141J	MP0290	MS32J	ND	ND
		467	4	Deleted	MA504J	MP3448	MS8J	CORDtransporter	SC3C3.21
		503	5	ML1539	MA310J	MP1795	ND	ND	ND
K	Mycosins, subtilisin-like cell-wall associated serine proteases, 43-49% homology	446	1	ML0041	ND	ND	MS65K	ND	ND
		550	2	ML0037 (pseudo)	MA177K	MP3886c	ND	ND	ND
		461	3	ML2528	MA141K	MP0291	MS32K	ND	ND
		455	4	Deleted	MA439K	MP3449	MS8K	CORDsub	SC3C3.17c and SC3C3.08
		585	5	ML1538	MA310K	MP1796	ND	ND	ND
L	2 x N-terminal transmembrane protein, 16-27% homology	462	1	ML0042	ND	ND	MS65L	ND	ND
		537	2	ML0038 (pseudo)	MA177L	MP3885c	ND	ND	ND
		331	3	ML2527	MA81L	MP0292	MS32L	ND	ND
		-	4	No Duplication	No Duplication	No Duplication	No Duplication	No Duplication	No Duplication
		406	5	ML1537	MA310L	MP1797	ND	ND	ND

Other region-specific genes of known functions (not assigned to a family):

Region 5
(not present in *M. smegmatis*, *C. diphtheriae*
and *S. coelicolor*)

Rv1785c - **Probable member of the cytochrome P450 family (pseudogene in *M. leprae*)**
Rv1786 - **Probable ferredoxin (pseudogene in *M. leprae*)**

Other region-specific genes of unknown functions (not assigned to a family):

Region 1
(deleted in *M. avium* and *M. paratuberculosis*,
not present in *C. diphtheriae* and *S. coelicolor*)

Rv3867 - **Unknown, annotated as part of MT3980 (Rv3866) in *M. tuberculosis* CDC1551 sequence with a frameshift (functional in *M. leprae*)**
Rv3878 - **Unknown, some similarity to PPE family, deleted with RD1 deletion region in *M. bovis* BCG (pseudogene in *M. leprae*)**
Rv3879c - **Unknown, repetitive, highly proline-rich N-terminus, deleted with RD1 deletion region in *M. bovis* BCG (pseudogene in *M. leprae*)**
Rv3880c - **Unknown (functional in *M. leprae*)**
Rv3881c - **Unknown (pseudogene in *M. leprae*)**

Region 4
(not present in *S. coelicolor*)

Rv3446c - **Unknown, may contain a possible ABC transporter signature (deleted in *M. leprae*)**

* = Names of genes of these organisms were given arbitrarily by the authors of this paper; ** = Gene is present in the sequence, but not annotated (name given arbitrarily by authors of this paper); ND = Not detected - not necessarily absent from genome but possibly not detected because of unfinished sequencing process; No Duplication = No duplication of this gene is present in this region; No Sequence Data = No sequence data is available for this organism, published deletion information is included (Mahairas *et al.*, 1996 and others); Deleted = Deleted from the genome of this particular species or strain (# = deleted in only some strains of this species); Frame = Frameshift; Stop = In frame stopcodon, (a) = Genes identified by BLAST as well as data obtained from Genbank, accession no. AJ250015; [b] = Gene not identified by BLAST, data obtained from (Mahairas *et al.*, 1996), Genbank accession no U34848 and AAC44033, [c] = Orthologues in *S. coelicolor* are equally similar to Family G and H, Stop$ = Stopcodon corresponds to stopcodon in *M. tuberculosis* H37Rv, which splits gene into Rv3870 and Rv3871; pseudo = confirmed pseudogene due to multiple frameshifts and stopcodons

3.3.2.2. *Mycobacterium leprae*

Figure 3.3 shows a schematic representation of the genomic organization of the respective gene families present in each of the five ESAT-6 gene cluster regions of *M. leprae*. The genome sequence of *M. leprae* contains functional copies of two of the five ESAT-6 gene cluster regions (region 1 and 3 – sharing between 50 and 70% similarity to *M. tuberculosis* H37Rv at protein level). Most of the genes from region 2 are deleted, while all the remaining genes of this region became pseudogenes due to extensive point mutations. This is in contrast to the genes from region 1 (which lies directly adjacent to region 2) and which contains no pseudogenes. It is thus conceivable that these clusters should function as a unit, and that genes could become non-functional when part of the unit is disrupted. Furthermore, all of the genes immediately flanking the putative functional regions as well as five of the eight genes only present in one of the regions as depicted in Table 3.3 (the Rv1785c, Rv1786, Rv3878, Rv3879c and Rv3881c orthologues ML1542, ML1541, ML0046, ML0045 and ML0043), are probable pseudogenes, indicating that the genes present in the functional clusters are being maintained as a unit.

3.3.2.3. *Mycobacterium avium* and *Mycobacterium paratuberculosis*

The genomes of the *M. avium* strain 104 and the closely related species *M. paratuberculosis* (or *M. avium* subsp. *paratuberculosis*) has revealed four of the five ESAT-6 gene cluster regions (sharing between 65 and 75% similarity to *M. tuberculosis* H37Rv at protein level), with region 1 being absent in both species (Figure 3.4). Closer inspection of the gene sequence surrounding region 1 in both these species has revealed a deletion of the region containing region 1 and some upstream flanking genes (from the Rv3861 gene orthologue up to and including the Rv3883c orthologue). This deletion coincided with the insertion of a ± 2 292 bp sequence containing the genes for a putative hydroxylase (± 818 bp) and the *sigI* sigma factor (± 824 bp). The presence of this sequence in both genomes (99% DNA sequence identity) indicates that the insertion-deletion may have occurred before the divergence of the two species. The genes from the remaining ESAT-6 gene cluster regions that are present in *M. avium* and *M. paratuberculosis* contain no stop codons or frameshifts and thus appear to be functional.

Figure 3.3. Schematic representation of the genomic organization of the genes present in the five ESAT-6 gene cluster regions of *Mycobacterium leprae*. ORF's are represented as blocked arrows showing the direction of transcription, with the different colors reflecting the specific gene family and the length of the arrow reflecting the relative lengths of the genes as in Figure 3.1. Black arrows indicate unconserved genes present in these regions, while open arrows indicate pseudogenes. Annotations of *M. leprae* genes are according to Cole *et al.* (1998).

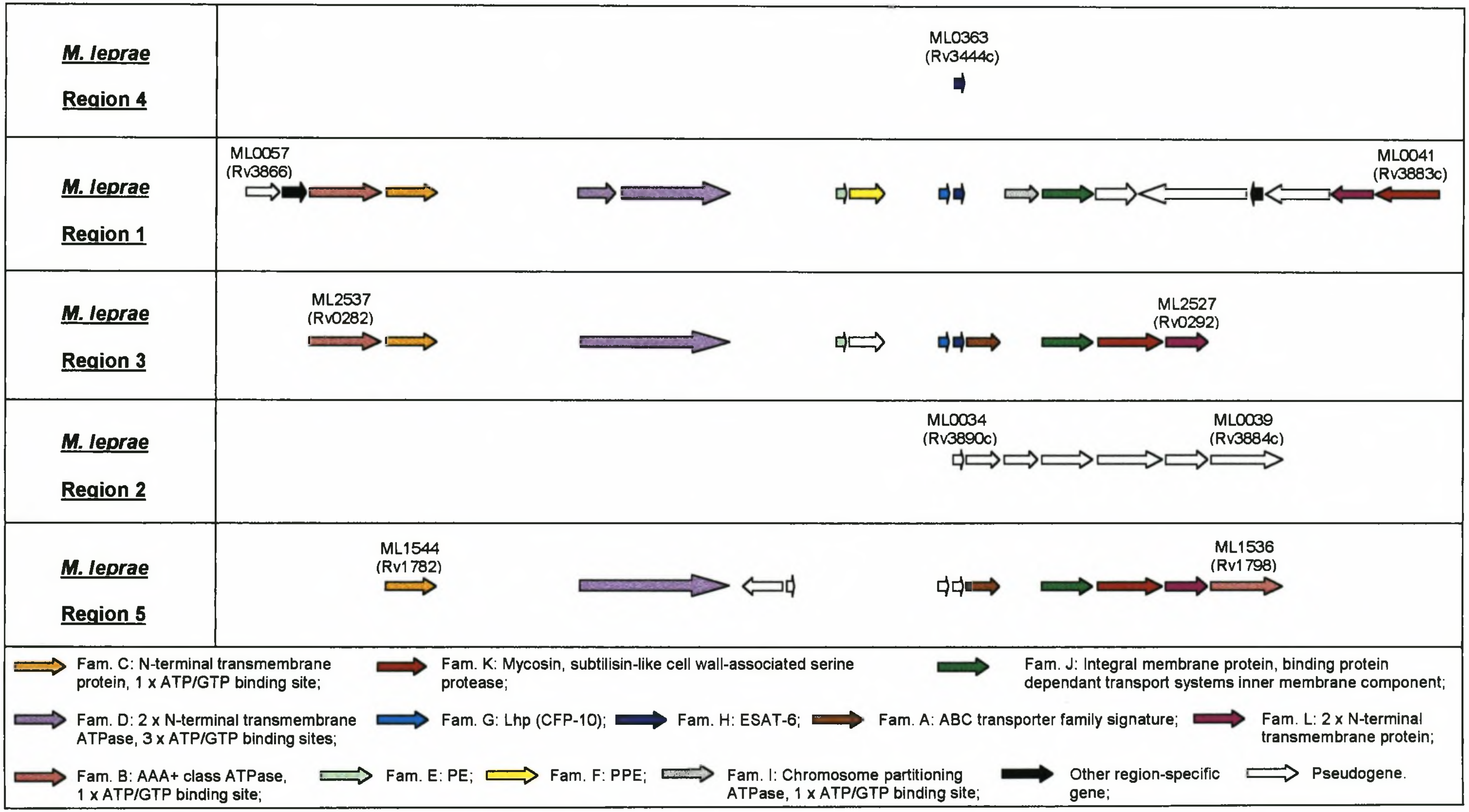
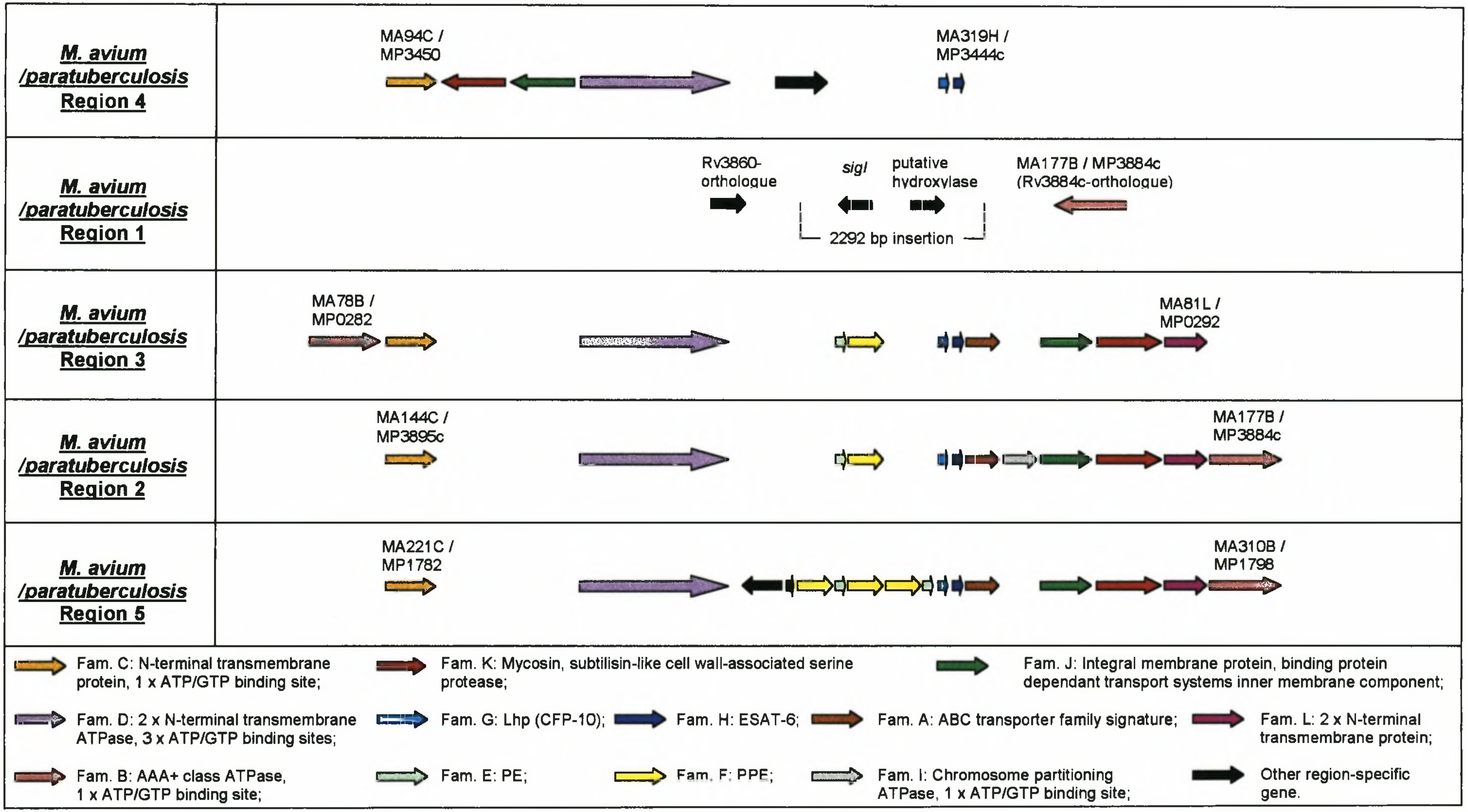

M. leprae
Region 4
ML0363
(Rv3444c)
M. leprae
Region 1
ML0057
(Rv3866)
ML0041
(Rv3883c)
M. leprae
Region 3
ML2537
(Rv0282)
ML2527
(Rv0292)
M. leprae
Region 2
ML0034
(Rv3890c)
ML0039
(Rv3884c)
M. leprae
Region 5
ML1544
(Rv1782)
ML1536
(Rv1798)
Fam. C: N-terminal transmembrane protein, 1 x ATP/GTP binding site;
Fam. K: Mycosin, subtilisin-like cell wall-associated serine protease;
Fam. J: Integral membrane protein, binding protein dependant transport systems inner membrane component;
Fam. D: 2 x N-terminal transmembrane ATPase, 3 x ATP/GTP binding sites;
Fam. G: Lhp (CFP-10);
Fam. H: ESAT-6;
Fam. A: ABC transporter family signature;
Fam. L: 2 x N-terminal transmembrane protein;
Fam. B: AAA+ class ATPase, 1 x ATP/GTP binding site;
Fam. E: PE;
Fam. F: PPE;
Fam. I: Chromosome partitioning ATPase, 1 x ATP/GTP binding site;
Other region-specific gene;
Pseudogene.

Figure 3.4. Schematic representation of the genomic organization of the genes present in the four ESAT-6 gene cluster regions of *Mycobacterium avium* and *Mycobacterium paratuberculosis*, as well as the flanking genes of the region 1 deletion. ORF's are represented as blocked arrows showing the direction of transcription, with the different colors reflecting the specific gene family and the length of the arrow reflecting the relative lengths of the genes as in Figure 3.1. Black arrows indicate unconserved genes present in these regions. *M. avium* and *M. paratuberculosis* genes were arbitrarily annotated by the authors of this paper.


M. avium /paratuberculosis Region 4
MA94C / MP3450
MA319H / MP3444c
M. avium /paratuberculosis Region 1
Rv3860-orthologue
sigI
putative hydroxylase
MA177B / MP3884c (Rv3884c-orthologue)
2292 bp insertion
M. avium /paratuberculosis Region 3
MA78B / MP0282
MA81L / MP0292
M. avium /paratuberculosis Region 2
MA144C / MP3895c
MA177B / MP3884c
M. avium /paratuberculosis Region 5
MA221C / MP1782
MA310B / MP1798
Fam. C: N-terminal transmembrane protein, 1 x ATP/GTP binding site;
Fam. K: Mycosin, subtilisin-like cell wall-associated serine protease;
Fam. J: Integral membrane protein, binding protein dependant transport systems inner membrane component;
Fam. D: 2 x N-terminal transmembrane ATPase, 3 x ATP/GTP binding sites;
Fam. G: Lhp (CFP-10);
Fam. H: ESAT-6;
Fam. A: ABC transporter family signature;
Fam. L: 2 x N-terminal transmembrane protein;
Fam. B: AAA+ class ATPase, 1 x ATP/GTP binding site;
Fam. E: PE;
Fam. F: PPE;
Fam. I: Chromosome partitioning ATPase, 1 x ATP/GTP binding site;
Other region-specific gene.

3.3.2.4. *Mycobacterium smegmatis*

The genome sequence of the avirulent, fast-growing mycobacterial species *M. smegmatis* contains three of the five ESAT-6 gene cluster regions, namely region 1, 3, and 4 (sharing between 60 and 75% similarity to *M. tuberculosis* H37Rv at protein level), with region 2 and 5 being absent (Figure 3.5). No deletions, frameshifts or stop codons were identified in any of the genes in the regions that are present and therefore it is concluded that these regions are functional.

3.3.3. ESAT-6 gene cluster identification in bacteria other than the mycobacteria

3.3.3.1. *Corynebacterium diphtheriae*

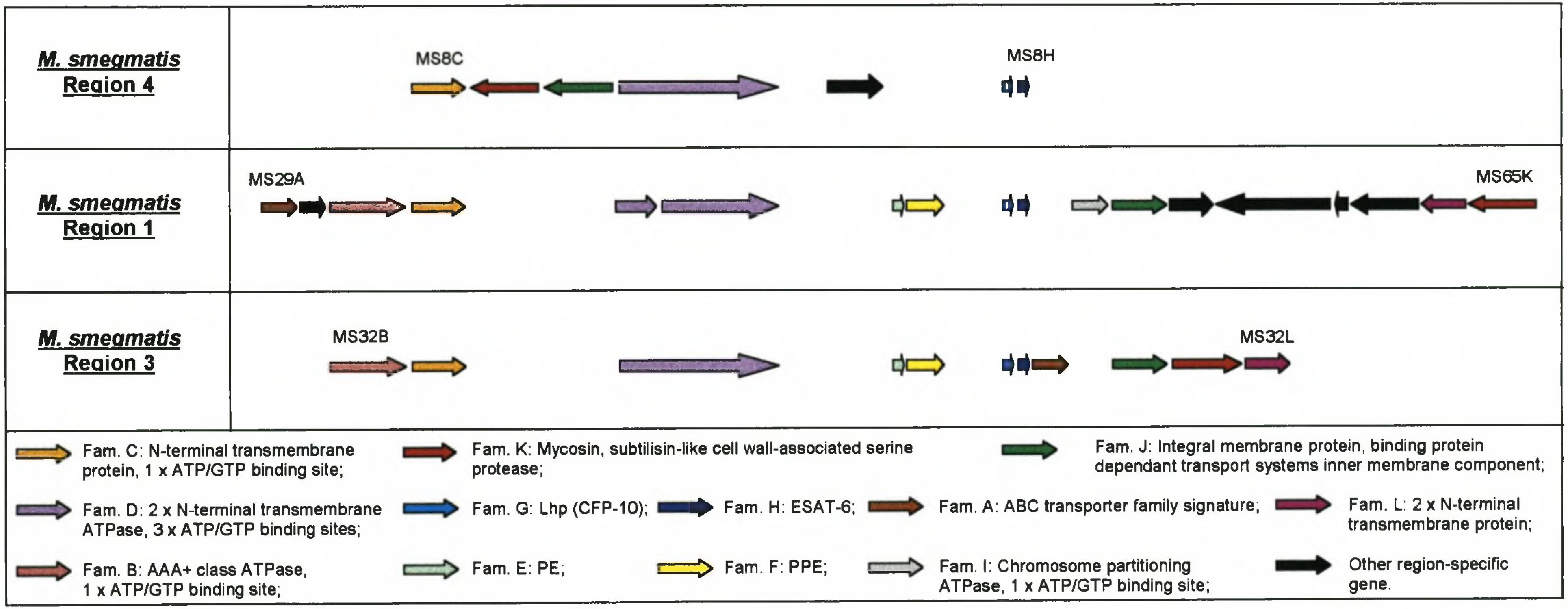
The genome sequence of the closely related *C. diphtheriae* has revealed a copy of the region 4 ESAT-6 gene cluster (Figure 3.1, see Table 3.4 for percentage similarity between sequences), situated in the same genomic location as in the mycobacteria (indicated by the large stretch of flanking genes homologous to the genes flanking region 4 in *M. tuberculosis* H37Rv). All the genes present within this cluster appear to be fully functional as no deletions, stop codons or frameshifts were identified. No duplications of the gene cluster could be detected in the genome of this organism.

Table 3.4. Similarity of *M. tuberculosis* H37Rv Region 4- encoded proteins to proteins encoded by the *C. diptheriae* and *S. coelicolor* regions

M. tuberculosis		**Percentage similarity**	
region 4 proteins	**Family**	***C. diptheriae***	***S. coelicolor***
Rv3450c	C	47%	36%
Rv3447c	D	53%	57%
Rv3445c	G	47%	47 and 51%*
Rv3444c	H	58%	41 and 44%*
Rv3448	J	33%	45%
Rv3449	K	49%	45 and 47%

* Orthologues in *S. coelicolor* are equally similar to Family G and H.

Figure 3.5. Schematic representation of the genomic organization of the genes present in the three ESAT-6 gene cluster regions of *Mycobacterium smegmatis*. ORF's are represented as blocked arrows showing the direction of transcription, with the different colors reflecting the specific gene family and the length of the arrow reflecting the relative lengths of the genes as in Figure 3.1. Black arrows indicate unconserved genes present in these regions. *M. smegmatis* genes were arbitrarily annotated by the authors of this paper.


M. smegmatis Region 4
MS8C
MS8H
M. smegmatis Region 1
MS29A
MS65K
M. smegmatis Region 3
MS32B
MS32L
Fam. C: N-terminal transmembrane protein, 1 x ATP/GTP binding site;
Fam. K: Mycosin, subtilisin-like cell wall-associated serine protease;
Fam. J: Integral membrane protein, binding protein dependant transport systems inner membrane component;
Fam. D: 2 x N-terminal transmembrane ATPase, 3 x ATP/GTP binding sites;
Fam. G: Lhp (CFP-10);
Fam. H: ESAT-6;
Fam. A: ABC transporter family signature;
Fam. L: 2 x N-terminal transmembrane protein;
Fam. B: AAA+ class ATPase, 1 x ATP/GTP binding site;
Fam. E: PE;
Fam. F: PPE;
Fam. I: Chromosome partitioning ATPase, 1 x ATP/GTP binding site;
Other region-specific gene.

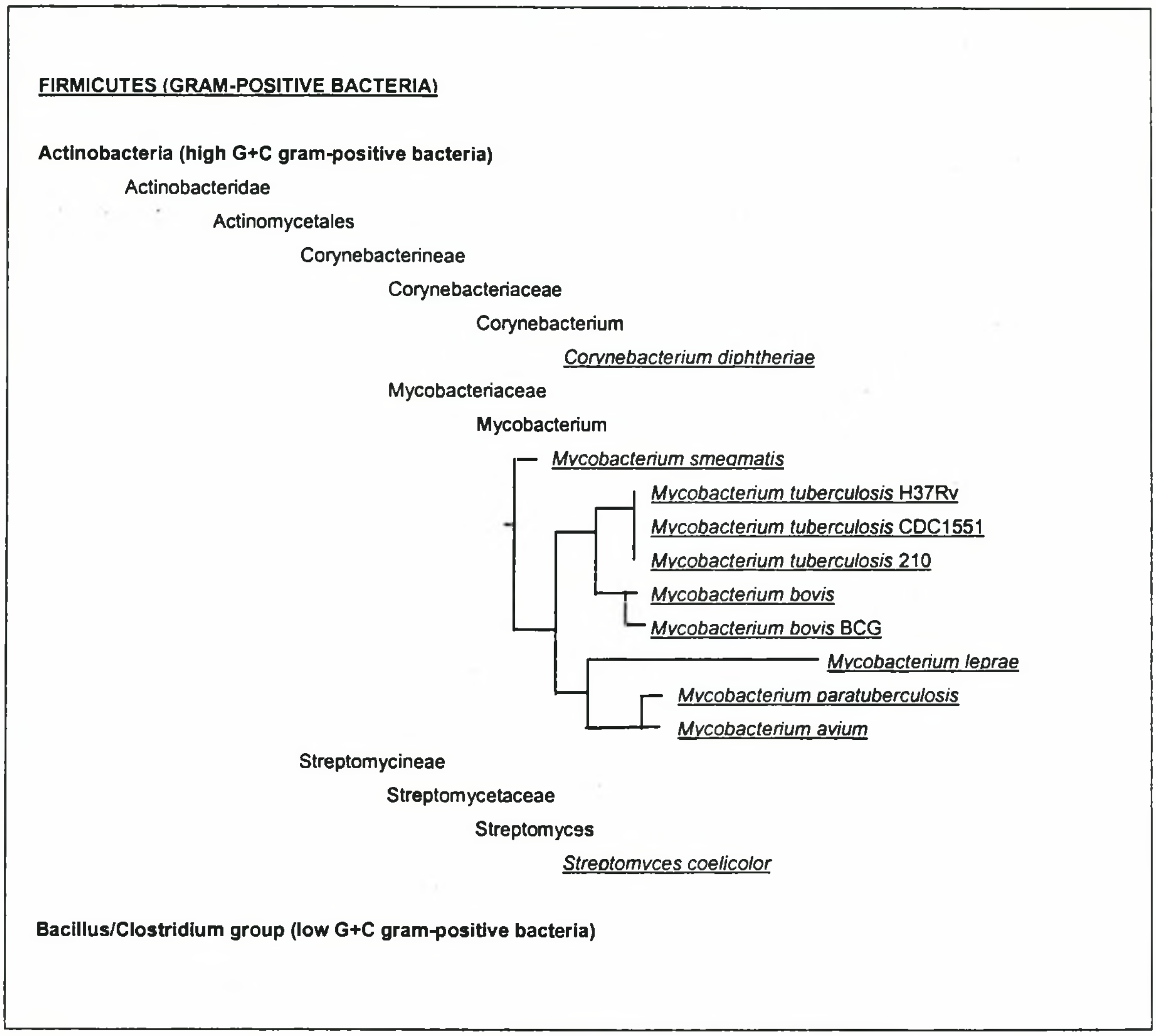
3.3.3.2. *Streptomyces coelicolor*

The *S. coelicolor* genome has revealed distinct orthologues for four of the six most conserved genes from the ESAT-6 gene cluster regions located in close proximity to each other (Figure 3.1). These genes show the highest similarity to the corresponding orthologues in region 4 of *M. tuberculosis* (see Table 3.4 for percentage similarity between sequences). There is also a very distinct orthologue (SC3C3.03c) of the region 1 family I gene (Rv3876) in the *S. coelicolor* region. There is no homologue for this gene in region 4 of *M. tuberculosis*. A sequence similarity search using the sequences of the other two proteins of region 4, namely ESAT-6 (Rv3444c) and CFP-10 (Rv3445c), has also revealed some similarity to two small genes situated within the same region in the genome of *S. coelicolor* (Table 3.4 and Figure 3.1). These genes (SC3C3.10 and SC3C3.11) encode small proteins (124 and 103aa) of unknown function, are very similar to each other, and lie adjacent to each other, similar to the observation for the ESAT-6/CFP-10 operon. The sequences of both of these proteins also contain the motif W-X-G, a feature present in most of the ESAT-6 and CFP-10 proteins. The higher degree of similarity between the genes from region 4 of the mycobacteria (and *C. diptheriae*) and those present in the region in *S. coelicolor* suggests that region 4 may be the ancestral region in the mycobacteria, although a number of differences between these regions do exist.

3.3.4. Taxonomy

It is evident from the taxonomy (Figure 3.6) of the different species of bacteria in which copies of the ESAT-6 gene clusters could be found, that the presence of these clusters appear to be a specific characteristic of the high G+C gram-positive Actinobacteria, and that multiple copies thereof are only found in the mycobacteria. No copies of the clusters could be found in the completed genome sequence of *Bacillus subtilus* and that of other related species, which also form part of the Firmicutes (gram-positive bacteria), but fall under the Bacillus/Clostridium group (low G+C gram-positive bacteria). No copies of these clusters could be found in the genomes of any other bacteria or organism outside of the Firmicutes and thus the ESAT-6 gene clusters appear to be unique to the Actinobacteria.

Figure 3.6. Taxonomical position of the bacterial species that have the ESAT-6 gene clusters present in their genomes. This indicates that the ESAT-6 gene clusters seem to be a feature of only the high G+C gram-positive bacteria (Actinobacteria) and that the presence of multiple copies of the gene clusters seems to be a characteristic only found in the mycobacteria. Phylogenetic relationships of members of the genus *Mycobacterium* indicated are based on 16S rRNA gene sequence information (Shinnick and Good, 1994).


FIRMICUTES (GRAM-POSITIVE BACTERIA)
Actinobacteria (high G+C gram-positive bacteria)
Actinobacteridae
Actinomycetales
Corynebacterineae
Corynebacteriaceae
Corynebacterium
Corynebacterium diphtheriae
Mycobacteriaceae
Mycobacterium
Mycobacterium smegmatis
Mycobacterium tuberculosis H37Rv
Mycobacterium tuberculosis CDC1551
Mycobacterium tuberculosis 210
Mycobacterium bovis
Mycobacterium bovis BCG
Mycobacterium leprae
Mycobacterium paratuberculosis
Mycobacterium avium
Streptomycineae
Streptomycetaceae
Streptomyces
Streptomyces coelicolor
Bacillus/Clostridium group (low G+C gram-positive bacteria)

3.3.5. Phylogeny of the ESAT-6 gene cluster

To calculate the phylogenetic relationships between the five duplicated ESAT-6 gene cluster regions in *M. tuberculosis* and to identify the ancestral region, detailed phylogenetic analyses were performed on each of the six protein families which are present in all five of these regions (family C, D, G, H, J and K). Figure 3.7 shows an example of a part of one of the multiple protein sequence alignments (family C), showing the high level of sequence identity shared among the family members in regions of homology. Although the individual protein families differ in the amount of sequence identity, they are all clearly derived from duplication events. These multiple sequence alignments were used in the subsequent phylogenetic analyses of the ESAT-6 gene cluster regions.

Figure 3.7. Partial multiple sequence alignment of Family C protein sequences.

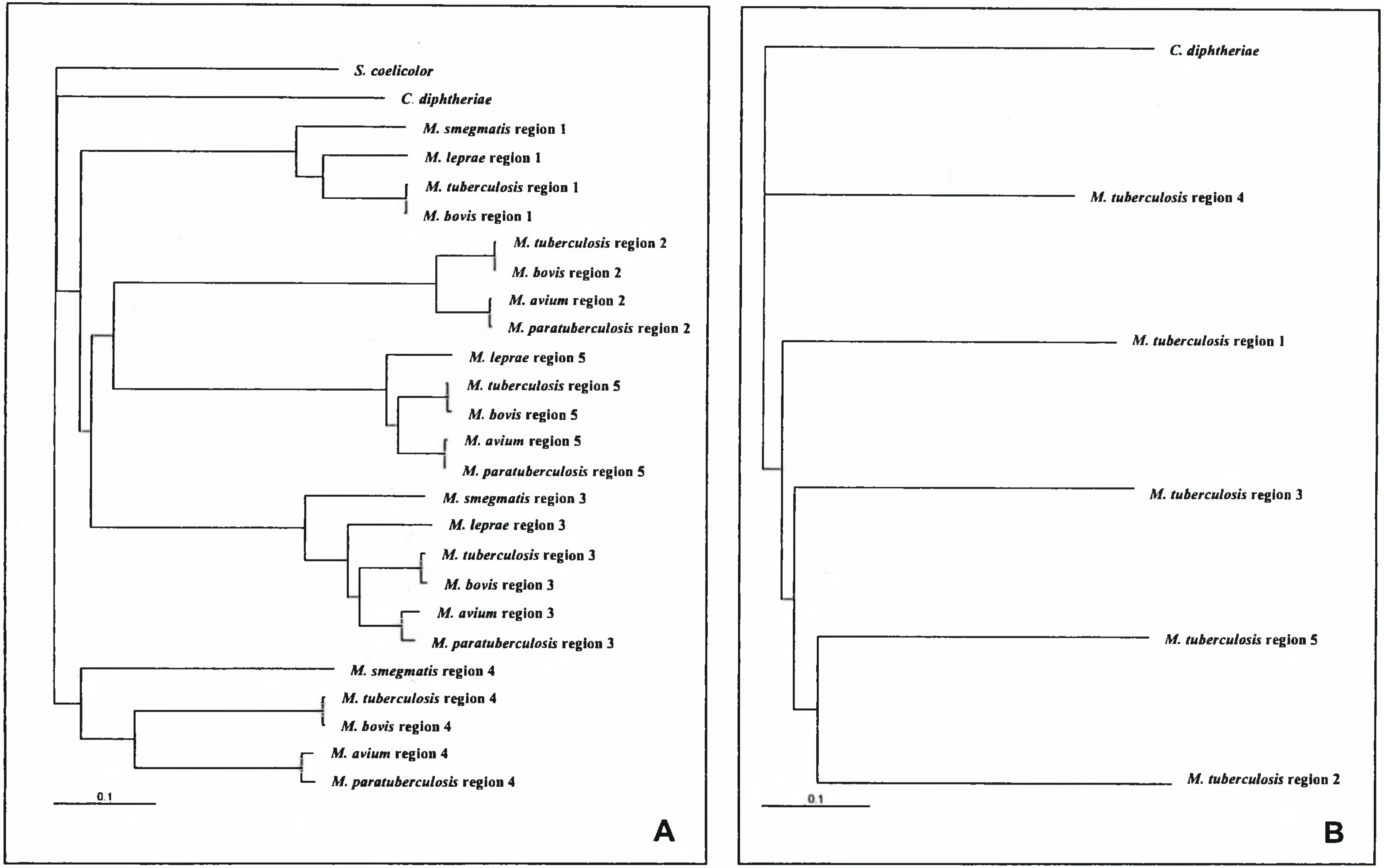
(*results not shown in published paper due to length restrictions*).

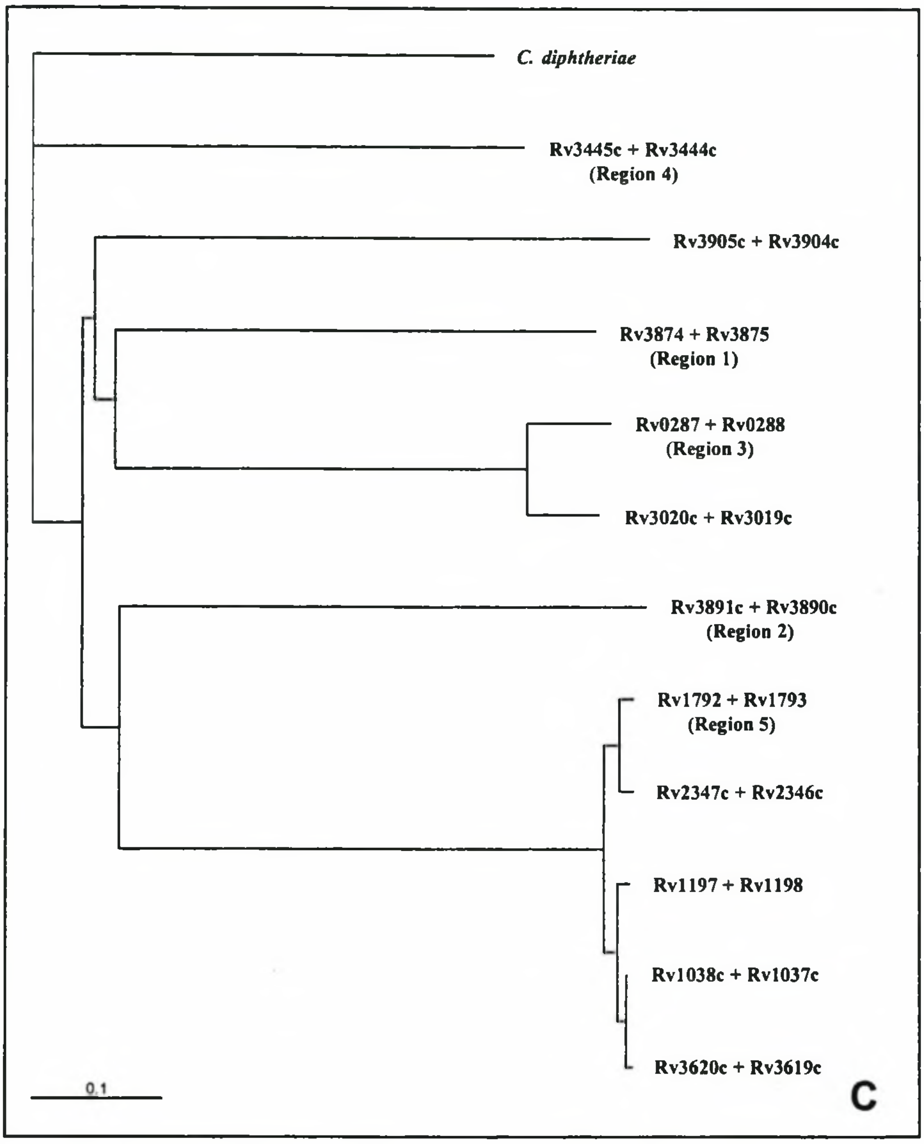
```
M.tuberculosis      region1   YVLLSG---QLHFVYNLTSARLVLGN-PANPATVKSSELS--KLPMGQTVGIPGAPYATP
M.bovis             region1   YVLLSG---QLHFVYNLTSARLVLGN-PANPATVKSSELS--KLPMGQTVGIPGAPYATP
M.leprae            region1   YVLLSG---HLYFVYNLTSARLALGK-PANPAAVKSSELT--KLPIGQTIGIPGAPYATP
M.smegmatis         region1   YVMLPGSN-QLRFVYNLTSARLVLGN-ASNPVAVKSEELN--RISKGQSIGIPGAPYATP
M.tuberculosis      region2   YARIDG---RLYFALNLTSARLATGT-AGQPTWVKPAEIA--KYPTGPLVGIPGAPAAMP
M.bovis             region2   YARIDG---RLYFALNLTSARLATGT-AGQPTWVKPAEIA--KYPTGPLVGIPGAPAAMP
M.avium             region2   YAKIDG---RLYFALNLTSARLATGT-ANQPTWVKPAEIA--KYPTGPLVGIPGAPAAMP
M.paratuberculosis  region2   YAKIDG---RLYFALNLTSARLATGT-ANQPTWVKPAEIA--KYPTGPLVGIPGAPAAMP
M.tuberculosis      region3   YVRVGE---QLHFVLNLTSARLIVGR-PVSPTTVKSTELD--QFPRGNLIGIPGAPERMV
M.bovis             region3   YVRVGE---QLHFVLNLTSARLIVGR-PVSPTTVKSTELD--QFPRGNLIGIPGAPERMV
M.leprae            region3   YVRVGD---ELHFVLNLTSATLIVGR-SVNPITVKSSELD--RFPRGNLIGIPGAPERMV
M.avium             region3   YVRVGD---DLHFVLNLTSARLIVGK-PVNPTTVKSAELD--RFPRGNLLGIPGAPERMV
M.paratuberculosis  region3   YVRVGE---DLHFVLNLTSARLITGR-PVDPTMVKSSELD--RFPRGNLIGIPGAPERMV
M.smegmatis         region3   YVRVGE---QLHFVLNLTSARLISGS-PDNPTMVKTSEID--KFPRGNLLGIPGAPERMV
M.tuberculosis      region4   YVRVDD---VWHFVLNLASARLIAAT-NANPQPVSESELG--HTKRGPLLGIPGAPQLLD
M.bovis             region4   YVRVDD---VWHFVLNLASARLIAAT-NANPQPVSESELG--HTKRGPLLGIPGAPQLLD
M.avium             region4   FVRVGD---TWHFVLNLASARLIAAS-AVDPLTIRQADLD--GSKRGFLLGIPGAPAFLG
M.paratuberculosis  region4   FVRVGD---TWHFVLNLASARLIAAS-AVDPLTIRQADLD--GSKRGPLLGIPGAPAFLG
M.smegmatis         region4   HVRIGD---TVHFVLNLASARLIVQN-PADPVLVDDAAID--AAPRGPLVGIPGAPARIH
C.diphtheriae                 FVRVDQ---QIHFVANLASARLVAGE-PAQPVKASDSILA--REHIGVPIGIPDAPRIVP
S.coelicolor                  VVLKTDGDTRLHFVLNLASARLLMKDGTYEVVQVGDDVLDSGEIPRGPILGIPYAPDRLP
M.tuberculosis      region5   YVRVGD---RLYFALNLASARLITGR-PDNPHLVRSSQIA--TMPRGPLVGIPGAPSSFS
M.bovis             region5   YVRVGD---RLYFALNLASARLITGR-PDNPHLVRSSQIA--TMPRGPLVGIPGAPSSFS
M.leprae            region5   YVRVGD---RLYFALNLASARLITGR-PDNPHAVRSSQIA--TLPHGPLVGIPGAPSELS
M.avium             region5   YVRVGD---RLYFALNLASARLITGR-PDNPHLVKSNQIA--SLPRGPMVGIPGAPSNFH
M.paratuberculosis  region5   YVRVGD---RLYFALNLASARLITGR-PDNPHLVKSNQIA--SLPRGPMVGIPGAPSNFH
                                          *  **:** *                :       *  :*** **
```

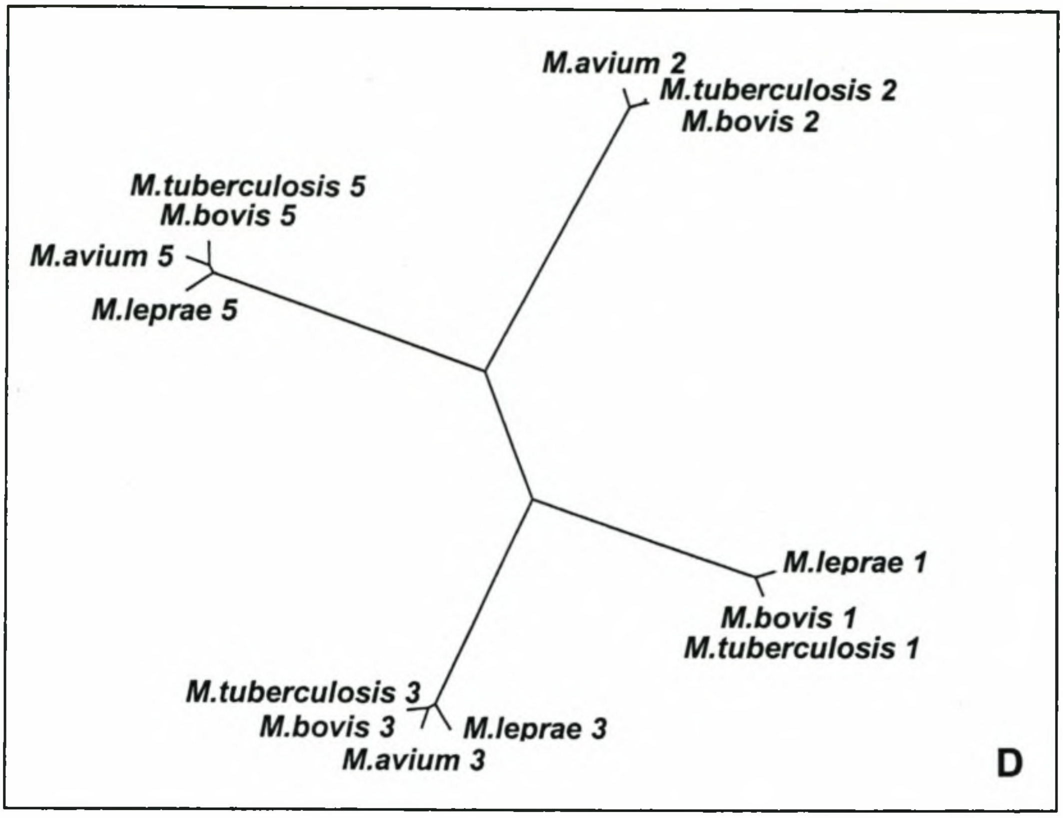
Figure 3.8A shows a neighbor-joining tree of the protein sequences of the ATP/GTP binding protein family (family D) from the ESAT-6 gene cluster regions of the mycobacteria and *C. diphtheriae*, with the protein orthologue of *S. coelicolor* as the outgroup. This tree is representative of all six trees that were drawn using the six families (data for the other trees are not shown). To confirm the results obtained with the *S. coelicolor* orthologues as outgroups, the same analyses were done using the *C. diphtheriae* orthologues as outgroups, with comparable results (results not shown). This tree topology was not due to systematic error as trees drawn using the FITCH algorithm gave the same results (results not shown). To confirm the basic structure of the trees and to verify that this structure is not influenced by the choice of outgroup, unrooted trees without any outgroup were constructed using the KITSCH algorithm, once again with comparable results (Figure 3.8D). To further verify the relationships among these clusters, the conserved sequences of all six proteins from *M. tuberculosis* were combined into one protein sequence and the same analysis was performed. The result of this analysis is presented in Figure 3.8B.

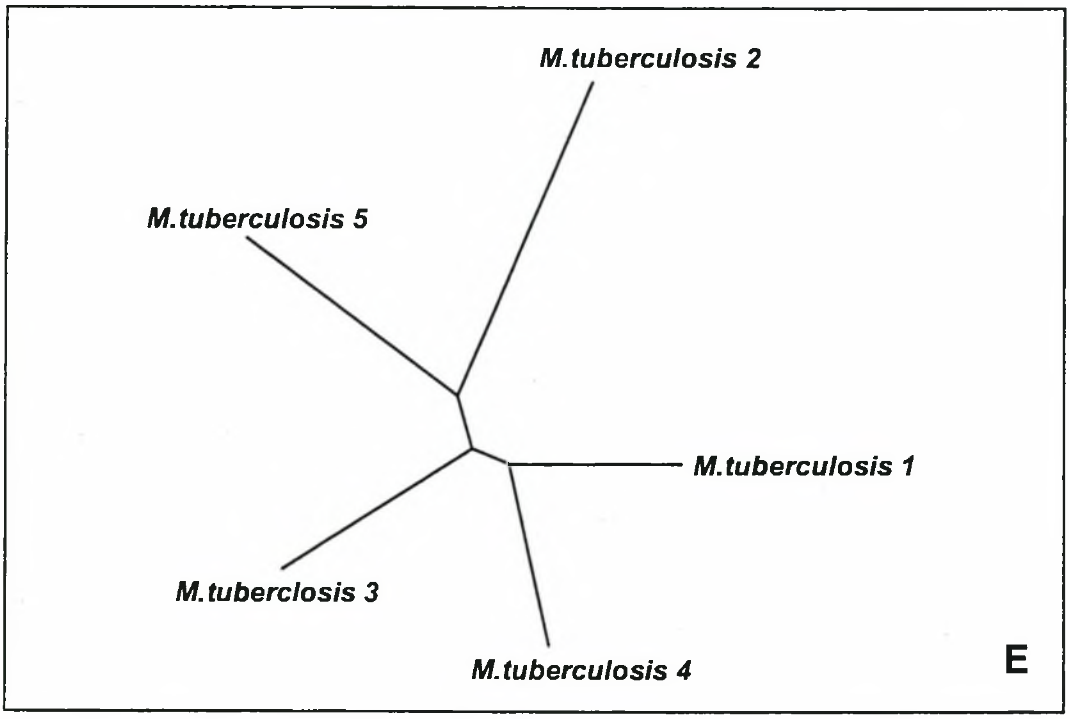
To investigate whether the non-conserved protein families (in other words those that are not present in region 4 of the mycobacteria, *C. diphtheriae* or *S. coelicolor*) show the same basic phylogenetic relationships as the conserved families (present in all five regions), an analysis was done on the AAA+ class ATPases family (family B). This family does not have a homologue in region 4 and there is also no *C. diphtheriae* or *S. coelicolor* orthologue to use as outgroup. The tree constructed from the data of this family clearly demonstrated once again that region 2 and 5, and region 1 and 3, respectively, are phylogenetically closer to each other (Figure 3.8E).

Figure 3.8. Phylogenetic trees showing the relationships between the five duplicated gene cluster regions. (A) Neighbor-joining phylogenetic tree of all available protein sequences of the ATP/GTP-binding protein family (family D in Table 3.3) with the protein orthologue of *Streptomyces coelicolor* as the outgroup. This tree is representative of all the trees drawn using the six most conserved proteins in these regions as well as using the protein orthologue of *Corynebacterium diphtheriae* as the outgroup. (B) Neighbor-joining phylogenetic tree of all six conserved proteins from the *M. tuberculosis* gene clusters combined into one protein per region. The combined protein of *C. diphtheriae* was used as the outgroup. (C) Neighbor-joining phylogenetic tree of the ESAT-6 and CFP-10 protein families combined (family G and H), using the combined protein of *C. diphtheriae* as the outgroup. (D) Unrooted phylogenetic tree, drawn without an outgroup, of all five protein sequences of the ATP/GTP binding protein family (family D) of *M. tuberculosis* H37Rv using the KITSCH algorithm. (E) Unrooted phylogenetic tree, drawn without an outgroup, of the AAA+ class ATPases family (Family B), which is one of the unconserved gene families found only in region 2 – 5 and not in *C. diptheriae* and *S. coelicolor* (*results for D and E not shown in published paper due to length restrictions*).


S. coelicolor
C. diphtheriae
M. smegmatis region 1
M. leprae region 1
M. tuberculosis region 1
M. bovis region 1
M. tuberculosis region 2
M. bovis region 2
M. avium region 2
M. paratuberculosis region 2
M. leprae region 5
M. tuberculosis region 5
M. bovis region 5
M. avium region 5
M. paratuberculosis region 5
M. smegmatis region 3
M. leprae region 3
M. tuberculosis region 3
M. bovis region 3
M. avium region 3
M. paratuberculosis region 3
M. smegmatis region 4
M. tuberculosis region 4
M. bovis region 4
M. avium region 4
M. paratuberculosis region 4
0.1
A
C. diphtheriae
M. tuberculosis region 4
M. tuberculosis region 1
M. tuberculosis region 3
M. tuberculosis region 5
M. tuberculosis region 2
0.1
B


C. diphtheriae
Rv3445c + Rv3444c
(Region 4)
Rv3905c + Rv3904c
Rv3874 + Rv3875
(Region 1)
Rv0287 + Rv0288
(Region 3)
Rv3020c + Rv3019c
Rv3891c + Rv3890c
(Region 2)
Rv1792 + Rv1793
(Region 5)
Rv2347c + Rv2346c
Rv1197 + Rv1198
Rv1038c + Rv1037c
Rv3620c + Rv3619c
0.1
C


M.avium 2
M.tuberculosis 2
M.bovis 2
M.tuberculosis 5
M.bovis 5
M.avium 5
M.leprae 5
M.leprae 1
M.bovis 1
M.tuberculosis 1
M.tuberculosis 3
M.bovis 3
M.leprae 3
M.avium 3
D


M.tuberculosis 2
M.tuberculosis 5
M.tuberculosis 1
M.tuberclosis 3
M.tuberculosis 4
E

Neighbor joining, FITCH, KITSCH and concatenated sequence comparison analyses all supported a single phylogeny that indicates that region 4 seems to be the most ancient of the mycobacterial ESAT-6 gene cluster regions. Region 4 is also the closest region to the *S. coelicolor* and *C. diphtheriae* regions. The order of duplication seems to extend from region 4, through 1 and 3 to regions 2 and 5. The phylogenetic relationships between corresponding clusters in the different mycobacteria are maintained throughout the different protein family trees, and are in agreement with the proposed phylogenetic order (or taxonomical position) of the mycobacterial species according to 16S rRNA data (see Figure 3.6).

As the genome of *M. tuberculosis* contains 11 copy pairs of the ESAT-6/CFP-10 genes that appear to be duplicated together, phylogenetic trees were constructed using the ESAT-6 or CFP-10 proteins separately (results not shown), or in combination as one ESAT-6/CFP-10 protein (Figure 3.8C). Using the combined *C. diphtheriae* ESAT-6/CFP-10 orthologue protein as outgroup, the same organization of duplication events was obtained with region 1, 3, 2 and lastly 5 being duplicated from the ancient region 4. The other copies of the ESAT-6/CFP-10 operon pairs that are present in the *M. tuberculosis* genome sequence, but are not part of the ESAT-6 gene cluster regions, seem to have arisen from singular duplication events originating from different cluster regions. It is interesting to note that the ESAT-6 and CFP-10 genes from region 5 seem to be highly prone to duplication, as there are four additional copies of these two genes present in the genome, compared to just one additional copy originating from region 4 and region 3, respectively. These four gene duplicates of ESAT-6 and CFP-10 from region 5 are also nearly identical (93 - 100% similarity at protein level), indicating their recent duplication.

8.4. Discussion

It was recently estimated in an *in silico* analysis of the genome sequence of *M. tuberculosis* H37Rv, that 52% of the proteome has been derived from gene duplication events (Tekaia *et al.*, 1999). One of these duplication events involves multiple copies of the genes of the secreted T-cell antigens ESAT-6 and CFP-10 (Andersen *et al.*, 1995, Sørensen *et al.*, 1995, Van Pinxteren *et al.*, 2000) together with a number of associated genes. A total of twelve gene families were identified in five regions (which were termed the ESAT-6 loci).

Phylogenetic analyses of the protein sequences of the six most conserved gene families, present within the five regions, predict that region 4 (Rv3444c to Rv3450c) is the ancestral region. Region 4 also contains the least number of proteins [only 6 compared to the 12 of region 1 (Rv3866-3883c) and region 2 (Rv3884c-3895c)], and does not contain the PE and PPE genes, which appear to may have been have been inserted into this region following the first duplication. Phylogenetic analyses using different methods and protein family data also suggests that subsequent duplications took place in the following order: region 1 (Rv3866-3883c) → 3 (Rv0282-0292) → 2 (Rv3884c-3895c) → 5 (Rv1782-1798). Furthermore, these analyses support the taxonomical order observed for the mycobacteria, with *M. smegmatis* being taxonomically the farthest removed from *M. tuberculosis*. The presence of a copy of region 4 and its flanking genes in *C. diphtheriae* strengthens the taxonomical data that implies that the corynebacteria and mycobacteria have a common ancestor. It appears that *C. diphtheriae* have diverged from the mycobacteria before the multiple duplications of the ESAT-6 gene cluster, as only one copy of this cluster could be identified in the genome of this organism.

The loss of region 1 from the genomes of the species *M. avium* and *M. paratuberculosis* (belonging to the *M. avium* complex) is confirmed by clinical data showing that HIV-sero-negative patients infected with mycobacteria belonging to the *M. avium* complex, do not respond to ESAT-6 from region 1, but do recognize PPD and *M. avium* sensitins (Lein *et al.*, 1999). The ESAT-6 and CFP-10 genes encoded by region 1 are also not found in *M. bovis* BCG and have thus been a focus of recent research efforts because of their application as diagnostic markers to differentiate between BCG vaccination and *M. tuberculosis*, *M. bovis*, or *M. avium* infection (Van Pinxteren *et al.*, 2000, Vordermeier *et al.*, 2000, etc.). In this study we have found several copies of ESAT-6 and CFP-10 (with differing degrees of similarity) in the genomes of different mycobacteria (80% and 71% protein sequence similarity for ESAT-6 and CFP-10 respectively from region 1 in avirulent *M. smegmatis*) as well as orthologues in species outside of the mycobacteria, therefore care should be taken when using these proteins for diagnostic purposes. It would be important to look at the protein sequence similarity between the copies of ESAT-6 and CFP-10 of different virulent and environmental mycobacterial species before a member of these immunodominant protein families can be chosen as a definite marker of *M. tuberculosis* infection. Studies to determine the IFN-gamma production in response to ESAT-6 and CFP-10 from environmental mycobacteria (for example *M. smegmatis*) by peripheral blood mononuclear cells from infected patients have not been done. Until these results are available indicating that the T-cell responses against these proteins are not comparable to those found with the *M. tuberculosis* proteins, care should be taken with claims regarding the potential diagnostic value of these antigens.

Most of the sequences of the genes belonging to the ESAT-6 gene cluster regions contain no stop codons or frameshifts and thus appear to be functional. This is significant when placed into the context of a bacterium like *M. leprae*, since it is hypothesized that the genome of *M. leprae* may contain the minimal gene set required by a pathogenic mycobacterium (Brosch *et al.*, 2000b, Wixon, 2000, Cole *et al.*, 2001) and that the activities of some functional genes once present in the genome of *M. leprae* have been silenced (they became pseudogenes through multiple stop codon mutations and frameshifts) because it is no longer needed for its intracellular survival (Cole *et al.*, 1998). It appears as if *M. leprae* contains at least two functional copies of the ESAT-6 gene cluster in its genome (region 1 and 3). The *M. leprae* ESAT-6 copy from region 1 (the L45-antigen or L-ESAT antigen from clone L45) was shown to be strongly reactive to sera from leprosy patients (Sathish *et al.*, 1990), further providing experimental data that at least one of the cluster regions are definitely functional in *M. leprae*.

As most of the genes present within the ESAT-6 gene cluster regions encode for proteins that are predicted to be associated with transport and energy-providing systems, we hypothesize that these proteins may be involved in the secretion of a substrate across the mycobacterial cell wall. It is well known that the T-cell antigens ESAT-6 and CFP-10 are found in short term culture filtrates (ST-CF) of *M. tuberculosis*, although the mechanism by which secretion occurs are unknown, as these proteins do not possess any ordinary *sec*-dependant secretion signals (Andersen *et al.*, 1995, Sørensen *et al.*, 1995, Berthet *et al.*, 1998). Therefore it is possible that the genes present within the ESAT-6 gene cluster regions function together to provide a system for the secretion of the ESAT-6 and CFP-10 proteins. There is evidence for the processing of the TB10.4 protein (the ESAT-6 family member belonging to region 3) to a lower molecular weight product (Skjøt *et al.*, 2000), suggesting a possible role for the cell wall-associated mycosin proteases (Brown *et al.*, 2000) in the hypothesized transport system. Most of region 1 is situated in the RD1 deletion region of *M. bovis* BCG, possibly explaining the absence of expression of the mycosin-1 gene (Rv3883c) in BCG (Brown *et al.*, 2000). The hypothesis that a dependent functional relationship exists between the genes contained in these regions is further supported by the *M. leprae* sequence data, which shows that a deletion of part of the ESAT-6 gene cluster region 2 apparently caused the remaining genes in the region to become pseudogenes, or vice versa. Furthermore, Wards and coworkers (2000) produced an *M. bovis* knockout mutant of the ATPase gene Rv3871 (family D) in the ESAT-6 gene cluster region 1, resulting in a strain that did not sensitize guinea pigs to an ESAT-6 skin test. These results indicate a close relationship between the genes contained within these regions.

Wards *et al.* (2000) showed that an *esat-6/cfp-10* knockout mutant of *M. bovis* was less virulent than its parent if gross pathology, histopathology and mycobacterial culture of tissues were taken into account. These results, combined with the fact that multiple copies of the ESAT-6 gene clusters are found in all the mycobacteria, clearly indicate that they form an important part of the mycobacterial genomic composition. The presence of multiple duplications of the ESAT-6 gene cluster regions in the mycobacteria may be a significant difference between the members of this genus and other high G+C gram positives. Although the function of this cluster is presently unknown, there is sufficient evidence to indicate that it is of importance to the mycobacteria, and needs to be investigated further.

NOTE ADDED IN PROOF: *Since the publication of the results described in the preceding chapter the whole genome sequence of another high G+C Gram positive organism belonging to the genus Corynebacterium, namely Corynebacterium glutamicum, has been completed (http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/genom_table_cgi). The genome sequence of C. glutamicum was subsequently also analyzed to determine the presence of the ESAT-6 gene cluster regions. The results of this analysis indicate that there is only one copy of the ESAT-6 gene cluster region present in the genome of C. glutamicum, which is an orthologue of region 4. This result supports the data obtained from the C. diphtheriae genome sequence analyses as well as the phylogenetic analyses indicating that region 4 is the most ancient ESAT-6 gene cluster. The genomes of a further two high G+C gram positive organisms, namely Thermobifida fusca (http://www.jgi.doe.gov/JGI_microbial/html/thermobifida/thermob_homepage.html) and Clavibacter michiganensis (http://www.sanger.ac.uk/Projects/C_michiganensis/), as well as a further three members of the genus Mycobacterium, namely Mycobacterium ulcerans, Mycobacterium miroti (http://www.pasteur.fr/recherche/unites/Lgmb/mycogenomics.html) and Mycobacterium marinum (http://www.sanger.ac.uk/Projects/M_marinum/), are in the process of being sequenced and will undoubtedly provide further clues to the distribution and evolution of the ESAT-6 gene cluster regions in the near future.*

ADDENDA TO CHAPTER THREE

ADDENDUM 3A

INDIVIDUAL PROTEIN FAMILIES, IMMUNOLOGICAL ASPECTS AND LINKS TO PATHOGENICITY

"There is a dread disease which so prepares the victim, as it were, for death; which so refines it of its grosser aspect, and throws around familiar looks, unearthly indications of the coming change – a dread disease, in which the struggle between soul and body is so gradual, quite, and solemn, and the result so sure, that day by day, and grain by grain, the mortal part wastes and withers away, so that the spirit grows light and sanguine with its lightening load, and, feeling immortality at hand, deems it but a new term of mortal life; a disease in which death takes the glow and hue of life, and life the gaunt and grisly form of death; a disease which medicine never cured, wealth warded off, or poverty could boast exemption from; which sometimes moves in giant strides, and sometimes at a tardy pace; but, slow or quick, is ever sure and certain."

Nicholas Nickleby – Charles Dickens (1870)

NOTE: The results presented in the following addendum will be submitted as part of a review article for peer review and publication as: "**The mycobacterial ESAT-6 gene clusters**, Gey van Pittius, N.C., Warren, R.M., and Van Helden, P.D."

3A.1. Introduction

It has been known for many years that molecules secreted by *M. tuberculosis* during the early phases of infection can be targets for a protective immune response and could thus possibly be utilized in a novel antituberculosis vaccine (Andersen *et al.*, 1991a, Andersen *et* al., 1991b, Pal and Horwitz, 1992;). A number of studies have shown that the two most antigenic fractions of short-term culture filtrates (ST-CF's) are the low-molecular weight fractions of secreted proteins ranging in molecular mass from 3 to 12 kDa and from 25 to 31 kDa (Andersen and Heron, 1993, Boesen *et al*, 1995). These studies have shown that T lymphocytes producing high levels of gamma-interferon are specifically directed to the abovementioned fractions. A number of small, potently immunogenic proteins have been identified from the low molecular mass fraction, which includes members of the ESAT-6 and CFP-10 protein families. These are potent T-cell antigens of between 90 and 125aa that are secreted without any obvious secretion signal. Recently, it has been shown that the genes encoding for these proteins are situated within a cluster of genes named the ESAT-6 gene cluster regions (Tekaia *et al.*, 1999, Gey van Pittius *et al.*, 2001). These clusters contain genes encoding for energy-providing and transport associated proteins and have thus been hypothesized to function together to form an energy-dependant active transport secretion system for the secretion of the immunologically-important ESAT-6 and CFP-10 protein families (Tekaia *et al.*, 1999, Gey van Pittius *et al.*, 2001).

This study aims to provide an extensive review of all currently available data concerning the proteins encoded by the ESAT-6 gene cluster regions of *M. tuberculosis*, as there is a wealth of unprocessed data concerning this T-cell antigen family and its proposed biosynthetic gene clusters.

3A.2. Distribution in the genus Mycobacterium

The distribution of the five ESAT-6 gene clusters within different bacterial species has recently been described in a detailed analysis of publicly available whole genome sequencing data (Gey van Pittius *et al.*, 2001). The results of this study indicate that the ESAT-6 gene clusters appear to be a feature of the Actinobacteria (high G+C Gram positive bacteria), with multiple duplications being limited to members of the genus Mycobacterium.

In addition to the whole genome sequencing information, several recent publications have described experiments done on genes or regions of DNA within the ESAT-6 gene cluster regions (Sorenson *et al.*, 1995, Harboe *et al.*, 1996, Mahairas *et al.*, 1996, Gormley *et al.*, 1997, Alderson *et al.*, 2000, Colangeli *et al.*, 2000, Skjøt *et al.*, 2000). ESAT-6 gene cluster region 1 (Rv3866-Rv3883c) has enjoyed preference throughout these studies because of the presence of the immunologically important ESAT-6 (Rv3875) and CFP-10 (Rv3874) genes (Skjøt *et al.*, 2001) and the RD1 deletion region within this cluster. RD1 is commonly thought to be the primary deletion that occurred during the serial passage of *M. bovis* by Calmette and Guérin between 1908 and 1921 (Calmette and Guérin, 1920, Calmette, 1927, Calmette, 1928, Guérin, 1928, Guérin, 1948, Guérin, 1980), and is thus thought to possibly be responsible for the primary attenuation of *M. bovis* to *M. bovis* BCG (Behr *et al.*, 1999, Brosch *et al.*, 2000a). As no immunological significance has so far been attributed to the ESAT-6 and CFP-10 copies within Region 4 (Rv3444c-Rv3450c)(Skjøt *et al.*, 2001), this region has been largely ignored in the past and is thus the only one of the five ESAT-6 gene cluster regions for which no experimental data is available in the literature. This is despite the fact that this is probably the progenitor ESAT-6 gene cluster (from where subsequent duplications took place), and that this is the only one of the ESAT-6 gene cluster regions also present in the Corynebacteria (Gey van Pittius *et al.*, 2001). By performing a comprehensive analysis of the Southern, Western and PCR results from all of the abovementioned publications, it is possible to obtain further insight into the distribution of the ESAT-6 gene clusters in different members of the genus Mycobacterium (Table 3A.1). Figure 3A.1 summarizes the results from this data on a phylogenetic tree of the mycobacteria, revealing that the multi-duplication of the ESAT-6 clusters are conserved in (although not necessarily restricted to) the slow-growing pathogenic mycobacterial complexes encompassing *M. tuberculosis* and *M. avium*.

It is interesting to note how the experimental results showing the absence of region 1 in *M. avium* (Sorenson *et al.*, 1995, Harboe *et al.*, 1996, Gormley *et al.*, 1997, Skjøt *et al.*, 2000, Colangeli *et al.*, 2000), compliment the whole genome sequencing data which confirmed the deletion of this region in the organism (Gey van Pittius *et al*, 2001). The sequencing data also revealed a corresponding absence of this region in the genome of the closely related species *M. paratuberculosis*. Furthermore, the experimental data revealed the absence of region 1 in *M. intracellulare* (Sorenson *et al.*, 1995, Harboe *et al.*, 1996, Skjøt *et al.*, 2000), indicating that the deletion of region 1 is probably a feature shared by all members of the MAI complex (*M. avium, paratuberculosis* and *intracellulare*). This is confirmed by clinical data indicating that there are no *in vitro* gamma interferon responses by peripheral blood mononuclear cells to Region 1 ESAT-6 in patients infected with a member of the *M. avium* complex, although they do respond to PPD and *M. avium* sensitins (Lein *et al.*, 1999).

Many authors (for example Picken *et al.*, 1988, Ramos, 1994) include the very similar *M. scrofulaceum* species into this complex (describing it as the MAIS complex). According to Shinnick and Good (1994), this is incorrect, because this species is phylogenetically and phenetically distinct from *M. avium* and *M. intracellulare* and it is thus not a member of the complex. What is very interesting to note though, is the observation that the *M. scrofulaceum* species is also one of the very few species that does not contain a copy of region 1, similar to the other members of the *M. avium* complex, while the species which separate the *M. avium* complex from *M. scrofulaceum* on the proposed phylogenetic tree (Figure 3A.1, for example *M. gastri* and *M. kansasii*), do contain this region. Although the absence of region 1 in *M. scrofulaceum* could be due to a separate deletion event, this observation may indicate a phylogenetically closer relationship between *M. scrofulaceum* and the members of the MAI complex than what is proposed by Shinnick and Good (1994) and could thus be an important observation in the taxonomical classification of the species.

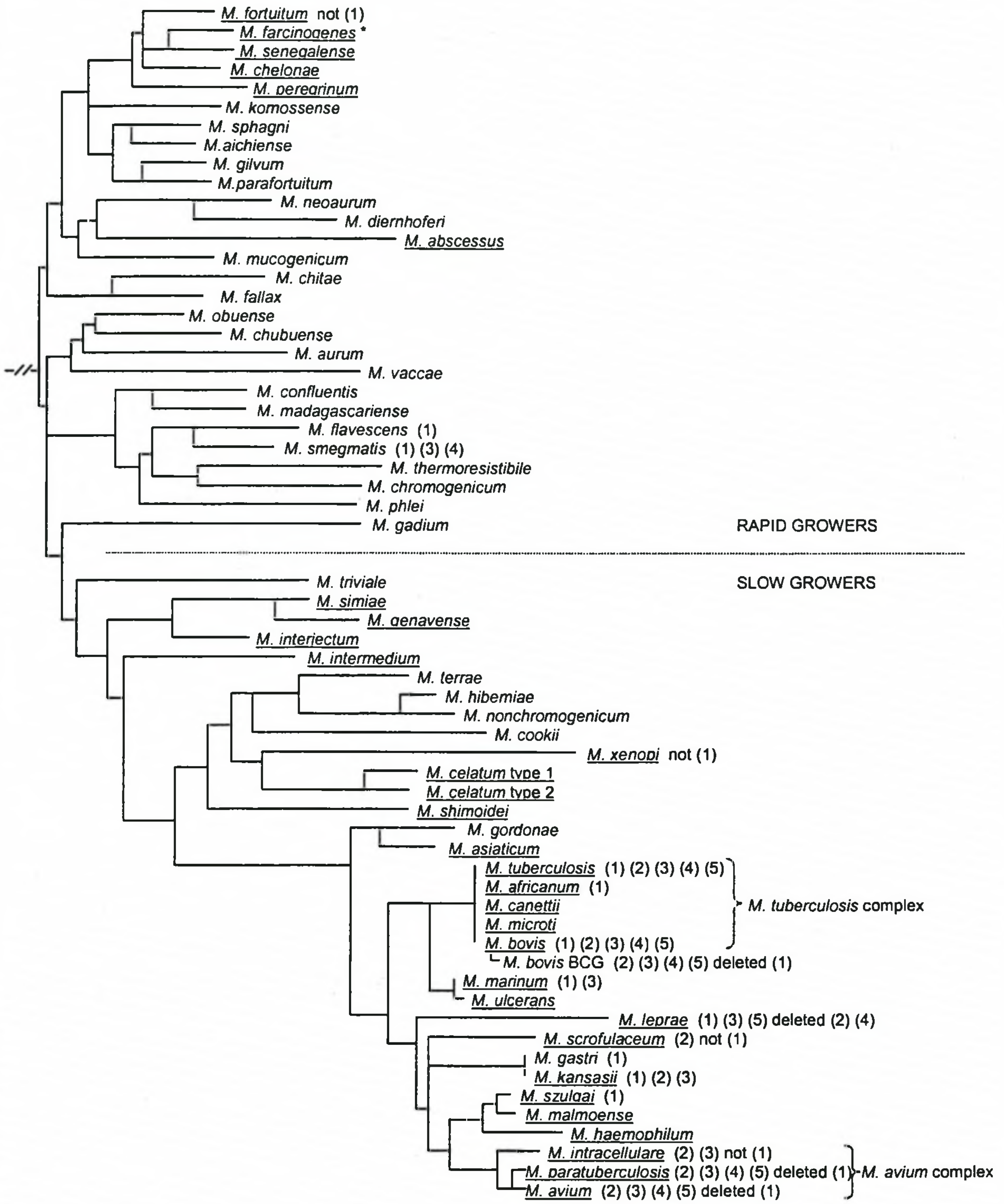
In summary, the results derived from the analysis of data from experimental approaches confirm the *in silico* results indicating the presence of different numbers of the ESAT-6 gene clusters in the genomes of members of the genus *Mycobacterium*.

Table 3A.1. Evidence for the presence of the ESAT-6 gene cluster regions in the mycobacteria*

Species and Strain	Pathogenicity [a]	ESAT-6 Gene Cluster Region [b]			
		Region 1	Region 2	Region 3	Region 5
M.africanum	Pathogenic	Present (d) (f) (g)			
M.asiaticum	Pathogenic	Not detected (d)			
M.avium	Pathogenic	Not detected (d) (f) (g) (h) (j)	Present (i)	Present (h)	Not detected (e)
M.bovis ATCC19210	Pathogenic	Present (j)			
M.bovis Branch	Pathogenic	Present (d)			
M.bovis KML	Pathogenic		Present (i)		
M.bovis MNC 27	Pathogenic	Present (g) (h)		Present (h)	
M.bovis NADL	Pathogenic	Present (d)			
M.bovis Ravenel	Pathogenic	Present (d)			
M.bovis BCG Brazil	Non-pathogenic (c)	Not detected (j)			
M.bovis BCG Connaught	Non-pathogenic (c)	Not detected (d) (j)			
M.bovis BCG Danish 1331	Non-pathogenic (c)	Not detected (f) (g) (h)		Present (h)	
M.bovis BCG Glaxo 1077	Non-pathogenic (c)	Not detected (g)			
M.bovis BCG Japan 172	Non-pathogenic (c)	Not detected (d)			
M.bovis BCG Montreal	Non-pathogenic (c)	Not detected (d)			
M.bovis BCG Moreau	Non-pathogenic (c)	Not detected (g)			
M.bovis BCG Pasteur 1173P2	Non-pathogenic (c)	Not detected (d) (g) (j)	Present (i)		
M.bovis BCG Tice	Non-pathogenic (c)	Not detected (g)			
M.bovis BCG Tokyo	Non-pathogenic (c)	Not detected (g) (h)		Present (h)	
M.bovis BCG Russia	Non-pathogenic (c)	Not detected (d) (g)			
M.chelonae	Pathogenic				Not detected (e)
M.flavescens	Non-pathogenic	Present (g)			
M.fortuitum	Pathogenic	Not detected (d) (f) (g) (h)	Not detected (i)	Not detected (h)	Not detected (e)
M.gastri	Non-pathogenic	Present (d)			
M.gordonae	Non-pathogenic		Not detected (i)		Not detected (e)
M.heamophilum	Pathogenic	Not detected (d)			
M.intracellulare	Pathogenic	Not detected (f) (g) (h)	Present (i)	Present (h)	
M.kansasii	Pathogenic	Present (d) (f) (g) (h)	Present (i)	Present (h)	
M.leprae	Pathogenic	Not detected (g)			Not detected (e)
M.malmoense	Pathogenic	Not detected (d)			
M.marinum	Pathogenic	Present (f) (g) (h)	Not detected (i)	Present (h)	
M.paratuberculosis	Pathogenic		Present (i)		
M.phlei	Non-pathogenic	Not detected (d)	Not detected (i)		
M.scrofulaceum	Pathogenic	Not detected (d) (f) (g) (h)	Present (i)	Not detected (h)	Not detected (e)
M.simiae	Pathogenic	Not detected (d)			
M.smegmatis	Non-pathogenic	Not detected (j)	Not detected (i)		Not detected (e)
M.szulgai	Pathogenic	Present (f) (g) (h)		Not detected (h)	
M.terrae	Non-pathogenic	Not detected (d)	Not detected (i)		
M.triviale	Non-pathogenic	Not detected (d)			
M.tuberculosis CSU#93	Pathogenic	Present (d)			
M.tuberculosis Erdman	Pathogenic	Present (g) (j)			Present (e)
M.tuberculosis H37Ra	Non-pathogenic (c)	Present (d) (f) (g) (j)	Present (i)		Present (e)
M.tuberculosis H37Rv	Pathogenic	Present (d) (f) (g) (h) (j)	Present (i)	Present (h)	Present (e)
M.tuberculosis R1609	Pathogenic	Present (f)			
M.tuberculosis W	Pathogenic	Present (d)			
M.ulcerans	Pathogenic	Not detected (d)			
M.vaccae	Non-pathogenic				Not detected (e)
M.xenopi	Pathogenic	Not detected (f) (g) (h)		Not detected (h)	

* Based on previously published Southern blotting, Western blotting and PCR data of selected genes and regions within the gene clusters
(a) = Shinnick and Good, 1994; (b) = No work has been done on any of the genes within region 4; (c) = Attenuated strains, potentially hazardous; (d) = Southern blotting data using *mtsa-10* (*cfp-10*) and *esat-6* as probes (Colangeli *et al.*, 2000); (e) = Southern blotting data using *mtb9.9a* as probe (Alderson *et al.*, 2000); (f) = Western blotting data using monoclonal anti-ESAT-6 antibodies (HYB 76-8) as probe (Sorenson *et al.*, 1995); (g) = Southern blotting and PCR data using *esat-6* as probe (Harboe *et al.*, 1996); (h) Southern blotting data using *tb10.4* and *cfp-10* as probes (Skjøt *et al.*, 2000); (i) = Southern blotting data using *Pan* promoter sequence as probe (Gormley *et al.*, 1997); (j) = Southern blotting data using RD1 deletion region specific probe (Mahairas *et al.*, 1996).

Figure 3A.1. Taxonomical positions of members of the genus Mycobacterium with the presence and absence of the different ESAT-6 gene cluster regions indicated next to the species. Data was obtained from experimental results (Table 3A.1) or from whole genome sequencing data (Gey van Pittius *et al.*, 2001). (1) = Region 1; (2) = Region 2; (3) = Region 3; (4) = Region 4; (5) = Region 5; not (..) = absent from the genome of this species and deleted (..) = confirmed as deleted from the genome of this species. A region has only been indicated as being absent or deleted if this has been shown by whole genome sequencing data or by experimental data from more than one publication. Underlined species are pathogens. * = *M. farcinogenes* is a slow growing mycobacterium. The taxonomical relationships between members of the genus Mycobacterium was constructed using sequence information of 16S rRNA genes as adapted from Pitulle *et al.* (1992), Shinnick and Good (1994) and Springer *et al.* (1996).


M. fortuitum not (1)
M. farcinogenes *
M. senegalense
M. chelonae
M. peregrinum
M. komossense
M. sphagni
M.aichiense
M. gilvum
M.parafortuitum
M. neoaurum
M. diernhoferi
M. abscessus
M. mucogenicum
M. chitae
M. fallax
M. obuense
M. chubuense
M. aurum
M. vaccae
M. confluentis
M. madagascariense
M. flavescens (1)
M. smegmatis (1) (3) (4)
M. thermoresistibile
M. chromogenicum
M. phlei
M. gadium
RAPID GROWERS
SLOW GROWERS
M. triviale
M. simiae
M. genavense
M. interjectum
M. intermedium
M. terrae
M. hiberniae
M. nonchromogenicum
M. cookii
M. xenopi not (1)
M. celatum type 1
M. celatum type 2
M. shimoidei
M. gordonae
M. asiaticum
M. tuberculosis (1) (2) (3) (4) (5)
M. africanum (1)
M. canettii
M. microti
M. bovis (1) (2) (3) (4) (5)
M. bovis BCG (2) (3) (4) (5) deleted (1)
M. tuberculosis complex
M. marinum (1) (3)
M. ulcerans
M. leprae (1) (3) (5) deleted (2) (4)
M. scrofulaceum (2) not (1)
M. gastri (1)
M. kansasii (1) (2) (3)
M. szulgai (1)
M. malmoense
M. haemophilum
M. intracellulare (2) (3) not (1)
M. paratuberculosis (2) (3) (4) (5) deleted (1)
M. avium (2) (3) (4) (5) deleted (1)
M. avium complex

3A.3. Individual protein families

To obtain a clearer understanding of the potential function of the ESAT-6 gene clusters, it is important to review all current information available for the individual protein families belonging to these clusters (see Chapter 3, Figure 3.1 and Table 3.3). These families are discussed in the following section according to the arbitrary alphabetical names given previously (Gey van Pittius *et al.*, 2001).

3A.3.1. Family A - Unknown

All members of this family are of unknown function. The average size of the proteins is approximately 288 aa. One member (Rv1794) contains a PS00211 ABC transporter family signature that is totally absent in the other family members. This protein has been detected in the bacterial cytoplasmic fraction with a molecular weight of 31 kDa. Another member (Rv0289) contains a motif (ALRTGTGKT) which has one mismatch to the PS00017 ATP/GTP-binding site motif A (P-loop) ([A] x4 GK [T]). This motif is found in proteins that bind ATP or GTP and is a glycine-rich region, which typically forms a flexible loop between a beta-strand and an alpha-helix. This loop interacts with one of the phosphate groups of the nucleotide and is generally referred to as the 'P-loop'. These motifs are among others found in ATP-binding proteins involved in active transport (ABC transporters). The Family A member of region 1 (Rv3866) is situated next to Rv3867 (a hypothetical gene of unknown function). The orthologue of Rv3867 is predicted to be part of the ORF of the Rv3866 gene orthologue (MT3980) in the genome sequence of *M. tuberculosis* CSU#93, and to have undergone a frameshift, which could indicate an error in the original ORF prediction and annotation of the *M. tuberculosis* H37Rv genome sequence. It is interesting to note that the gene sequence of the orthologue in *M. leprae* (ML1143A) also contains a frameshift (although it is not in the same position in the sequence). It is thus possible that this gene does not play a major role in the function of the gene cluster (highlighted by the fact that it is absent in region 4) or that the shortened version of this gene (only Rv3866) contains all the necessary domains to still be able to perform the original function.

3A.3.2. Family B - AAA$^+$ class ATPases

This family of proteins is classified under the AAA$^+$ class ATPases (COG database - Tatusov *et al.*, 2000 - maintained by the NCBI at http://www.ncbi.nlm.nih.gov/COG/). The AAA$^+$ class ATPases are a family of chaperone-like ATPases associated with the assembly, operation and disassembly of protein complexes (Neuwald *et al.*, 1999). The hexameric structure, which is often associated with members of this class, can form a hole through which DNA or RNA can be transported. The average size of the proteins from the mycobacterial Family B is approximately 608aa. They are all hydrophobic proteins and each contains one PS00017 ATP/GTP binding site motif A (P-loop) (see Family A for description of motif). Their N-termini appear to be mycobacterial specific, but their C-termini are highly similar to members of the CbxX/CfqX family, which are ATP-binding proteins of unknown function (Tekaia *et al.*, 1999). The closest known homologue is the *Bacillus subtilis* spore formation protein spoVK (also named spoVJ) which is required for spore coat formation (Fan *et al.*, 1992). One member (Rv0282) has been detected in the bacterial cytoplasmic fraction at a molecular size of between 65 kDa and 80 kDa.

3A.3.3. Family C - Unknown

All members of this family are of unknown function. The average size of the proteins is approximately 497 aa. There is a hydrophobic stretch near the N-terminus of all members corresponding to an N-terminal transmembrane region (Figure 3A.2). One member (Rv3450c) contains a binding-protein-dependent transport systems inner membrane component signature (with a single amino acid mismatch). Another family member (Rv0283) contains one PS00017 ATP/GTP binding site motif A (P-loop) (that is totally absent from all other members) and has recently been identified to be a secreted protein containing an N-terminal secretion signal (Wiker *et al.*, 2000). The authors made use of a *phoA* fusion library and identified an alternative starting codon on this protein other than the annotated one so that the N-terminal transmembrane region moved sufficiently close to the N-terminus to be predicted as a signal peptide. It is possible that all the members of this family might have an alternative start site, resulting in the N-terminal transmembrane region to change to a signal peptide. It is clear, though, from the multiple sequence alignment (Figure 3A.3) that the originally predicted N-terminal region YRRGFVTRHQVTGWRFVM**RR**IAAGIA (that was deleted by the authors of this paper) is a region of relatively high sequence homology between the proteins in this family. If this proposed alternative start site holds true, the significance of this homologous stretch of residues before the start codon is unknown, but it could indicate the presence of a conserved regulatory region upstream of the genes.

Figure 3A.2. TMHMM profile for Rv0283. This result is representative of all members of Family C and shows clearly the single N-terminal transmembrane region.

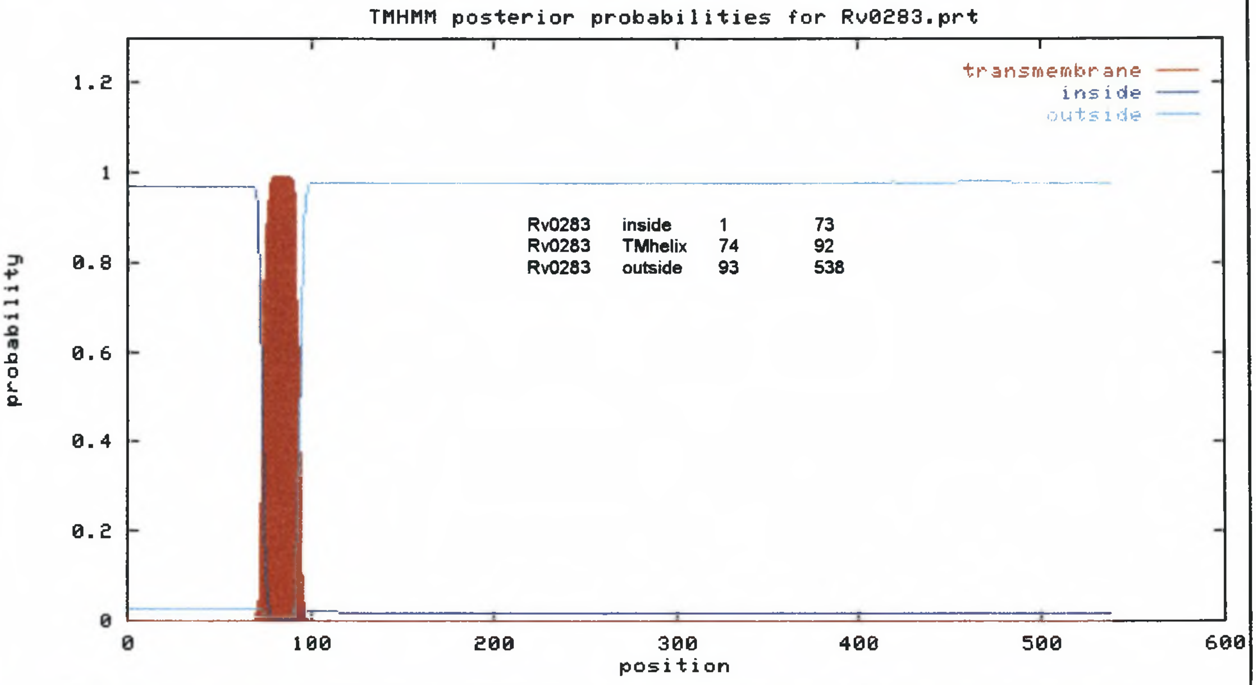


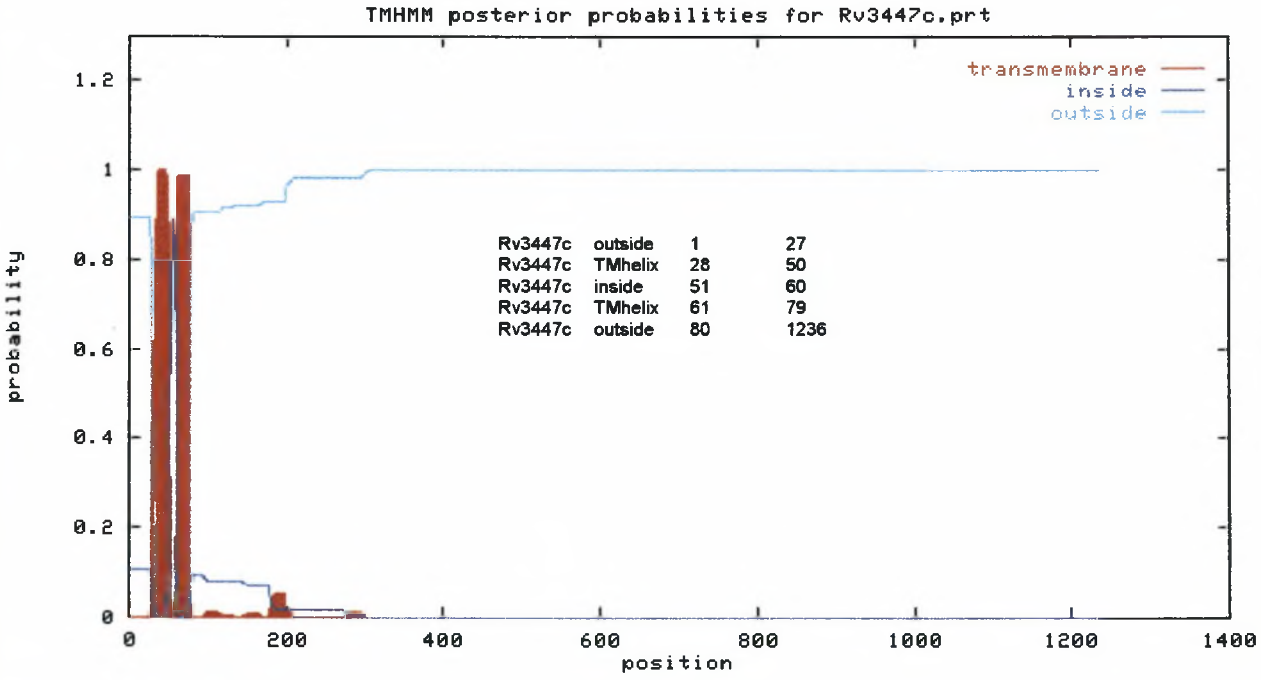
Figure 3A.3. Partial sequence alignment of N-terminal part of Family C protein sequences. Arrow indicates new experimentally defined starting amino acid.

```
                                                                             ↓
S.coelicolor       ----------------------------------MASRRDELNAYTFAKRRLLAAFL QPS
M.tb region 4      -------------------------------VPSPATTWLHVSGYRFLLRRIECALL FGD
M.tb region 1      -----------------------------MGLR--LTTKVQVSGWRFLLRRLEHAIV RRD
M.tb region 3      MTNQQHDHDFDHDRRSFASRTPVNNNPDKVVYRRGFVTRHQVTGWRFVMRRIAAGIA LHD
M.tb region 2      -------------------------------MPLSLSNRDQNSGHLFYNRRLRAATT RFS
M.tb region 5      -------------------VAEESRGQRGSGYGLGLSTRTQVTGYQFLARRTAMALT RWR
                                                                 *  **

S.coelicolor       PSGTEEGAPKPLRTVVPSLVAGALTLAVFGAWGMFQPTAPSGWDEPGARVIVGKQSTTRY
M.tb region 4      VCAATGALRARTTSLALGCVLAIVAAMGCAFVALLRPQSALGQAP-----IVMGRESGAL
M.tb region 1      TRMFDDPLQFYSRSIALGIVVAVLILAGAALLAYFKPQGKLGGTS-----LFTDRATNQL
M.tb region 3      TRMLVDPLRTQSRAVLMGVLIVITGLIGSFVFSLIRPNGQAGSNA-----VLADRSTAAL
M.tb region 2      VRMKHDD-RKQTAALALSMVLVAIAAGWMMLLNVLKPTGIVGDSA-----IIGDRDSGAL
M.tb region 5      VRMEIEPGRRQTLAVVASVSAALVICLGALLWSFISPSGQLNESP-----IIADRDSGAL
                                                       *

S.coelicolor       VVLKTDGDTRLHPVLNLASARLLMKDGTYEVVQVGDDVLDSGEIPRGPILGIPYAPDRLP
M.tb region 4      YVRVDD---VWHPVLNLASARLIAAT-NANPQPVSESELG--HTKRGPLLGIPGAPQLLD
M.tb region 1      YVLLSG---QLHPVYNLTSARLVLGN-PANPATVKSSELS--KLPMGQTVGIPGAPYATP
M.tb region 3      YVRVGE---QLHPVLNLTSARLIVGR-PVSPTTVKSTELD--QFPRGNLIGIPGAPERMV
M.tb region 2      YARIDG---RLYPALNLTSARLATGT-AGQPTWVKPAEIA--KYPTGPLVGIPGAPAAMP
M.tb region 5      YVRVGD---RLYPALNLASARLITGR-PDNPHLVRSSQIA--TMPRGPLVGIPGAPSSFS
                               *  ** ****                        *   *** **
```

3A.3.4. Family D - ATPases

This family of proteins is classified under the DNA segregation ATPase FtsK/SpoIIIE and related proteins (COG database - Tatusov *et al.*, 2000 - maintained by the NCBI at http://www.ncbi.nlm.nih.gov/COG/). The average size of the proteins is approximately 1333aa. They all contain two N-terminal transmembrane regions (Figure 3A.4) as well as three PS00017 ATP/GTP binding site motif A (P-loop) in the C-terminus that are separated by ~350 and 230 residues (Tekaia *et al.*, 1999). The closest known homologues are the cell division protein FtsK, the DNA translocase stage III sporulation protein spoIIIE and the *Bacillus subtilus* yukA (or yueA). These are all membrane-associated proteins containing hydrophobic N-terminal transmembrane regions connected to a highly homologous C-terminal region that is situated in the cytoplasm (Wu *et al.*, 1997, Wang *et al.*, 1998) and contains the ATP binding sites responsible for ATP hydrolyses to drive translocation (of DNA in this case).

Figure 3A.4. TMHMM profile for Rv3447c. This result is representative of all members of Family D and shows clearly the two N-terminal transmembrane regions.



Rv3447c	outside	1	27
Rv3447c	TMhelix	28	50
Rv3447c	inside	51	60
Rv3447c	TMhelix	61	79
Rv3447c	outside	80	1236

3A.3.5. Family E - PE

The PE family is a large family of 99 duplicated Gly-Ala-rich proteins of variable sequences found distributed throughout the genome of *M. tuberculosis*. All these proteins contain a conserved N-terminal segment of ~110 aa with the motif Pro-Glu (PE) in positions 8 - 9 in most cases (Tekaia *et al.*, 1999). The average size of the PE proteins found in the ESAT-6 gene clusters is approximately 95 aa, although the sizes of other family members vary greatly.

3A.3.6. Family F - PPE

The PPE family is a large family of 67 duplicated Gly-Ala-rich proteins of variable sequences found distributed throughout the genome of *M. tuberculosis*. All these proteins contain a conserved N-terminal segment of ~180 aa containing the motif Pro-Pro-Glu (PPE) (Tekaia *et al.*, 1999). The average size of the PPE proteins found in the ESAT-6 gene clusters is approximately 398 aa, although the sizes of other family members vary greatly.

3A.3.7. Family G - lhp (CFP-10) and Family H - ESAT-6

Family G and Family H are both families of small (~10 kDa), very potent T-cell antigens of unknown function, secreted from early growth without any ordinary secretion signals (Andersen *et al.*, 1995, Sørensen *et al.*, 1995, Berthet *et al.*, 1998, Van Pinxteren *et al.*, 2000). In addition to the five duplications in the ESAT-6 gene clusters, there are six other sub-duplications of only the *lhp* and *esat-6* genes in the genome of *M. tuberculosis* H37Rv (see Chapter 3, Figure 3.2). As these are mostly flanked on one or both sides with PE, PPE and/or an insertion sequence (IS), it signifies either a propensity to co-duplicate with PE and PPE, or a susceptibility to IS-mediated transfer (Tekaia *et al.*, 1999).

One member of Family G, named the "L45 homologous protein" or *lhp* (Rv3874), encode the "culture filtrate protein-10" (CFP-10 or renamed "*M. tuberculosis*-specific antigen-10", MTSA-10 by Colangeli *et al.*, 2000) and was shown to form part of an operon with a member of Family H named the "6kDa early-secreted antigenic target" (ESAT-6 or Rv3875, Berthet *et al.*, 1998). The genes of all the other members of Family G are also situated directly next to a member of Family H, suggesting that these genes may all form part of operon structures.

Because of the sequence similarity observed between members of these two families, as well as the fact that their genes are always situated directly adjacent to each other, it is hypothesized that they have arisen through gene duplication. Thus, the 23 genes from Families G and H are collectively classified under the combined ESAT-6 family (Berthet *et al.*, 1998, Skjøt *et al.*, 2001). A list of the family members and alternative names are provided in Table 3A.2.

There is evidence that at least three ESAT-6 family members (Rv1038c or TB11, Rv3875 or ESAT-6 and Rv0288 or TB10.4) are N-terminally cleaved (Skjøt *et al.*, 2000, Peter Andersen and Karin Weldingh, personal communication), indicating a possible function for the mycosin proteases (Family K) present within the ESAT-6 gene clusters.

Table 3A.2. Members of the ESAT-6 family (Subfamily G)

No.	ORF no.	Region	Gene Name	Protein name	*M. tb* H37Rv	*M. tb* CSU#93	*M. bovis*	*M. leprae*	*M. avium*	*M. paratb*	*M. smegmatis*	Reference
1	Rv3445c	region 4			Yes	Yes	Yes	Deleted	Yes	Yes	Yes	
2	Rv3874	region 1	*lhp, mtsa-10, mtb11 or ORF-6*	LHP, CFP-10, Mtb11 or MTSA-10	Yes	Yes	Yes	Yes	Deleted	Deleted	Yes	Berthet *et al.*, 1998; Alderson *et al.*, 2000, Dillon *et al.*, 2000, Colangeli *et al.*, 2000; Mustafa, 2001
3	Rv3905c	---			Yes	Yes	Yes	?	Yes	Yes	Yes	
4	Contig565	---			No	No	No	?	Yes	Yes	?	
5	Rv0287	region 3	*tb9.8*	TB9.8	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Skjøt *et al.*, 2001
6	Rv3020c	---			Yes	Yes	Yes	?	?	?	?	
7	Rv3891c	region 2			Yes	Yes	Yes	Deleted	Yes	Yes	?	
8	Rv1792	region 5			Yes	Yes	Yes	Stop + frame	Yes	Yes	?	
9	Rv2347c	---			Yes	Yes	Deleted (RD7)	?	?	?	?	
10	Rv1197	---		u1756c	Yes	Yes	Yes	Yes	?	?	?	Genbank accession number U15180
11	MLCB1701.07c			MLCB1701.07c	No	No	No	Yes	?	?	?	Genbank accession number AL049191
12	Rv1038c	---	*tb11.0*	TB11.0	Yes	Yes	Yes	Stop + frame	?	?	?	Rosenkrands *et al.*, 2000a, 2000b
13	Rv3620c	---			Yes	Yes	Deleted (RD9)	?	?	?	?	
14	MT2421	---			Deleted	Yes	Deleted	?	?	?	?	

Note: (a) The name "L45 homologous protein" (*lhp*) is in fact a misnomer, as the L45 gene of *M. leprae* is a Family H (ESAT-6) orthologue (see Genbank accession no X90946 and family H below) and the authors actually described the gene lying adjacent to the L45 gene (the Family G CFP-10 orthologue). The confusion in the naming of the *lhp* gene was probably the result from the fact that the clone containing L45 also contains the *M. leprae* orthologues of Rv3874 and Rv3876.

(b) Rv1792 contains an in-frame stopcodon and is probably a pseudogene.

(c) Rv3020c has been incorrectly classified as a member of the PE family of proteins in the original annotation of the *M. tuberculosis* H37Rv genome (Cole *et al.*, 1998).

(d) *M. avium* contains one extra duplicate of Family G not found in the other mycobacteria. This gene is situated on contig 565 and seems to be a duplication of the *M. avium* Rv3905c orthologue.

(e) MLCB1701.07c is an *M. leprae* region 5 duplicate and do not have any homologue on the *M. tuberculosis* genome sequence. It seems to be a recent duplication of the Rv1197 orthologue into a region a few thousand bases downstream of Rv1354.

(f) *M. tuberculosis* CSU#93 contains one more copy of the QILSS subfamily designated MT2421. This gene has been knocked out in the genome of *M. tuberculosis* H37Rv by the insertion of an IS 6110 transposon.

Table 3A.2. Continued - Members of the ESAT-6 family (Subfamily H)

No	ORF no.	Region	Gene Name	Protein name	*M. tb*	*M. tb*	*M. bovis*	*M. leprae*	*M. avium*	*M. paratb*	*M. smegmatis*	Reference
.					H37Rv	CSU#93						
1	Rv3444c	region 4			Yes	Yes	Yes	Deleted	Yes	Yes	Yes	
2	Rv3875	region 1	*esx, esat-6, L45, ORF-7, ORF1C*	ESAT-6 or L-ESAT	Yes	Yes	Yes	Yes	Deleted	Deleted	Yes	Andersen *et al.*, 1995, Genbank accession number X79562, Ahmad *et al.*, 1999, Genbank accession no X90946, Mahairas *et al.*, 1996
3	Rv3904c	—			Yes	Yes	Yes	?	Yes	Yes	Yes	
4	Contig565	—			No	No	No	?	Yes	Yes	?	
5	Rv0288	region 3	*tb10.4 or cfp-7*	TB10.4 or CFP-7	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Skjøt *et al.*, 2000, Genbank accession no. AJ002067,
6	Rv3019c	—	*tb10.3*	TB10.3	Yes	Yes	Yes	?	?	?	?	Skjøt *et al.*, 2001
7	Rv3017c	—	*tb12.9*	TB12.9	Yes	Yes	Yes	?	?	?	?	Skjøt *et al.*, 2001
8	Rv3890c	region 2			Yes	Yes	Yes	Deleted	Yes	Yes	?	
9	Rv1793	region 5	*mtb9.9a*	Mtb9.9A	Yes	Yes	Yes	Stop + frame	Yes	Yes	?	Alderson *et al.*, 2000
10	Rv2346c	—	*mtb9.9e*	Mtb9.9E	Yes	Yes	Partly deleted (RD7)	?	?	?	?	Alderson *et al.*, 2000
11	Rv1198	—	*mtb9.9c*	Mtb9.9C or u1756d	Yes	Yes	Yes	Yes	?	?	?	Alderson *et al.*, 2000, Genbank accession number U15180 and AAA62902
12	MLCB1701.06c			MLCB1701.06c	No	No	No	Yes	?	?	?	Genbank accession number AL049191
13	Rv1037c	—	*mtb9.9d*	Mtb9.9D	Yes	Yes	Yes	Stop + frame	?	?	?	Alderson *et al.*, 2000
14	Rv3619c	—	*mtb9.9b*	Mtb9.9B	Yes	Yes	Deleted (RD9)	?	?	?	?	Alderson *et al.*, 2000
15	MT2420	—			Deleted	Yes	Deleted	?	?	?	?	

Note: (a) *M. avium* contains one extra duplicate of Family G not found in the other mycobacteria. This gene is situated on contig 565 and seems to be a duplication of the *M. avium* Rv3904c orthologue. (b) MLCB1701.07c is an *M. leprae* region 5 duplicate and do not have any homologue on the *M. tuberculosis* genome sequence. It seems to be a recent duplication of the Rv1198 orthologue into a region a few thousand bases downstream of Rv1354. (c) *M. tuberculosis* CSU#93 contains one more copy of the Mtb9.9 subfamily designated MT2420. This gene has been knocked out in the genome of *M. tuberculosis* H37Rv by the insertion of an IS 6110 transposon.

Although the 23 members of the combined ESAT-6 family show very low protein sequence homology (< 22%), most of them contain one conserved primary sequence feature. This sequence feature is a [W]-x-[G] motif shared by 18 of the 23 members (with the other 5 containing slight variations thereof), situated at positions 43 and 45 (Rv3875 numbering, Figure 3A.5). The region surrounding these conserved tryptophan (W) and glycine (G) residues is also moderately conserved between the proteins. Conserved aromatic residues like tryptophan have been shown previously to play a very important role in protein activity. One example of this is the drastic decrease in activity that was observed with the removal of a single residue of tryptophan from mesentericin Y105 (Montville and Chen, 1998).

Figure 3A.5. Partial sequence alignment of the members of the ESAT-6 protein family. Amino acid residues in bold indicates potential family-specific motif [W]-x-[G].

Rv3619c	LTASDF--**WGG**-AGSAACQGFITQLGRNFQVIYEQA
Rv1037c	LTASDF--**WGG**-AGSAACQGFITQLGRNFQVIYEQA
Rv1198	LTASDF--**WGG**-AGSAACQGFITQLGRNFQVIYEQA
Rv2346c	LAAGDF--**WGG**-AGSVACQEFITQLGRNFQVIYEQA
Rv1793	LAAGDF--**WGG**-AGSVACQEFITQLGRNFQVIYEQA
Rv0288	AALQSA--**WQG**-DTGITYQAWQAQWNQAMEDLVRAY
Rv3019c	AVLSSA--**WQG**-DTGITYQGWQTQWNQALEDLVRAY
Rv3017c	TAPSRA--**CQG**-DLGMSHQDWQAQWNQAMEALARAY
Rv0287	MSAQAF--**HQG**-ESSAAFQAAHARFVAAAAKVNTLL
Rv3020c	MSAQAF--**HQG**-ESAAAFQGAHARFVAAAAKVNTLL
Rv1038c	QNISGAG-**WSG**-MAEATSLDTMTQMNQAFRNIVNML
Rv1792	QNISGAG-**WSG**-MAEATSLDTMT-MNQAFRNIVNML
Rv3620c	QNISGAG-**WSG**-MAEATSLDTMTQMNQAFRNIVNML
Rv1197	QNISGAG-**WSG**-MAEATSLDTMAQMNQAFRNIVNML
Rv2347c	QNISGAG-**WSG**-MAEATSLDTMAQMNQAFRNIVNML
Rv3874	GSLQGQ--**WRG**-AAGTAAQAAVVRFQEAANKQKQEL
Rv3890c	NALQEF--**FAG**-HGAQGFFDAQAQMLSGLQGLIETV
Rv3891c	NVMNPAT-**WSG**-TGVVASHMTATEITNELNKVLTGG
Rv3905c	GQMLGG--**WRG**-ASGSAYGSAWELWHRGAGEVQLGL
Rv3904c	TRLHVT--**WTG**-EGAAAHAEAQRHWAAGEAMMRQAL
Esat6	TKLAAA--**WGG**-SGSEAYQGVQQKWDATATELNNAL
Rv3444c	APLQQL--**WTR**-EAAAAYHAEQLKWHQAASALNEIL
Rv3445c	SGVPPSV-**WGG**-LAAARFQDVVDRWNAESTRLYHVL

Five of the duplications from both Family G and H are almost identical (suggestive of recent duplication events) and have arisen exclusively from duplications of the region 5 copies (Gey van Pittius *et al.*, 2001, Skjøt *et al*, 2001). These two subfamilies have been classified as the QILSS subfamily (for the members of family G, Cole *et al.*, 1998) and the MTIN or immunodominant Mtb9.9 subfamily (for the members of family H, see Figure 3A.6, Cole *et al.*, 1998, Alderson *et al.*, 2000).

Alderson *et al.* (2000) only identified 4 members of the Mtb9.9 subfamily present in *M. tuberculosis* strain Erdman (designated *mtb9.9a, b, c* and *d*). They further identified one more family member in the *M. tuberculosis* H37Rv genome sequence and named it *mtb9.9e* and also noted that the *mtb9.9b* gene identified in Erdman was not present on the H37Rv genome sequence. *M. tuberculosis* H37Rv actually have 5 members of this subfamily, with the members Rv1037 and Rv3619c being identical and having the same sequence as the *mtb9.9d* gene identified in Erdman. There are two possibilities to explain the fact that Alderson and coworkers saw another member of this family (*mtb9.9b*) and did not detect the second copy of *mtb9.9d* (Rv1037 or Rv3619c). Either the Erdman genome actually does contain one extra duplicate of this family (*mtb9.9b*) and the experimental methods they used could not distinguish between Rv1037 and Rv3619c, or one of the genes Rv1073 or Rv3619c have undergone changes in Erdman to result in the formation of *mtb9.9b*. A possible answer for this can be found in the genome sequence of *M. tuberculosis* CSU#93. The sequence for the Rv3619c orthologue found in CSU#93 (MT3721) has exactly the same sequence as the *M. tuberculosis* Erdman *mtb9.9b* identified by Alderson and coworkers as being novel to Erdman (Figure 3A.6). It is thus very likely that this is the same gene in all three *M. tuberculosis* strains, which has undergone changes in H37Rv.

Figure 3A.6. Multiple protein sequence alignment of the Mtb9.9 subfamily of the ESAT-6 protein family. The alignment indicates the extremely high level of amino acid sequence conservation between family members. Homologous residues are represented by a dot (.) and differences are indicated by the change in residue. Orthologs are grouped together.

```
M.tb H37Rv  Rv1793     MTINYQFG DVDAHGAMIR AQAASLEAEH QAIVRDVLAA GDFWGGAGSV
M.tb CSU93  MT1842     ........ .......... .......... .......... ..........
M.bovis     Rv1793     ........ .......... .......... .......... ..........
M.avium     Rv1793     .S...... .......L.. .......... ...I...... ..........

M.tb H37Rv  Rv1198     ........ .......... ...GL..... ...I...T.  S........A
M.tb CSU93  MT1236     ........ .......... ...GL..... .......... .........A
M.bovis     Rv1198     ........ ...D...... ...GL..... ...I...T.  S........A

M.tb H37Rv  Rv2346c    ........ .......... ...GL..... .......... ..........
M.tb CSU93  MT2411     ........ .......... ...GL..... .......... ..........
M.bovis     Rv2346c                                          ... ..........

M.tb CSU93  MT2420     ........ .......... ....A..... .......... ..........

M.tb H37Rv  Rv1037c    ........ .......... ...G...... ...IS...T. S........A
M.tb CSU93  MT1066     ........ .......... ...G...... ...IS...T. S........A
M.bovis     Rv1037c    ........ .......... .L.G...... ...IS...T. S........A

M.tb H37Rv  Rv3619c    ........ .......... ...G...... ...IS...T. S........A
M.tb CSU93  MT3721     ........ .......... .L.GL..... ...IS...T. S........A
M.tb Erdman mtb9.9b    ........ .......... .L.GL..... ...IS...T. S........A

M.leprae                 ...... EI.....A.. ....A..TT. ...LAT.RD. AE....Q..T


M.tb H37Rv  Rv1793     ACQEFITQLG RNFQVIYEQA NAHGQKVQAA GNNMAQTDSA VGSSWA
M.tb CSU93  MT1842     .......... .......... .......... .......... ......
M.bovis     Rv1793     .......... .......... .......... .......... ......
M.avium     Rv1793     .......... .......... ........T. .S...S.... ......

M.tb H37Rv  Rv1198     ...G...... .......... .......... .......... ......
M.tb CSU93  MT1236     ...G...... .......... .......... .......... ......
M.bovis     Rv1198     ...G...... .......... .......... .......... ......

M.tb H37Rv  Rv2346c    .......... .......... .......... .......... ......
M.tb CSU93  MT2411     .......... .......... .......... .......... ......
M.bovis     Rv2346c    .......... .......... .......... .......... ......

M.tb CSU93  MT2420     .......A.. ...A...Q.. ......I... .S........ ......

M.tb H37Rv  Rv1037c    ...G...... .......... .......... .......... ......
M.tb CSU93  MT1066     ...G...... .......... .......... .......... ......
M.bovis     Rv1037c    ...G...... .......... .......... .......... ......

M.tb H37Rv  Rv3619c    ...G...... .......... .......... .......... ......
M.tb CSU93  MT3721     ...G...... .......... .......... .......... ......
M.tb Erdman mtb9.9b    ...G...... .......... .......... .......... ......

M.leprae               .HEM..AD.. ....M..... .S......R. SSS..D..RS .S.A.S
```

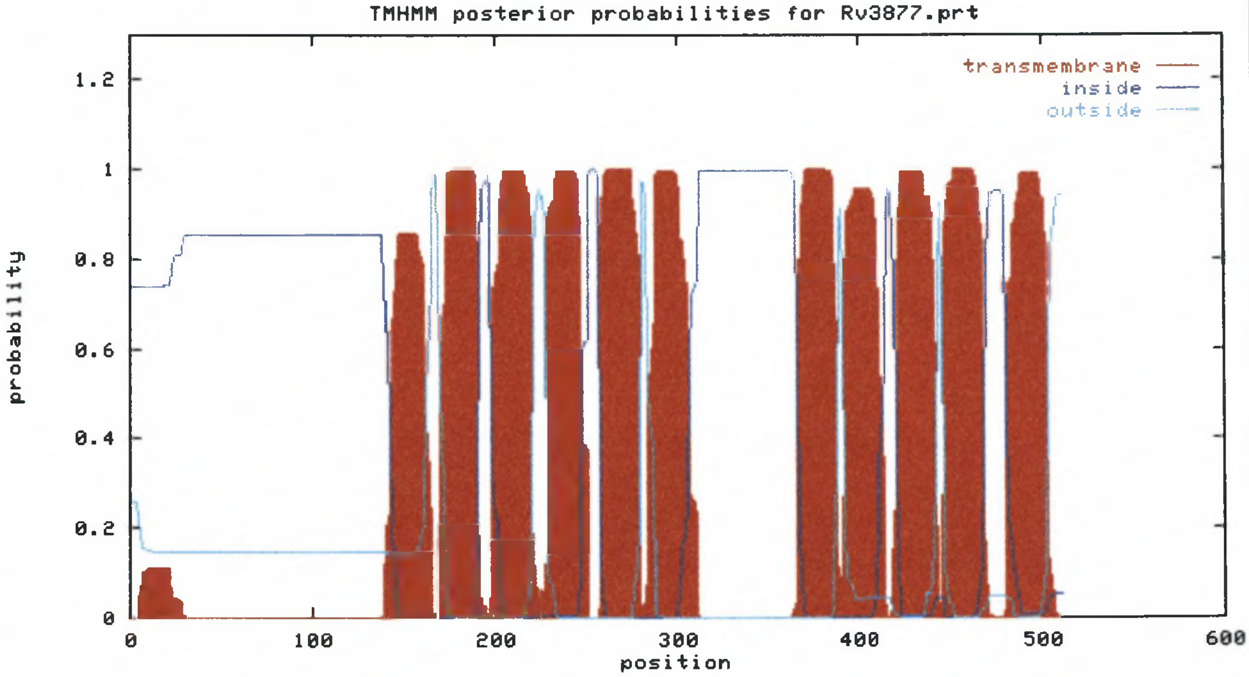
3A.3.8. Family I - ATPases involved in chromosome partitioning

This two-member family is classified under the "ATPases involved in chromosome partitioning" (COG database - Tatusov *et al.*, 2000 - maintained by the NCBI at http://www.ncbi.nlm.nih.gov/COG/). One gene (Rv3876) has a highly repetitive and proline-rich N-terminus and has a length of 666aa. Only the C-terminus of this protein shares homology to the other protein in the family, (Rv3888c), which is 341aa in length. This protein contains a transmembrane region inside the shared homology area, which only shows very weak transmembrane potential in Rv3876 in the corresponding position. These two proteins also share similarity to three other hypothetical *M. tuberculosis* proteins Rv0530, Rv2787 and Rv3860, which are not found situated in the ESAT-6 gene clusters but probably form part of this family. Two of these proteins have easily identifiable PS00017 ATP/GTP binding site motifs (Rv3860 and Rv2787). Rv3876 and Rv3888c have a shared region of very high homology, which corresponds to the ATP/GTP binding motifs of the other two proteins. This region contains just two amino acid changes, only one of which does not correspond to the classical ATP/GTP binding motif [(AG)-X-X-X-X-G-K-(ST)] so that it could probably still act as a motif for ATP/GTP binding in these proteins. (ATP/GTP binding motif in Rv2787 = VS**AKGGVGKT**TM, conserved corresponding region in Rv3888c V**SGKGGVGVT**TM).

3A.3.9. Family J - Putative transporters

A family of integral inner-membrane proteins containing 11 transmembrane regions (Figure 3A.7) and sharing weak similarity to known transporters. The average size of the proteins is approximately 492aa. One member (Rv3448) contains a PS00402 binding protein dependent transport systems inner membrane component signature, which seemed to have undergone some divergence in the sequences of the other members. Rv3877 contains a predicted signal peptide sequence, while Rv3448 and Rv1795 have weak N-terminal hydrophobic domains that may also act as membrane signals. Rv1795 is the only member that contains only 10 transmembrane regions as opposed to 11 in the other family members. Two members of this family (Rv3448 and Rv3887c) were shown to be deleted in certain clinical strains of *M. tuberculosis* (Kato-Maeda *et al.*, 2001).

Figure 3A.7. TMHMM profile for Rv3877. This result is representative of all members of Family J and shows clearly the eleven transmembrane regions.

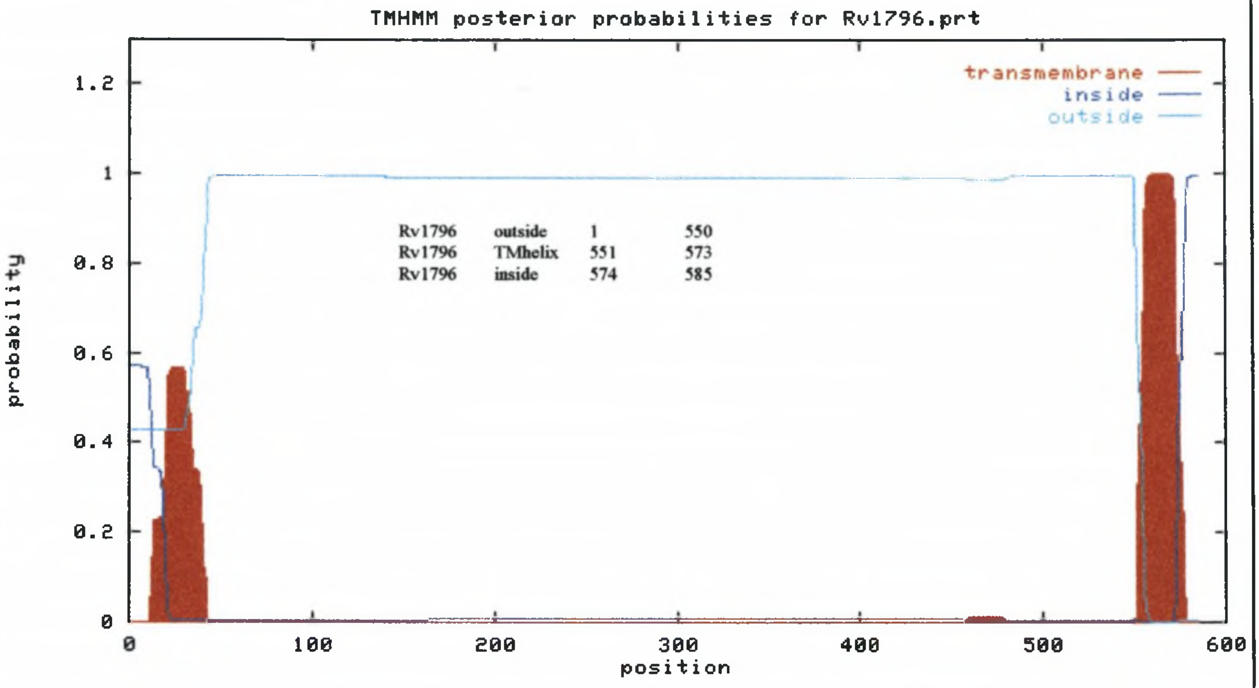


Rv3877	inside	1	142	Rv3877	TMhelix	143	161	Rv3877	outside	162	172
Rv3877	TMhelix	173	191	Rv3877	inside	192	197	Rv3877	TMhelix	198	220
Rv3877	outside	221	228	Rv3877	TMhelix	229	247	Rv3877	inside	248	257
Rv3877	TMhelix	258	280	Rv3877	outside	281	283	Rv3877	TMhelix	284	306
Rv3877	inside	307	365	Rv3877	TMhelix	366	388	Rv3877	outside	389	391
Rv3877	TMhelix	392	414	Rv3877	inside	415	420	Rv3877	TMhelix	421	442
Rv3877	outside	443	445	Rv3877	TMhelix	446	468	Rv3877	inside	469	480
Rv3877	TMhelix	481	503	Rv3877	outside	504	511				

3A.3.10. Family K - Mycosins - subtilisin-like serine proteases

Family K is a family of secreted, membrane anchored, cell wall associated subtilisin-like serine proteases - the mycosins (Brown *et al.*, 2000). The average size of the proteins is approximately 499 aa. All members contain an N-terminal signal sequence with a signal peptidase I cleavage site as well as a C-terminal hydrophobic domain, followed by a short positively charged segment, that could act as a transmembrane anchor (Figure 3A.8). All the mycosins contain the subtilase conserved active site residues Asp-His-Ser (see Chapter 2, Figure 2.2). One member of this family (Rv1796) was shown to elicit delayed-type hypersensitivity reactions only in guinea pigs immunized with live mycobacteria (Romain *et al.*, 1993), indicating that these proteins may be able to be shed from the cell wall surface.

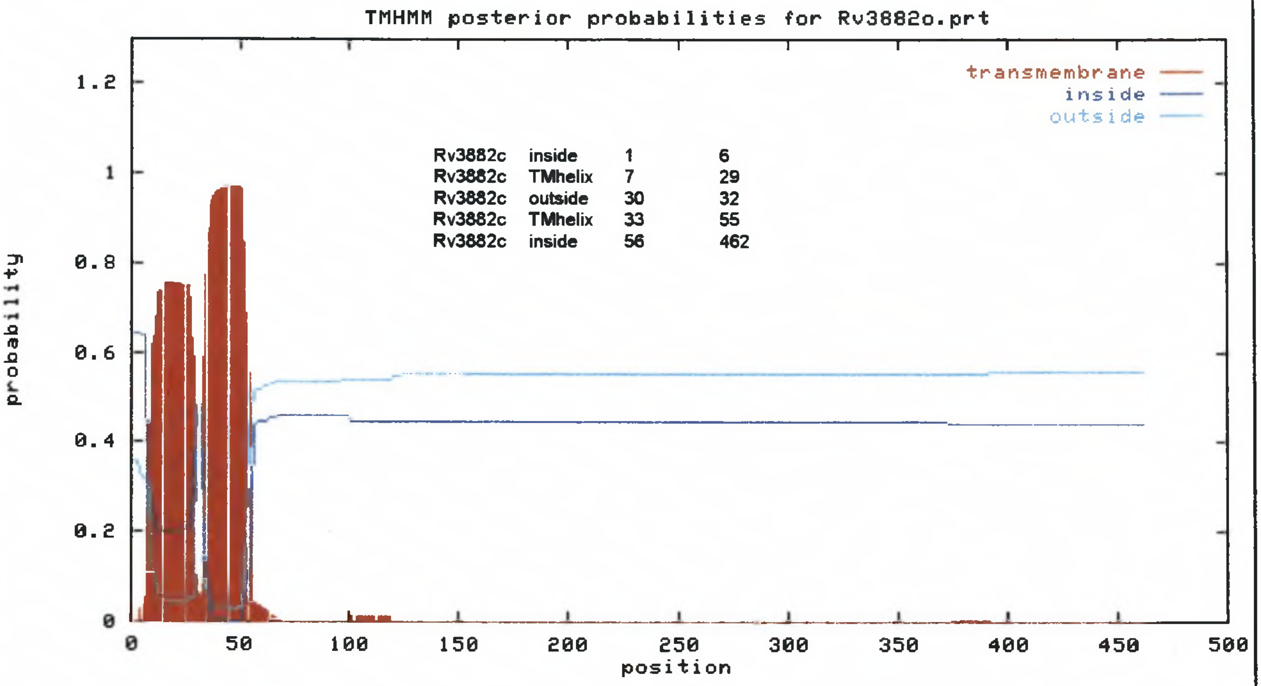
Figure 3A.8. TMHMM profile for Rv1796. This result is representative of all members of Family K and shows clearly the N-terminal signal peptide as well as the C-terminal transmembrane anchor.



3A.3.11. Family L - Unknown

A family of proteins of unknown function containing two N-terminal transmembrane regions, the first one of which is predicted to be a signal peptide sequence (Figure 3A.9). The average size of the proteins is approximately 434 aa. The first transmembrane region is situated very close to the second region so that it might be more likely to be a signal anchor. It is known that there is a problem with poor discrimination between signal peptides and uncleaved signal anchors when using computer prediction systems (Nielsen *et al.*, 1998).

Figure 3A.9. TMHMM profile for Rv3882c. This result is representative of all members of Family L and shows clearly the two N-terminal transmembrane regions.



Rv3882c	inside	1	6
Rv3882c	TMhelix	7	29
Rv3882c	outside	30	32
Rv3882c	TMhelix	33	55
Rv3882c	inside	56	462

3A.4. Immunological aspects of the ESAT-6 gene cluster

It is a well-known contradiction in terms that live mycobacteria generate a much more efficient protective immunity than killed bacteria (Andersen, 1997), but that both sensitize animals for a DTH reaction. This has been linked to the fact that live bacteria secrete peptides early in infection that are needed to recruit protective T-cells. The key antigenic determinants of the culture filtrate are smaller than the 10 kDa fraction (Boesen et al., 1995) and we now know that this fraction contains multiple copies of the small immunodominant proteins belonging to the ESAT-6 and CFP-10 families (Sørensen *et al.*, 1995, Berthet *et al.*, 1998, Alderson *et al.*, 2000). ESAT-6 elicits a high level of interferon-gamma from memory effector cells during the first phase of a protective immune response and this is important because mycobacterial diseases are generally characterized by strong Th1 responses and high levels of interferon-gamma (Andersen, 1997). At least eight of the twenty-three members of the greater ESAT-6 family (including the CFP-10 family) have already been shown experimentally to be immunologically relevant (Skjøt *et al.*, 2001). However, it is not known what advantage the bacterium obtains from the secretion of the multiple immunodominant copies of the ESAT-6 and the CFP-10 proteins. It may be speculated that spread of disease could be propagated by inducing a massive host immune response, resulting in intense inflammation, tissue destruction, caseous necrosis and the formation of cavitory lesions resulting in release of the bacteria from the damaged airways of susceptible individuals unable to contain the infection. In a comparison between the immunopathologies of schistosomiasis and tuberculosis, Doenhoff (1998) showed how in both diseases extensive immune-dependent granulomatous inflammation is an important step in facilitating the right conditions for efficient transmission of trapped infective bacteria. Transmission of *M. tuberculosis* depends almost entirely upon the bacteria being able to be released into the airways of an infected individual's lungs. It is well-known that during leprosy disease there is a massive destruction of the infected Schwann cells of the nervous system, mediated by the host's own immune system in response to *M. leprae* antigens, and that much of the pathology seen during tuberculosis infection is also the result of excessive host-mediated cellular immune and inflammatory responses against *M. tuberculosis* antigens (Brosch *et al.*, 2000b). This is confirmed by a recent report by Dannenberg and Collins (2001) in which they showed extensively that progressive pulmonary tuberculosis is not due to increasing numbers of viable bacteria, but rather to the continuous host response against constantly released mycobacterial products.

On the other hand, in resistant healthy individuals T-cell migration and activation would take place in response to the potent antigen release, with granuloma formation and subsequent control of the progression of disease, causing a long-term latent infection. The bacteria that survive in a host that controls the infection are potentially placed into an important growth phase whereby they are held in stasis for many years. This static phase may continue until a time point when the host's immune system is down-regulated (for example during an immunosuppressing disease like AIDS) and the reactivation and continued multiplication of the bacteria result in tuberculosis (Andersen, 1997). The clinical isolate CDC1551 (CSU#93 or Oshkosh strain) that was reported to be hypervirulent and having an unusually high rate of transmission (Valway *et al.*, 1998), was shown to be in fact less virulent than other clinical isolates (North *et al.*, 1999), but proved to be more transmissible (Bishai *et al.*, 1999). In addition to this, CDC1551 induced an earlier and more vigorous host response, causing earlier control of the growth in lungs and the establishment of chronic stable infection (Manca *et al.*, 1999). This is critical for the long-term outcome of disease, since individuals infected with this strain would survive longer and have a longer lifespan in which to spread the disease. It would thus make sense for a bacterium to secrete high levels of different immunodominant proteins like the ESAT-6 and CFP-10 families to obtain such an early, vigorous host response and the subsequent chronic stable infection.

Betts and coworkers (2000) have hypothesized that the presence of an extra pair of ESAT-6 and CFP-10 proteins in the genome of *M. tuberculosis* CSU#93 (Table 3A.2) may contribute to the inceased host immune response to this strain observed by Manca *et al.* (1999). It is interesting to note that the different mycobacterial species all contain varying amounts of copies of the ESAT-6 and CFP-10 family members in their genomes (Table 3A.2). It has been shown previously that the efficacy of BCG vaccination against leprosy was much greater than that obtained against tuberculosis (Ponnighaus *et al.*, 1992). This observation may be linked to the fact that the ESAT-6 and CFP-10 families form a major part of the antigenic nature of the mycobacteria and that the genome of *M. leprae* contains less copies of the ESAT-6 and CFP-10 families than *M. bovis* BCG, while *M.*

tuberculosis contains more copies than BCG. Thus, the amount of copies of the ESAT-6 and CFP-10 proteins in the different species might be an indication of the differences and variability of their host range and might have an influence on their level of pathogenicity. It must also be kept in mind that for an immunocompromised individual, nonpathogenic mycobacteria may not really exist (Shinnick and Good, 1994).

In addition to the amount of copies in the different genomes, the sequence homology of the orthologous copies also varies between different organisms. Although there is very little sequence divergence among the Mtb9.9 subfamily protein sequences (Figure 3A.4), Alderson *et al.* (2000) showed that even these small changes were enough to cause specificity of T-cells to each peptide in the subfamily and thus heterogeneity in their responses to the antigens. A number of studies (Ravn *et al.*, 1999, Mustafa *et al.*, 2000, Arend *et al.*, 2000a) have clearly indicated that multiple epitopes throughout the sequences of ESAT-6 as well as CFP-10 are recognized by T-cells. Skjøt *et al.* (2000) presented data that indicated that TB10.4 (Rv0288) from the ESAT-6 family induced significantly higher levels of interferon-gamma than ESAT-6 (Rv3875) in tuberculosis patients, while comparable levels were obtained with T-cell responses to CFP-10 and ESAT-6. Immunization of mice with ESAT-6-encoding DNA gave only reasonable protection (considerably less than BCG vaccination) to *M. tuberculosis* infection (Kamath *et al.*, 1999, Li *et al.*, 1999), but immunization with a multivalent combination DNA vaccine (containing the ESAT-6, MPT-64, MPT-63, and KatG constructs) generated a strong protective response comparable to the protection given by BCG (Morris *et al.*, 2000). It would be interesting to investigate the effect of a multi-subunit vaccine consisting of more than one ESAT-6 family member, thus broadening the total epitope population available for recognition. The fact that the broader ESAT-6 family consists of twenty-three potentially strong T-cell antigens, holds a lot of appeal for this strategy. This could be especially useful when vaccination is to be done in a genetically heterogeneous population, as it has been shown that different individuals sometimes recognize different epitopes and different members of these protein families (Ravn *et al.*, 1999, Skjøt *et al.*, 2000, Alderson *et al.*, 2000). It must be added, though, that it has been demonstrated recently that ESAT-6 is able to induce a very potent immune response which controls infection to the same level as BCG vaccination on its own, but only when vaccinated in combination with a very strong adjuvant (Brandt *et al.*, 2000). As there is no correlation between sequence homology and the immunodominance of proteins belonging to the ESAT-6 and CFP-10 families, the explanation for the consistently high level of T-cell responses to members of these families must be due to another factor. This is very likely to be the synchronized upregulation of all members during a certain stage of intracellular infection (Skjøt *et al.*, 2000). It has been suggested previously (Brandt *et al.*, 1996) that because of the fact that ESAT-6 is only produced in low quantities during growth *in vitro*, the fact that it is such a potent target *in vivo* could indicate that there is an upregulation of expression during growth in the macrophage. It is also interesting that Romain and coworkers (1993) found that one of the mycosin proteases (Rv1796 in region 5) elicited DTH responses in guinea pigs only when it was immunized with live *M. bovis* BCG and not with dead bacteria. As we know that the mycosins are expressed constitutively throughout the growth of *M. tuberculosis* (Brown *et al.*, 2000), this might also indicate an upregulation of the expression of the mycosins during *in vivo* replication.

3A.5. Potential links to pathogenicity

M. tuberculosis primarily reside within the granuloma during latent infection where there is an absence of sufficient levels of oxygen, and it has also been suggested that during these periods of low oxygen availability, the organism switches off the expression of certain unessential genes and goes into a state of spore-like inactivity (Imboden *et al.*, 1998). A cDNA clone of ESAT-6 was obtained in a partial *M. tuberculosis* cDNA expression library of genes expressed at reduced (5%) oxygen tension (Imboden *et al.*, 1998, GenBank accession number AA465071), which corresponds to the oxygen levels a bacterium would encounter in the vascular space. Although the oxygen levels in the granuloma might even be lower, it is interesting that ESAT-6 is still being expressed, indicating a potential important function for this protein during infection. Because the library was constructed from bacteria in the late log phase, it also indicates that it is expressed at least up to late log phase and not only in the early stages of development as was previously shown (Andersen *et al.*, 1995). This is confirmed by data that showed the presence of the *cfp-10/esat-6* RNA transcript in early (day 5) and late (day 16) cultures (Berthet *et al.*, 1998).

The RD1 deletion region of *M. bovis* BCG is a major region that is thought to have been deleted during the original attenuation of BCG from *M. bovis* (Mahairas *et al.*, 1996, Oettinger *et al.*, 1999). This region is present within ESAT-6 gene cluster region 1 and *M. bovis* BCG can thus be seen as a natural knockout of ESAT-6 and CFP-10 and some surrounding genes of region 1. The gene Rv3881c, as well as the mycosin associated with the ESAT-6 gene cluster region 1 (mycosin-1 or Rv3883c), are not expressed in *M. bovis* BCG, although both of these genes are present in the genome of this organism (Brown *et al.*, 2000, Mattow *et al.*, 2001). This is in contrast to the ESAT-6 gene cluster region 2 mycosin-2 gene (Rv3886c), lying only 3 ORF's downstream of mycosin-1, which are expressed efficiently in BCG (Brown *et al.*, 2000). As the Rv3881c and mycosin-1 genes lie 2 and 4 ORF's respectively downstream of the RD1 deletion region, it indicates that there are certain factors present within the RD1 deletion region that are needed for mycosin-1/Rv3881c expression, and points to a possible relationship between the genes present in this region. Wards *et al.* (2000) showed that a knockout of Rv3871 (a member of the ATPase protein family D which lies just inside the RD1 deletion region, 4 genes upstream of ESAT-6) resulted in a mutant showing similar loss of virulence in guinea pigs as an *esat-6* knockout. The authors speculate that this might be due to polar effects on the downstream *esat-6* gene. They further showed that an *esat-6/cfp-10* knockout of *M. bovis* was less virulent than its parent, if gross pathology, histopathology and mycobacterial culture of tissues was taken into account, indicating that the ESAT-6 and/or CFP-10 proteins are most likely associated with virulence. This is not limited only to the ESAT-6 protein Rv3875, as a family member of ESAT-6 named Rv0288 (TB10.4 or CFP-7) was previously shown to be highly downregulated (64%) in the attenuated *M. tuberculosis* strain H37Ra (Rindi *et al.*, 1999).

After the original deletion event that resulted in the RD1 deletion in *M. bovis* BCG, further deletions occurred with time in different substrains of BCG (RD2 and RD14 in Pasteur, RD8 in Frappier and Connaught, and RD16 in Moreau - numbering according to Behr *et al.*, 1999), resulting in substrains that showed a very high variability in protective efficiency and a decrease in virulence (Behr *et al.*, 1999, Gordon *et al.*, 1999a, Oettinger *et al.*, 1999). It is hypothesized that the deletion of certain transcriptional regulators (repressors and activators) from RD2, 14 and 16 may have had an influence on the adaptation to environmental change, such as *in vivo* infection (Behr *et al.*, 1999). One of these transcriptional regulators (Rv1773c in RD14) lies just upstream of the ESAT-6 gene cluster region 5 (Rv1782-Rv1798), but its influence on the transcription of the adjacent ESAT-6 gene cluster is unknown. Deletions specific to *M. bovis* (and accordingly also BCG) were also described, and it is interesting to note that two of these deletion regions (RD7 and RD9) span the ESAT-6 and CFP-10 family members Rv2346c and 47c, and Rv3619c and 20c, respectively. These two ESAT-6 genes (Rv2346c and Rv3619c) were recently shown to be part of the immunodominant Mtb9.9 subfamily of the ESAT-6 family (Alderson *et al.*, 2000). It is thus clear that *M. bovis* has a disadvantage compared to *M. tuberculosis* because it contains two less copies of the ESAT-6 and CFP-10 families. It is also tempting to speculate that the deletion of another pair of ESAT-6 and CFP-10 family members (for example with the RD1 deletion) might cause an attenuation of this strain similar to BCG, that would not be observed in *M. tuberculosis* due to the fact that it still has enough copies of the ESAT-6 and CFP-10 gene families. These genetic differences between *M. bovis* and *M. tuberculosis* may provide insights into the phenotypic differences between the two species (Behr *et al.*, 1999). *M. bovis* are not spread easily from person to person, and do not reactivate to the same level as *M. tuberculosis*, although the disease caused by both species is identical in all ways. It was also shown that both RD7 and RD9 (RD5 and RD8 according to Gordon *et al.*, 1999a) are deleted from the *M. microti* genome, although both regions are present on the genome of *M. africanum*. It is thus possible that these regions may effect host range and virulence and also be the cause of the phenotypical differences observed between members of the *M. tuberculosis* complex (Gordon *et al.*, 1999a).

By using a library of signature-tagged transposon mutants, Camacho *et al.* (1999) identified a number of genes that affected multiplication of *M. tuberculosis* within the lungs of mice. One of these attenuated mutants contained a transposon insertion into the gene Rv3018c, which is one of the PPE genes co-duplicated with an ESAT-6/CFP-10 operon singular copy (Rv3019c/20c). This insertion may have disrupted the operon structure and therefore may have an influence on the expression of the ESAT-6 and CFP-10 proteins situated within the region, giving a false impression that the Rv3018c gene is a virulence factor.

One factor which has to be explained if the ESAT-6 gene clusters were to be linked to virulence, is the presence of a copy of ESAT-6 gene cluster region 4 in the genomes of the Corynebacteria, as well as an ortholog of this region in the genome sequence of *S. coelicolor*. One explanation may be the existence of an earlier function for these proteins shared by all the high G+C Gram positives, after which duplication of the region resulted in evolution of function. The virulence properties may also be dependent on dosage or the amount of antigenicity of the respective ESAT-6 proteins. It is interesting to note that the corresponding ESAT-6 family proteins from region 4 have not been identified as immunologically relevant antigens of *M. tuberculosis* (Skjøt *et al.*, 2001), which may indicate that they have a low antigenicity. As the ESAT-6 and CFP-10 proteins have never been described in bacteria other than the mycobacteria, it would be interesting to look at the expression, localization and antigenicity of the orthologues in the Corynebacteria.

3A.6. Concluding Remarks

The function of the ESAT-6 gene clusters in the growth and pathogenicity of the mycobacteria is still unknown, but all the above information on the genes that form part thereof indicates that it may play a very important role in the growth of these organisms and in tuberculosis infection. We summarize the importance and indications of an association with virulence as follows:

(1) Multiple copies of the gene cluster. *M. tuberculosis* and *M. bovis* contain five duplications of this gene cluster on the genome, *M. avium* four copies and *M. leprae* three. These proteins must have some important or significant function to have been functionally retained as multiple copies on the genomes of these virulent mycobacteria.

(2) *M. leprae* clusters are fully functional. It is suggested that genes in the genome of *M. leprae* that are not absolutely necessary for intracellular survival became non-functional (in other words it contains multiple stopcodons and frameshifts) and that the genome of *M. leprae* contains the minimal gene set required by a pathogenic mycobacterium (Vissa and Brennan, 2001). Most of the genes in the three gene clusters in *M. leprae* contain no stopcodons or frameshifts (in other words seems to be functional) and there is experimental evidence for the successful secretion of the region 1 ESAT-6 homologue L-ESAT in *M. leprae*.

(3) ESAT-6 and CFP-10. These regions contain members of the ESAT-6 and CFP-10 families, which are small proteins of unknown function and are potent, secreted T-cell antigens. They have no detectable secretion signals, but seem to be actively secreted from early in infection.

(4) RD1 deletion region. Most of the genes of region 1 lie in the RD1 deletion region of the attenuated strain *M. bovis* BCG. It is commonly thought that the deletion of RD1 from *M. bovis* led to the attenuation of this organism to BCG. This indicates the importance of this region with regard to virulence.

(5) Other ESAT-6 family deletions. Other ESAT-6 and CFP-10 family members have been deleted in other RD regions in *M. bovis*.

(6) Downregulation in H37Ra. The ESAT-6 family member Rv0288 (belonging to gene cluster region3) is highly downregulated in the avirulent strain *M. tuberculosis* H37Ra.

(7) *Esat-6/cfp-10* knockout. A decrease in virulence has been shown to occur with infection with an *esat-6/cfp-10* knockout mutant in guinea pigs.

(8) Rv3871 knockout. A comparable decrease in virulence has been shown to occur in a knockout of Rv3871 (the 3 X ATP/GTP binding ATPase).

(9) Rv3018c knockout. Attenuation of virulence as measured by multiplication within mice lungs have been obtained by a knockout mutant of the PPE gene associated with a singular duplication of ESAT-6 and CFP-10.

(10)Expression under oxygen tension. ESAT-6 is expressed under reduced oxygen tension comparable to conditions within the vascular space, and also during late log phase.

(11) Extra pair of ESAT-6 and CFP-10 in strain CSU#93. Strain CSU#93, which has been shown to induce a more robust immune response than other *M. tuberculosis* strains, contains one more copy of ESAT-6 and CFP-10.

(12) One less copy in Erdman. There is evidence to suggest that *M. tuberculosis* strain Erdman, which has been shown to be less virulent than other *M. tuberculosis* strains, contains one less copy of ESAT-6 (and possibly also the associated CFP-10).

The multiplicity of these gene clusters, their immunological significance as well as the links to pathogenicity suggest that they have an important function in the mycobacteria and are worth further investigation.

ADDENDUM 3B

DIAGNOSTIC POTENTIAL

"Consumption, that great destroyer of human health and human life, takes the first rank as an agent of death. Any facts regarding a disease that destroys one-seventh to one-fourth of all that die, cannot but be interesting."

Lemuel Shattuck (1849)

NOTE: The views presented in the following addendum will be submitted as a counterpoint article for peer review and publication as: "**ESAT-6 and CFP-10: What is the diagnoses?**, Gey van Pittius, N.C."

3B.1. Introduction

There is a constant search for new and effective diagnostic tests for the determination of *M. tuberculosis* infection due to the non-specificity of the current tuberculin test (Andersen *et al.*, 2000). The ESAT-6 and CFP-10 proteins have been evaluated over the past few years and have shown promise as a tool to differentiate between BCG vaccination, *M. avium* infection and *M. tuberculosis* infection (Ravn *et al.*, 1999, Colangeli *et al.*, 2000, Arend *et al.*, 2000b). The major problem is that contrary to common belief (Harboe *et al.*, 1996, Elhay *et al.*, 1998, Ravn *et al.*, 1999, Arend *et al.*, 2000a, Arend *et al.*, 2000b, Van Pinxteren *et al.*, 2000, Andersen *et al.*, 2000, Arend *et al.*, 2001a, Arend *et al.*, 2001b), these proteins are not *M. tuberculosis* specific (Gey van Pittius *et al.*, 2001). They are also present in fast growing environmental mycobacterial species (with a high percentage of protein homology), the presence of which may therefore interfere with diagnostic tests based on these antigens (Gey van Pittius *et al.*, 2001). This should be especially evident in developing countries where environmental mycobacteria are present in large amounts (Vekemans *et al.*, 2001).

3B.2. Discussion

What is the definition of a "good diagnostic agent"? The most important criterion must surely be specificity towards the infecting organism, failure of which could lead to false or misleading diagnoses (Andersen *et al.*, 2000). Diagnostic tests currently in use for the diagnoses of *Mycobacterium tuberculosis* infection are based on the Mantoux skin test, making use of purified protein derivative (PPD or tuberculin), but because the protein constituents of PPD are shared between several non-pathogenic environmental mycobacterial species as well as the vaccine strain *M. bovis* BCG, it is poorly specific and does not perform well as a diagnostic tool (Chaparas *et al.*, 1970, Huebner *et al.*, 1993, Andersen *et al.*, 2000). Thus, novel approaches for the specific diagnoses of *Mycobacterium tuberculosis* infection are constantly being evaluated. Recently, a great deal of interest has been shown towards two small, secreted proteins of *M. tuberculosis*, namely ESAT-6 (encoded by the gene *esx* or *esat-6*, Sørensen *et al.*, 1995) and CFP-10 (encoded by the gene *lhp*, Bethet *et al.*, 1998). These proteins are potent T-cell antigens (Skjøt *et al.*, 2000) and their genes are absent from the genomes of *M. bovis* BCG (the tuberculosis vaccine strain, Mahairas *et al.*, 1996) as well as *M. avium* (an agent of opportunistic infections, Gey van Pittius *et al.*, 2001). In *M. bovis* BCG this is due to a 9505 bp deletion named RD1, commonly thought to have resulted from the serial passage of *M. bovis* by Calmette and Guèrin between 1908 and 1921 (Mahairas *et al.*, 1996). Consequently, the ESAT-6 and CFP-10 antigens from the RD1 deletion region have been the focus of recent research efforts because of their application as potential diagnostic markers to differentiate between *M. bovis* BCG vaccination, and *M. tuberculosis*, *M. bovis* or *M. avium* complex infection (Ravn *et al.*, 1999, Colangeli *et al.*, 2000, Arend *et al.*, 2000b). *Esat-6* and *lhp* belong to a family of 21 other genes distributed throughout the genome of *M. tuberculosis* (Cole *et al*, 1998, Gey van Pittius *et al.*, 2001, Skjøt *et al.*, 2001, Addendum 3A). These other copies of the *esat-6* and *lhp* genes are also situated adjacent to each other, suggesting that they were co-duplicated (Gey van Pittius *et al.*, 2001). The protein sequence homology between the different copies within each of these two protein families varies between 15 to 27% for ESAT-6 and between 9 to 32% for the CFP-10 proteins, respectively (Gey van Pittius *et al.*, 2001). Due to the very low percentage homology of the paralogs in a specific organism, it would be safe to say that the other copies of ESAT-6 and CFP-10 within a specific organism should not interfere with a potential diagnostic test based on the two antigens from ESAT-6 gene cluster region 1.

The most recent evaluation of the diagnostic potential of an assay for *M. tuberculosis* infection based on the ESAT-6 and CFP-10 proteins (from the RD1 deletion region) was performed by Arend and colleagues (2001a). In this study the authors refer to ESAT-6 and CFP-10 as being *M. tuberculosis*-specific antigens absent from most environmental mycobacteria, due to the well-known absence of these proteins from *M. bovis* BCG and *M. avium*. This assumption was further based on previously published Southern blotting results that indicated the absence of these genomic domains from most environmental mycobacteria (Sorenson *et al.*, 1995).

We propose that this statement is incorrect. In a recent comparative genomic analysis we have found that the genomes of all members of the mycobacteria that are currently being sequenced contain copies of different members of the ESAT-6 and CFP-10 families (Gey van Pittius *et al.*, 2001). More specifically, a copy of the ESAT-6 and CFP-10 proteins of the RD1 deletion region (Rv3874 and Rv3875) could be found in the genomes of *M. leprae* and even the distantly-related, non-pathogenic, fast-growing, environmental mycobacterium *M. smegmatis*. This data is further supported by results dating back to 1995, which showed that the genes for these proteins are also present in other pathogenic mycobacteria (*M. africanum*, *M. kansasii*, *M. marinum*, *M. szulgai* - Sorenson *et* al., 1995 and *M. bovis* - Harboe *et al.*, 1996), as well as the slow-growing non-pathogenic mycobacterium *M. gastri* (Colangeli *et al.*, 2000) and the fast-growing non-pathogenic environmental species *M. flavescens* (Harboe *et al.*, 1996). The similarity between the orthologs in *M. smegmatis* and the so-called "*M. tuberculosis*-specific antigens ESAT-6 and CFP-10" is 80% and 71%, respectively (see Figure 3B.1 for sequence alignment). Therefore, it is likely that these proteins share epitopes that could be recognized by the T-cells and which may result in a similar T-cell response. Furthermore, given the evolutionary history of the mycobacteria (see Addendum 3A, Figure 3A.1) and the presence of the RD1 deletion region-specific ESAT-6 and CFP-10 in *M. smegmatis* and the other mentioned species, it is highly plausible that this region would be present in the genomes of most other environmental mycobacteria. The homology between ESAT-6 and CFP-10 of the many environmental mycobacterial strains phylogenetically more closely related to *M. tuberculosis* may even be higher than that between the antigens of *M. tuberculosis* and *M. smegmatis*. It is therefore imperative to study the extent of sequence similarity between the ESAT-6 and CFP-10 proteins of different pathogenic and non-pathogenic environmental mycobacteria before a member of these immunodominant families can be considered as a specific marker for *M. tuberculosis* infection. Furthermore, studies will have to be initiated to determine the relative influence of secreted ESAT-6 and CFP-10 from environmental mycobacteria on the T-cell responses from suspected infected individuals. Interferon gamma production in response to ESAT-6 and CFP-10 from environmental mycobacteria by peripheral blood mononuclear cells from infected patients has to our knowledge not been done. This is surprising given the fact that numerous studies have already been performed on the use of these antigens as diagnostic tools (Ravn *et al.*, 1999, Andersen *et al.*, 2000, Arend *et al.*, 2000a, Arend *et al.*, 2000b, Dillon *et al.*, 2000, Skjøt *et al.*, 2000, Ulrichs *et al.*, 2000, Van Pinxteren *et al.*, 2000, Arend *et al.*, 2001a, Arend *et al.*, 2001b, Vekemans *et al.*, 2001, Vordermeier *et al.*, 2001). The only explanation for this omission must be the misleading absence of these antigens from *M. bovis* BCG and *M. avium*. Until results are obtained which indicate that the host cellular immune response is able to distinguish between the ESAT-6 and CFP-10 proteins secreted from either environmental mycobacteria or *M. tuberculosis*, claims regarding the potential diagnostic use of these antigens in the diagnoses of *M. tuberculosis* infection have to be treated with caution. Even if proved that the response is specific, it is still misleading to term the ESAT-6 and CFP-10 antigens *M. tuberculosis*-specific.

The presence of ESAT-6 and CFP-10 in the genomes of other mycobacteria may also explain the significant proportion of PPD-positive individuals showing memory T-cell responses towards ESAT-6 and CFP-10 without evident disease (Ravn *et al.*, 1999). These people may have come in contact with environmental mycobacteria and could thus be sensitized with the ESAT-6 and CFP-10 secreted by the organisms, causing cross-reactivity when the assays are performed. In a recent paper by Vekemans and colleagues (2001) it was found that tuberculosis contacts but not patients in the Gambia have higher interferon gamma responses to ESAT-6 than do community controls. In spite of this, 30% of the community controls produced interferon gamma in response to ESAT-6, which was proportionately similar to the number observed in the patient group. This is also only 8% less than the 38% of the community control individuals who had a positive skin response to tuberculin, indicating that there is a highly plausible that these responses were caused by exposure to environmental mycobacteria. The authors conclude that an ESAT-6 interferon gamma assay "will be of limited use in the diagnoses of tuberculosis in countries where tuberculosis is endemic".

It is clear that the promising results obtained with ESAT-6 and CFP-10 in industrialized countries will be of little benefit to those people living in developing countries (where the real need for these tests lies), underlining the fact that studies done on tuberculosis in industrialized countries can not necessarily always be applied directly to the developing world. Thus, considering the amount of specificity of ESAT-6 and CFP-10 to *M. tuberculosis*, can we really say that they fulfil the main criterion of a good diagnostic agent?

Figure 3B.1. Multiple protein sequence alignment between RD1 deletion region-specific ESAT-6 and CFP-10 orthologs from *M. tuberculosis* and *M. smegmatis*. Although studies have indicated the presence of multiple T-cell epitopes scattered throughout the ESAT-6 protein sequence (Ulrichs *et al*., 1998, Mustafa *et al*., 2000), the positions of predominantly recognized epitopes are indicated. Data for epitopes were obtained from Brandt *et al*. (1996), Ulrichs *et al*. (1998), Ravn *et al*. (1999) and Mustafa *et al*. (2000).

ESAT-6

Mouse T-cell epitope (aa 1-20)

Human T-cell epitope (aa 1-30) - Germany

Human T-cell epitope (aa 42-75) - Ethiopia

Mouse T-cell epitope (aa 51-70)

```
M. tb     MTEQQWNFAGIEAAASAIQGNVTSIHSLLDEGKQSLTKLAAAWGGSGSEAYQGVQQKWDA
M. smeg   MTEQVWNFAGIEGGASEIHGAVSTTAGLLDEGKASLTTLASAWGGTGSEAYQAVQARWDS
          **** *******..** *:* *::  .****** ***.**:****:******.** :**:
```

Human T-cell epitope (aa 72-95) - Kuwait and Denmark

```
M. tb     TATELNNALQNLARTISEAGQAMASTEGNVTGMFA 95
M. smeg   TSNELNLALQNLAQTISEAGQTMAQTEAGVTGMFA 95
          *:.*** ******:*******:**.**..******
```

CFP-10

```
M. tb     MAEMKTDAATLAQEAGNFERISGDLKTQIDQVESTAGSLQGQWRGAAGTAAQAAVVRFQE
M. smeg   MAAMNTDAAVLAKEAANFERISGELKGVIAQVESTGSALAAQMVGQAGTAAQAALARFHE
          ** *:****.**:**.*******:**  * *****..:* .*  * ********:.**:*

M. tb     AANKQKQELDEISTNIRQAGVQYSRADEEQQQALSSQM 98
M. smeg   AAAKQVQELNEISANIHTSGTQYTSTDEDQAGTLASSM 98
          ** ** ***:***:**: :*.**: :**:*  :*:*.*
```

CHAPTER FOUR

OPERON STRUCTURE

".....the genome is considered as a mosaic of independent molecular blueprints for the building of individual cellular constituents. In the execution of these plans, however, coordination is evidently of absolute survival value."

F. Jacob and J. Monod (1961)

NOTE: The results presented in the following chapter will be submitted for peer review and publication as: "**Mycosin-3, a Subtilisin-like Serine Protease *of Mycobacterium tuberculosis*, is expressed as part of the ESAT-6 gene cluster region 3 operon along with members of the ESAT-6, CFP-10, PE and PPE multigene families.** Gey van Pittius, N.C., Warren, R.M., and Van Helden, P.D."

4.1. Introduction

Gene clusters are prominent features of bacterial chromosomes. Prokaryotic gene clusters are typically composed of functionally related genes (Overbeek *et al.*, 1999). These functionally coupled genes are mostly situated adjacent to each other on the same strand with intergenic gaps of no more than 300 bp. The probability of a functional relationship between the genes in a cluster further increases if the cluster is conserved between different bacterial species (Overbeek *et al.*, 1999). Examples of these types of gene clusters have been known for many years and include the genes for lactose utilization (*lac*), galactose utilization (*gal*), histidine biosynthesis (*his*) and tryptophan biosynthesis (*trp*) (Lawrence and Roth, 1996). Gene clusters are also found in the genome of *M. tuberculosis*, and include for example the *mce* operons and the various operons in which the ATP-binding cassette (ABC) transporter superfamilies are situated (Tekaia *et al.*, 1999). We have previously described a gene cluster situated in the genome of *Mycobacterium tuberculosis* and other mycobacteria, as well as in the genomes of members of the Corynebacteria and Streptomyces (Gey van Pittius *et al.*, 2001). This gene cluster is duplicated five times in the genome of *M. tuberculosis* and contains members of the important T-cell antigen ESAT-6 gene family, leading to the clusters being designated the ESAT-6 loci (Tekaia *et al.*, 1999). The gene organization within the ESAT-6 clusters and the conservation thereof between species have been described in detail (Gey van Pittius *et al.*, 2001). Visual inspection of the close proximity of the genes within the clusters indicates that they may constitute one or more operons, which is supported by the fact that they are conserved between different bacterial species. Furthermore, it has already been demonstrated that the *esat-6* gene from the ESAT-6 gene cluster region 1 forms part of an operon with the gene *lhp*, which is situated directly adjacent to it (Berthet *et al.*, 1998). Although the functions of the genes situated in the clusters remain unknown, it is hypothesized that these genes may encode proteins involved in the active transport of the members of the ESAT-6 protein family (Tekaia *et al.*, 1999, Gey van Pittius *et al.*, 2001). These proteins are secreted without ordinary secretion signals and are potent T-cell antigens of *M. tuberculosis*. As the relationships between the genes within the clusters may shed light on their function, we decided to investigate one of the ESAT-6 gene cluster regions (region 3: Rv0282-Rv0292) to determine if these clusters are of an operonic nature and to identify the promoter driving the expression of this cluster. This region includes a member of the previously described mycosin subtilisin-like serine protease family (mycosin-3, Brown *et al.*, 2000). The ESAT-6 protein encoded by this cluster (Rv0288, TB10.4 or CFP-7) was shown to be a potent secreted T-cell antigen (Skjøt *et al.*, 2000) and was previously shown to be highly downregulated (64%) in the attenuated *M. tuberculosis* strain H37Ra (Rindi *et al.*, 1999). It is thus of utmost importance to study and identify the regulatory mechanisms controlling the expression of this gene and the cluster it is situated in.

In this study, we demonstrate that the ESAT-6 gene cluster region 3 is expressed as one single polycistronic RNA. In addition to this we have cloned various intergenic regions from this cluster and subsequently identified the promoter driving the expression of this region. This work opens the way for the study of the mechanisms controlling the expression and regulation of the ESAT-6 antigen family and their putative secretion system.

4.2. Materials and Methods

4.2.1. *DNA sequence analyses*

All DNA sequence information was obtained from the publicly available finished and unfinished genome sequence databases. These databases are accessible on the internet at the URL's listed in Table 4.1. Multiple sequence alignments of promoter regions were done with the program ClustalW 1.5 on the ClustalW WWW server at the European Bioinformatics Institute website (http://www2.ebi.ac.uk/clustalw/; Thompson *et al.*, 1994).

Table 4.1. Genome sequencing project data used in this study:

Organism	**Website(s)**
Mycobacterium tuberculosis H37Rv	http://genolist.pasteur.fr/TubercuList/
Mycobacterium tuberculosis CDC1551	http://www.tigr.org/tigr-scripts/CMR2/GenomePage3.spl?database=gmt
Mycobacterium tuberculosis 210	http://www.tigr.org/cgi-bin/BlastSearch/blast.cgi?
Mycobacterium bovis AF2122/97	http://www.sanger.ac.uk/Projects/M_bovis/
Mycobacterium leprae TN	http://genolist.pasteur.fr/Leproma
Mycobacterium avium 104	http://www.tigr.org/cgi-bin/BlastSearch/blast.cgi?organism=m_avium
Mycobacterium paratuberculosis K10	http://www.cbc.umn.edu/ResearchProjects/AGAC/Mptb/Mptbhome.html
Mycobacterium smegmatis mc^2 155	http://www.tigr.org/cgi-bin/BlastSearch/blast.cgi?organism=m_smegmatis

4.2.2. *Bacterial strains*

Escherichia coli JM109 was used as a host for all cloning experiments. *Mycobacterium tuberculosis* H37Rv (laboratory strain; ATCC 25618) was used for the characterization of operon structure and *M. smegmatis* mc^{2}155 (Snapper *et al.*, 1990) as a heterologous mycobacterial cloning host to study promoter activity.

4.2.3. *Media and culture conditions*

E. coli was grown on solid or liquid Luria-Bertani (LB) medium as described by Sambrook *et al.* (1989). Mycobacterial strains were grown at 37°C for 2 days with shaking (200 rpm, *M. smegmatis*) or 14 days with stirring (*M. tuberculosis*) in Middlebrook 7H9 broth (Difco) supplemented with filter-sterile ADC (0.5% BSA, 0.2% glucose, 0.015% catalase) and containing 0.05% Tween 80 (Sigma). All work on *M. tuberculosis* H37Rv was performed in a Biosafety Level III facility. Kanamycin (50 μg/ml, Roche) and Ampicillin (50 μg/ml, Roche) were added to bacterial cultures when antibiotic selection was required. For transformant selection and *β*-galactosidase activity detection on solid media, *M. smegmatis* cells were grown on Middlebrook 7H11 agar supplemented with filter-sterile OADC (0.005% oleic acid, 0.5% BSA, 0.2 % glucose, 0.02% catalase, 0.085% NaCl) and containing 0.05% Tween 80 (Sigma). For *β*-galactosidase activity assays in liquid media, mycobacterial cultures (10 ml) were grown to an optical density of 1. The cells were pelleted by centrifugation, resuspended in 500 μl phosphate-buffered saline (PBS) and sonicated at 4.5 setting in a Misonix cup sonicator on ice for a total of 5 minutes (15-second bursts with 30-second intervals).

4.2.4. Primers

Table 4.2 contains a list of all oligonucleotide primers used in this study. Primers were chosen according to length, Tm, G+C content and length of the product, to efficiently amplify the complete selected intergenic region. To avoid problems with RT-PCR reactions due to length restrictions of products, all primers were chosen to result in a product less than 300 bp.

Table 4.2. List of RT-PCR and PCR oligonucleotide primers

Name of primer	Primer sequence (from 5' to 3')	Length of primer	Tm * [°C]	G+C (%)	Intergenic region covered	Length of product
T028182f	gttcgcccgcaacaccct	18 bp	60	66.7	From 40 bp inside Rv0281 to 18 bp inside Rv0282	281 bp
T028182r	ttcacctacgcccgccat	18 bp	58	61.1		
T028283f	caatggtcggtttgtgcg	18 bp	56	55.6	From 187 bp inside Rv0282 to 24 bp inside Rv0283	207 bp
T028283r	gtggtcgtgctgctggttc	19 bp	62	63.2		
T028384f	cggtgctgtcgctgtttgt	19 bp	60	57.9	From 125 bp inside Rv0283 to 121 bp inside Rv0284	242 bp
T028384r	gtcgtagcagtgacggtggg	20 bp	66	65.0		
T028485f	ccgacagtgatagtccaacct	21 bp	64	52.4	From 98 bp inside Rv0284 to 172 bp inside Rv0285	270 bp
T028485r	acgccctgtgcactgaac	18 bp	58	61.1		
T028586f	agctacctggccggtgatg	19 bp	62	63.2	From 57 bp inside Rv0285 to 157 bp inside Rv0286	216 bp
T028586r	ccagtagcccactgagttctgc	22 bp	70	59.1		
T028687f	gaccgcaaccaaagaacgc	19 bp	60	57.9	From 184 bp inside Rv0286 to 40 bp inside Rv0287	273 bp
T028687r	gaggccaccaactgtgggata	21 bp	66	57.1		
T028788f	caggcgaatctgggtgag	18 bp	58	61.1	From 78 bp inside Rv0287 to 35 bp inside Rv0288	142 bp
T028788r	aacatcgcggggtagttgta	20 bp	60	50.0		
T028889f	acccatgaagccaacacca	19 bp	58	52.6	From 69 bp inside Rv0288 to 195 bp inside Rv0289	274 bp
T028889r	aacaccctgcggcgataa	18 bp	56	55.6		
T028990f	tgggtctccaccttcagcc	19 bp	62	63.2	From 135 bp inside Rv0289 to 68 bp inside Rv0290	249 bp
T028990r	gccatctcggtcaacctgct	20 bp	64	60.0		
T029091f	tgtggtggcactgaacccg	19 bp	62	63.2	From 154 bp inside Rv0290 to 28 bp inside Rv0291	182 bp
T029091r	gccgccagacacgcaaat	18 bp	58	61.1		
T029192f	gtgctggtcgggctcacag	19 bp	64	68.4	From 66 bp inside Rv0291 to 105 bp inside Rv0292	171 bp
T029192r	gcacgtaatcgcgtgtcga	19 bp	60	57.9		
T387374f	gcaggagcgtgaagaagac	19 bp	60	57.9	From 55 bp inside Rv3873 to 94 bp inside Rv3874	241 bp
T387374r	cctggtcgatctgggtttt	19 bp	58	52.6		

*Tm were calculated using the following formula: [4x (G+C)] + [2x (A+T)].

4.2.5. RNA preparation

All procedures involving RNA were performed under the strictest RNase-free conditions. Total RNA was extracted from mycobacterial cultures by firstly pelleting the cells in 50 ml centrifuge tubes at 4 000 rpm for 5 minutes. Thereafter, the culture supernatant was removed and Trizol (GibcoBRL) was immediately added to the cell pellet (1ml of Trizol per 20 ml volume of original culture). Cells were resuspended by pipetting and transferred to Blue FastPrep tubes containing silicone beads (Bio101). Cells were ribolyzed in a FastPrep bead-beater (Bio101) at speed setting 6.5 for 45 seconds. Cells were placed on ice for 5 minutes immediately after ribolyzing, and centrifuged at 13 000 rpm for 1 minute in a microfuge. The supernatant was removed from the beads/cell debris and transferred to a 1.5 ml microfuge tube containing 500 µl chloroform. The tube was vortexed, and placed on ice for 2 minutes with intermittent vortexing. The tube was centrifuged at 13 000 rpm for 5 minutes, and the aqueous phase was transferred to a 1.5 ml microfuge tube containing 500 µl isopropanol. Tubes were incubated at -20°C overnight. RNA was pelleted by centrifugation at 13 000 rpm for 30 minutes at 4°C. The pellet was washed with 1 ml of 70% ethanol and briefly air-dried. Nuclease-free H_2O (200 µl per 20 ml volume of original culture, Promega) containing DNAseI buffer, 1 U/µl RNasin and 10 mM DTT was added to the tube to resuspend the RNA. This was incubated on ice for 30 minutes with intermittent pipetting (no vortexing to avoid shearing of the RNA). DNAseI was added at 0.1 U/µl and incubated at 37°C for 30 minutes, after which stop solution was added (20 µl per 200 µl), and the RNA purified using the QIAGEN RNeasy mini-kit, as described by the manufacturer. Briefly, 350 µl buffer RLT and 250 µl ethanol were added per 100 µl RNA. This was applied to a minicolumn, washed twice with buffer RLT, and once with buffer RPE. RNA was eluted with nuclease-free H_2O. A second round of DNaseI treatment and column purification was performed. The integrity of the RNA was assayed on a 1 % TAE agarose gel. RNA stocks were stored in single-use aliquots at -20°C with 1 U/µl RNasin and 10 mM DTT. RNA was confirmed to be free of any DNA contamination by PCR analysis.

4.2.6. RT-PCR analyses

RT-PCR reactions were performed using the Roche Titan One-Tube RT-PCR System kit in combination with the Promega HotstarTaq PCR System, as described by the manufacturer. Two separate master mixes were prepared, and only mixed together immediately prior to cycling. Master mix 1 contained the RNA template, dNTP's (200 μM final concentration), primers (50 pmol each), DTT (10 mM final concentration), RNasin (1 U/μl) and water to a final volume of 25 μl. Master mix 2 contained 5X RT-PCR buffer (with $MgCl_2$ to a final concentration of 1.5 mM), Reverse Transcriptase enzyme mix (1 μl), HotstarTaq PCR enzyme mix (0.2 μl), Q-buffer (10 μl) and water to a final volume of 25 μl. Reverse transcription and cycling were performed without interruption in a Perkin Elmer GeneAmp 2400 under the conditions indicated in Table 4.3. Results of RT-PCR reactions were visualized on a 2 % TAE agarose gel.

Table 4.3. RT-PCR cycling parameters

Number of Cycles	Reaction, temperature and time duration
1 cycle	reverse transcription at 50°C for 30 minutes
1 cycle	template denaturing at 94°C for 2 minutes
1 cycle	Hotstart Taq polymerase activation at 95°C for 15 minutes
35 cycles	denaturing at 94°C for 30 seconds annealing at *x*°C* for 30 seconds elongation at 72°C for 30 seconds
1 cycle	elongation at 72°C for 7 minutes

**x*°C is the optimum anealing temperature for the specific primer pair.

4.2.7. DNA manipulations for promoter cloning

All DNA manipulations (plasmid extraction, DNA cloning and restriction digestions) were performed essentially as described by Sambrook *et al.*, (1989). Details of the plasmids and plasmid constructs with the sequence positions relevant to their construction are presented in Table 4.4. *lacZ* operon transcriptional fusions were constructed using a mycobacterial-*E. coli* shuttle vector from the pJEM series, pJEM15 (Timm *et al.*, 1994), a gift from J. Rauzier (Institut Pasteur, Paris, France). Oligonucleotide pairs T387374f and r, T028182f and r, T028687f and r, and T028990f and r, were used to PCR amplify intergenic regions (see Table 4.1) from *M. tuberculosis* H37Rv DNA (a gift from J.T. Belisle, Colorado State University, USA). Intergenic regions chosen for amplification were selected according to DNA sequence analysis of the ESAT-6 gene cluster region 3. Intergenic region Rv3873-Rv3874 was chosen as a positive control for promoter activity by an ESAT-6 region promoter, as an ESAT-6 operon promoter has been identified from this region (region 1) previously (Berthet *et al.*, 1998). Intergenic region PCR products were cloned into the T-vector pGemT-Easy (Promega), resulting in pGemT7374, pGemT8182, pGemT8687 and pGemT8990 respectively. Inserts were excised using *Eco*RI, and cloned into the *Eco*RI site of pBluescript II KS (Stratagene), resulting in pBlue7374, pBlue8182, pBlue8687 and pBlue8990 respectively. *Bam*HI / *Kpn*I fragments were excised from these constructs and finally cloned into the corresponding sites in the mycobacterial-*E.coli* promoter-probe shuttle vector pJEM15, creating transcriptional fusions of the intergenic regions with a promoterless *lacZ* operon. These constructs were named pJEM7374, pJEM8182, pJEM8687 and pJEM8990 respectively. All essential constructs created during the cloning of the intergenic regions were verified by DNA sequencing. Transformation *of M. smegmatis* with the pJEM constructs was performed using electroporation, as described previously (Jacobs *et al.*, 1991), and transformants were selected on Kanamycin-containing Middlebrook 7H11 agar plates.

4.2.8. β-Galactosidase assays

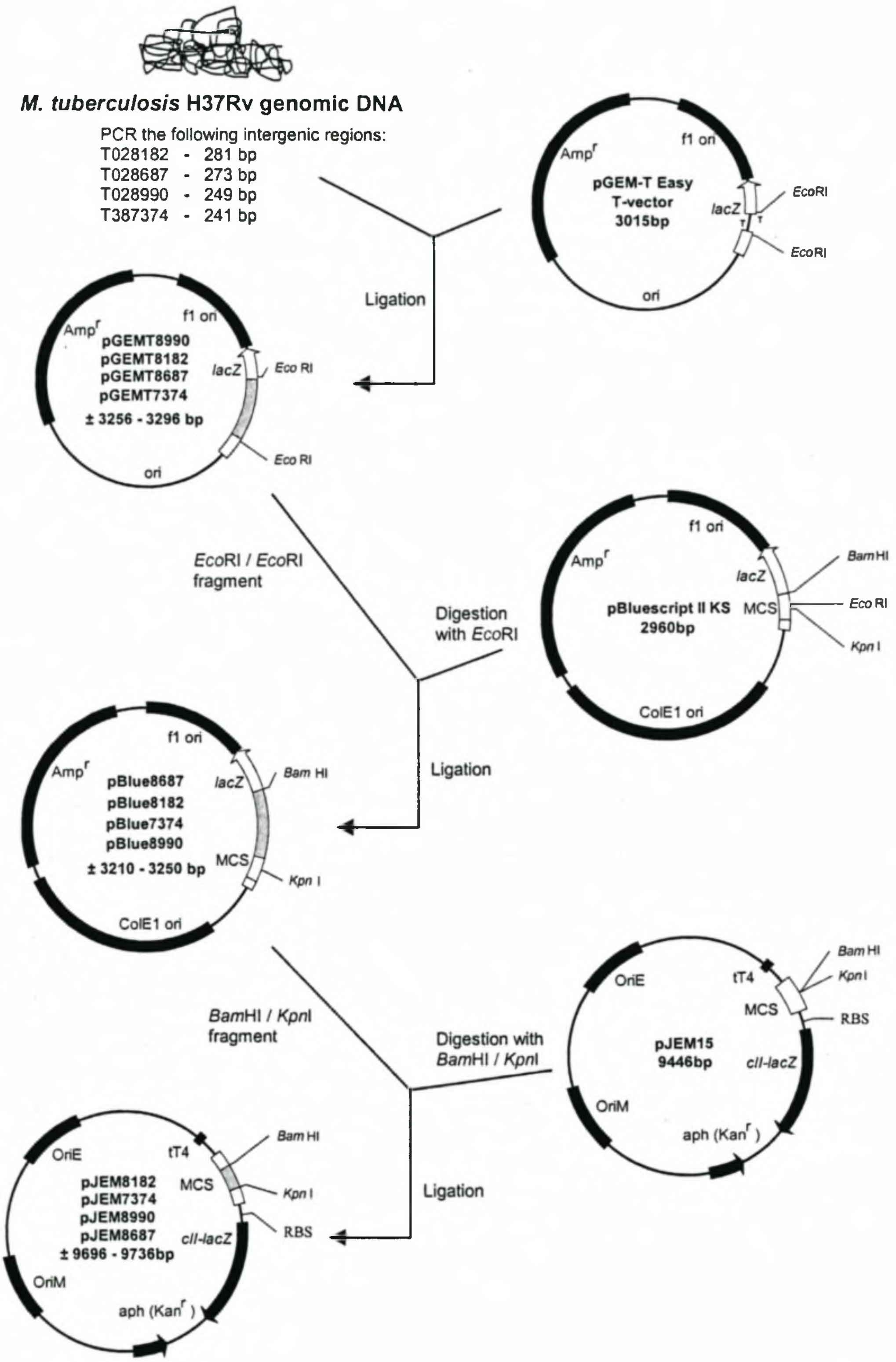
β-Galactosidase activity of *M. smegmatis* transformants were detected on solid media by plating transformants on Middlebrook 7H11 media containing 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-Gal; 0.001%). β-Galactosidase activity was quantitatively assayed in liquid media as described previously (Pardee et al., 1959, Timm et al., 1994) using sonicated extracts (see Media and culture conditions) of *M. smegmatis* transformed with different pJEM constructs. This assay makes use of the β-galactosidase substrate *o*-nitrophenyl-β-D-galactoside (ONPG) and one unit of β-galactosidase is defined as producing 1 μMole *o*-nitrophenol per minute from *o*-nitrophenyl-β-D-galactoside at 28°C, pH 7.0 (1 μMole/ml *o*-nitrophenol has an optical density at 420 nm of 0.0075). β-Galactosidase activity was spectrophotometrically measured at OD_{420} after a 30 minute incubation in equal amounts of protein (6.6 μg). Protein concentrations of sonicated extracts were determined spectrophotometrically using the BioRad protein concentration assay. All activity assays were performed in triplicate. To calculate the units of β-galactosidase we used the formula: 1Unit = 200 x OD_{420}/mg of protein/minute, as was described by Timm *et al.* (1994) and Berthet *et al.* (1998).

Table 4.4. Plasmids and constructs used in this study

Plasmids	Characteristics*	Source or reference
pGemT-Easy	PCR Cloning T-vector, Amp^r	Promega
pGemT7374	241 bp *M. tuberculosis* H37Rv PCR product using primers T387374f and T387374r, cloned into pGemT-Easy, Amp^r	This study
pGemT8182	281 bp *M. tuberculosis* H37Rv PCR product using primers T028182f and T028182r, cloned into pGemT-Easy, Amp^r	This study
pGemT8687	273 bp *M. tuberculosis* H37Rv PCR product using primers T028687f and T028687r, cloned into pGemT-Easy, Amp^r	This study
pGemT8990	249 bp *M. tuberculosis* H37Rv PCR product using primers T028990f and T028990r, cloned into pGemT-Easy, Amp^r	This study
pBluescript II KS	Cloning vector, Amp^r	Stratagene
pBlue7374	*Eco*RI fragment from pGemT7374 subcloned into pBluescript II KS, Amp^r	This study
pBlue8182	*Eco*RI fragment from pGemT8182 subcloned into pBluescript II KS, Amp^r	This study
pBlue8687	*Eco*RI fragment from pGemT8687 subcloned into pBluescript II KS, Amp^r	This study
pBlue8990	*Eco*RI fragment from pGemT8990 subcloned into pBluescript II KS, Amp^r	This study
pJEM15	Promoter-probe vector used for creating transcriptional fusions with *lacZ*, Kan^r	Timm *et al*., 1994
pJEM7374	*Bam*HI / *Kpn*I fragment from pBlueT7374 subcloned into pJEM15, Kan^r	This study
pJEM8182	*Bam*HI / *Kpn*I fragment from pBlueT8182 subcloned into pJEM15, Kan^r	This study
pJEM8687	*Bam*HI / *Kpn*I fragment from pBlueT8687 subcloned into pJEM15, Kan^r	This study
pJEM8990	*Bam*HI / *Kpn*I fragment from pBlueT8990 subcloned into pJEM15, Kan^r	This study

* Amp^r, Ampicillin resistance; Kan^r, Kanamycin resistance

Figure 4.1. Generation of the pJEM-promoter clones. Intergenic regions were PCR amplified from *M. tuberculosis* genomic DNA and cloned into the T-vector pGEM-T Easy. Inserts were subcloned through pBluescript II KS into the mycobacterial-*E. coli* promoter-probe shuttle vector pJEM15.

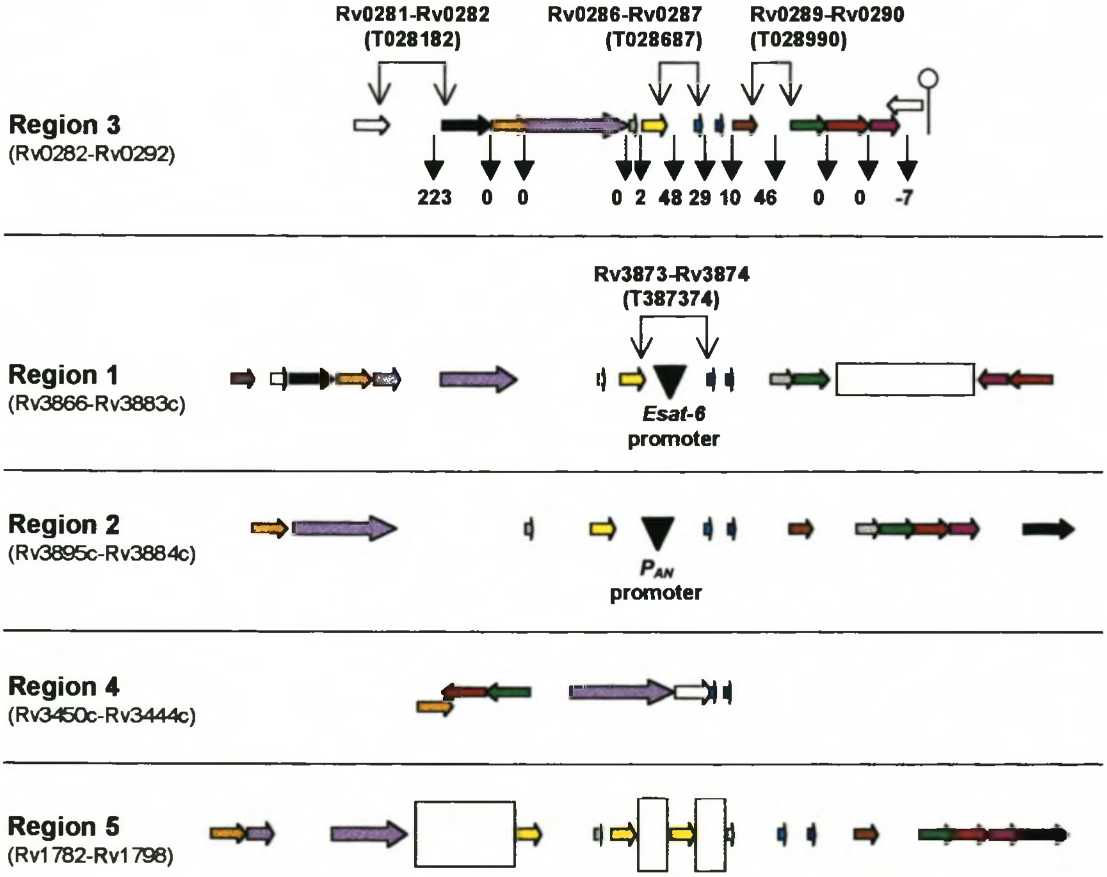


4.3. Results

4.3.1. Gene organization in the M. tuberculosis ESAT-6 gene cluster region 3

The whole genome sequence of *M. tuberculosis* H37Rv revealed that the ESAT-6 gene cluster region 3 contains 11 open reading frames (ORF's) all situated on the same strand (Figure 4.2).

Figure 4.2. Genomic organization of the *Mycobacterium tuberculosis* ESAT-6 gene clusters. Distances between genes are representative of actual intergenic distances. Intergenic distance lengths for region 3 are indicated. Large intergenic distances (containing non-conserved genes in some cases) are indicated by a boxed space. Homologous genes from different clusters are shown in the same colours, with non-conserved genes indicated in white. Promoters that were previously identified experimentally are indicated by a black triangle (Murray *et al.*, 1992, Berthet *et al.*, 1998). A putative Rho-dependent terminator downstream of region 3 is indicated by a stem-loop structure. Intergenic regions from region 1 (T387374) and 3 (T028182, T028687 and T028990) that were cloned in this study are indicated by arrows.

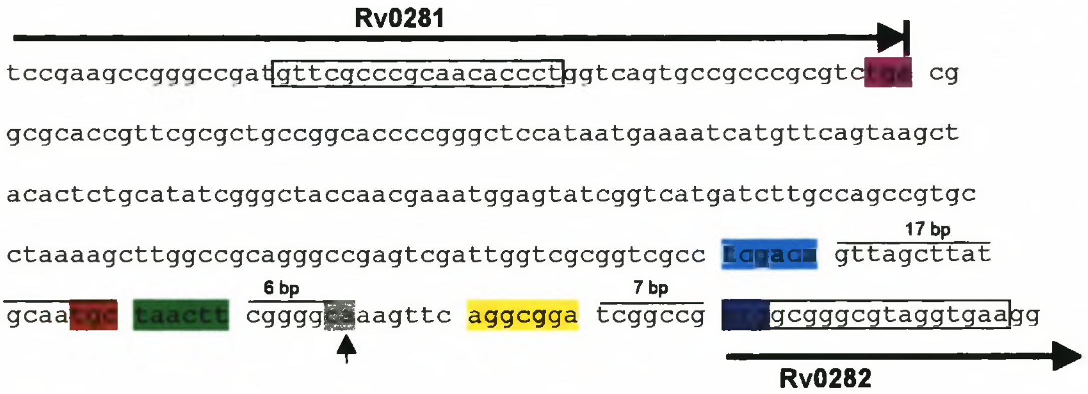


A visual analysis of the juxtaposition of these 11 genes strongly suggests that they are co-transcribed and thus constitute a single ESAT-6 gene cluster operon (Figure 4.3). There are little or no intercistronic regions between the 11 ORF's (the largest being 48 bp), but a 223 bp region in which no ORF could be detected, was found situated upstream of the first gene Rv0282. The length of this region and the fact that it is situated upstream of the first gene of the gene cluster supported the results obtained by an *in silico* analysis of the region and which revealed the presence of a putative promoter sequence (Figure 4.4).

Figure 4.3. Intergenic distances and Stop/Start codons for genes in ESAT-6 gene cluster region 3. Start codons are indicated in bold, stop codons of the previous genes are underlined and putative ribosome binding sites are highlighted

Gene	Overlap of stop/start codons	Preceding intergenic region size	Position of start codon	Position of stop codon	Protein length
Rv0282	AGGCGGATCGGCCG **ATG** GCGGGC	223 bp	342130 bp	344023 bp	631 aa
Rv0283	TCGGTGCGGGC **ATG**A CGAACCAG	0 bp	344022 bp	345636 bp	538 aa
Rv0284	GAGGGAGGGTACCG **GTG**A GCAGA	0 bp	345635 bp	349625 bp	1330 aa
Rv0285	AGGGGAGTCAGTC **ATG**A CGTTGC	0 bp	349624 bp	349930 bp	102 aa
Rv0286	TGGGCGGC TGA GC **ATG** GCCGC	2 bp	349935 bp	351474 bp	513 aa
Rv0287	AG TAA CCGAATTCCGAATCACGT GGACCCGTACGGGTCGAAAGGAGAG ATGTT **ATG** AGCCTTTTGGATGCT	48 bp	35 1525 bp	351816 bp	97 aa
Rv0288	GGTTC TGA TCGAACCCTGCTGAC CGAGAGGACTTGTG **ATG** TCGCAA	29 bp	351848 bp	352136 bp	96 aa
Rv0289	GC TAG CTCGCGCTAC **ATG** GAT	10 bp	352149 bp	353034 bp	295 aa
Rv0290	CG TAA TCAGAAACCAGAAAGTGA GCACGATGTCCCAGGAACGGTCCCG CTG **ATG** TCCGGCACCGTCATGCA	46 bp	353083 bp	354499 bp	472 aa
Rv0291	GGTGCTCAACAG **ATG**A TCCGTGC	0 bp	354498 bp	355881 bp	461 aa
Rv0292	GGAGCCCACCGA **ATG**A ACCCGAT	0 bp	355880 bp	356873 bp	331 aa

Figure 4.4. Putative region 3 promoter element. Nucleotide sequence of the 223 bp intergenic region between the open reading frames Rv0281 and Rv0282, indicating the motifs of the putative ESAT-6 gene cluster region 3 promoter. The annotated positions of the genes Rv0281 and Rv0282 are indicated by horizontal arrows. Vertical arrow indicates putative transcription initiation site. The T028182 primers used for RT-PCR and cloning of the region are boxed. The *E. coli* consensus sequence motifs with inter-motif distances and percentages of base occurrence are included at the bottom of the figure.


Rv0281
tccgaagccgggccgatgttcgcccgcaacaccctggtcagtgccgcccgcgtctga cg
gcgcaccgttcgcgctgccggcaccccgggctccataatgaaaatcatgttcagtaagct
acactctgcatatcgggctaccaacgaaatggagtatcggtcatgatcttgccagccgtgc
17 bp
ctaaaagcttggccgcagggccgagtcgattggtcgcggtcgcc tcgaca gttagcttat
6 bp
7 bp
gcaatgc taactt cggggcaaagttc aggcgga tcggccg atggcgggcgtaggtgaagg
Rv0282

🟪 Stop codon for Rv0281

🟦 Start codon for Rv0282

🟦 Putative –35 motif (homologous to –35 region of P_{AN} promoter)

🟥 Putative extended -10 motif (TGn)

🟩 Putative -10 motif (Pribnow box)

▨ Putative transcriptional start site

🟨 Putative ribosome binding site (RBS or Shine-Delgarno sequence)

Motif	-35		-10		+1	RBS		Start codon
% of occurrence	85 83 81 61 69 52		89 89 50 65 65 100		42 78			
E. coli consensus sequence:	T T G A C A	16 - 19 bp	T A T A A T	4 - 9 bp	C A/G	AGGAGG	4 - 7 bp	ATG

The proposed promoter of region 3 was named P_{ESREG3}, the **ES**AT-6 gene cluster **reg**ion **3** promoter. The promoter-specific motifs of P_{ESREG3} (-35 sequence, -10 sequence, transcriptional start site), as well as the intermotif distances are typical of a mycobacterial promoter (Mulder *et al.*, 1997). There is a putative ribosome binding site identified seven basepairs upstream of the annotated start codon of Rv0282. In addition to this, a signature (TGC) characteristic of an extended -10 motif (TGn) was found situated directly upstream from the putative -10 motif. This motif is found predominantly in promoters that have the ability to still function in the absence of its -35 region (Kenney and Churchward, 1996). In the case were the -35 region is deleted, the RNA polymerase typically binds to the extended -10 region alone and is able to initiate transcription from this point.

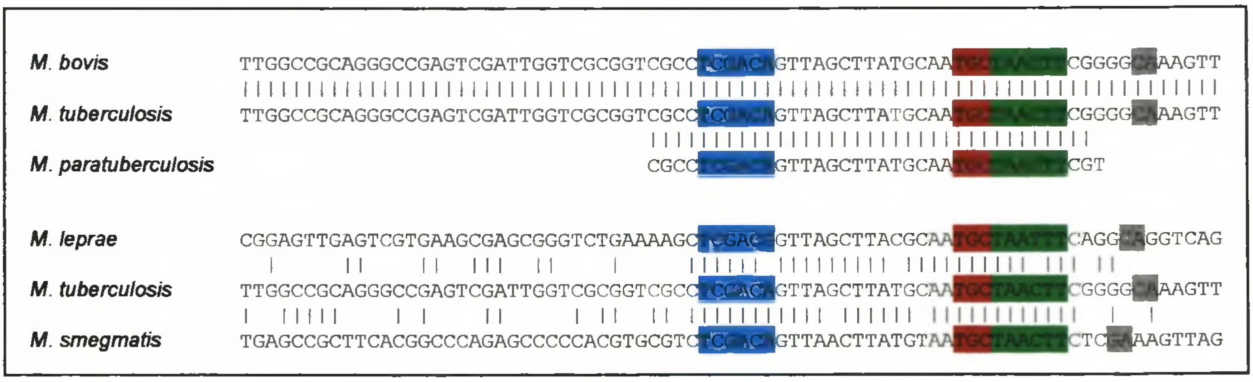
Only two promoters have been previously identified in the ESAT-6 gene clusters, the P_{AN} promoter situated in the ESAT-6 gene cluster region 2 (Murray *et al.*, 1992, Genbank accession no AJ250015), and the *esat-6* promoter from region 1 (Berthet *et al.*, 1998). An alignment of the motif sequences of the two previously identified promoters with that of the putative region 3 promoter revealed that P_{ESREG3} seem to be closest in homology to the P_{AN} promoter from region 2, with a striking 100% sequence homology found in the -35 region (Figure 4.5).

Figure 4.5. ESAT-6 gene cluster promoter alignment. Alignment of the *esat-6* promoter from the ESAT-6 gene cluster region 1 and the P_{AN} promoter of the ESAT-6 gene cluster region 2 with the putative P_{ESREG3} promoter from ESAT-6 gene cluster region 3.

	-35		-10		+1		RBS		Start codon
E. coli **consensus**									
	TTGACA	16-19	TATAAT	4-9	CG/A		AGGAGG	4-7	ATG
ESAT-6 gene clusters									
Region 1 (*esat-6*)	AGGACG	15	TAATGA	8	CT	52	GAGAGA	12	ATG
Region 2 (P_{AN})	TCGACA	17	TACACT	7	CA	37	AAGGAG	8	GTG
Region 3 (P_{ESREG3})	TCGACA	17	TAACTT	6	CA	6	AGGCGG	7	ATG

The 223 bp region in *M. tuberculosis* H37Rv containing P_{ESREG3} was also aligned to the same intergenic region (orthologous to Rv0281-Rv0282) from other mycobacteria. The sequences of the corresponding *M. tuberculosis* CDC1551, *M. tuberculosis* 210, and *M. bovis* regions were exactly the same as that of *M. tuberculosis* H37Rv (Figure 4.7). In the more distantly-related mycobacteria, *M. leprae* and *M. smegmatis*, the region upstream from the putative promoter is unconserved. This is in contrast to the sequence surrounding the putative P_{ESREG3} promoter region, which shows a high percentage of homology to that of *M. tuberculosis* H37Rv in both organisms (Figure 4.7). Although only limited sequence was available for the *M. paratuberculosis* region, a part of the region corresponding to the putative promoter region was identified and also showed a high percentage of homology to the *M. tuberculosis* region. The inability to extract any surrounding sequences from this region in *M. paratuberculosis* is indicative of a low level of homology of the rest of the region, in agreement with what was observed in *M. leprae* and *M. smegmatis*. The exclusive conservation of this part of the intergenic region supports the hypothesis that this region is important for the transcription mechanisms and that it contains the P_{ESREG3} promoter for the ESAT-6 gene cluster region 3.

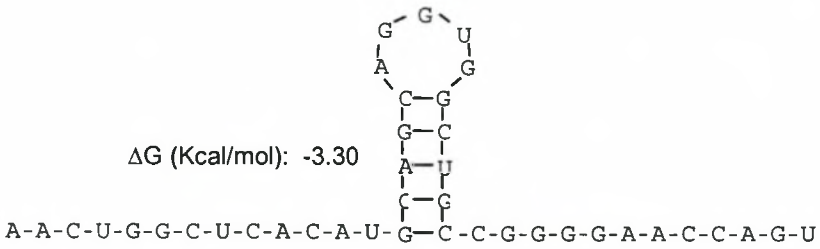
Figure 4.7. Multiple sequence alignment of the ESAT-6 gene cluster region 3 putative P_{ESREG3} promoter regions from different mycobacteria. Putative -35 sequence is highlighted in blue, extended -10 sequence in red, -10 sequence in green and transcriptional start site in grey.



```
M. bovis               TTGGCCGCAGGGCCGAGTCGATTGGTCGCGGTCGCCTCGACAGTTAGCTTATGCAATGCTAACTTCGGGGCAAAGTT
M. tuberculosis        TTGGCCGCAGGGCCGAGTCGATTGGTCGCGGTCGCCTCGACAGTTAGCTTATGCAATGCTAACTTCGGGGCAAAGTT
M. paratuberculosis                                        CGCCTCGACAGTTAGCTTATGCAATGCTAACTTCGT

M. leprae              CGGAGTTGAGTCGTGAAGCGAGCGGGTCTGAAAAGCTCGAGGGTTAGCTTACGCAATGCTAATTTCAGGCAGGTCAG
M. tuberculosis        TTGGCCGCAGGGCCGAGTCGATTGGTCGCGGTCGCCTCGACAGTTAGCTTATGCAATGCTAACTTCGGGGCAAAGTT
M. smegmatis           TGAGCCGCTTCACGGCCCAGAGCCCCCACGTGCGTCTCGACAGTTAACTTATGTAATGCTAACTTCTCGAAAGTTAG
```

An analysis of all the intergenic regions of ESAT-6 gene cluster region 3 revealed that there are no clearly identifiable transcriptional terminators or attenuators present within this region. This includes the upstream 223 bp intergenic region containing the putative promoter for the region (between Rv0281 and Rv0282). An RNA secondary structure possibly representing a weak rho-dependent terminator-like structure could be identified at a position 51 bp after the stop codon of the last gene (Rv0292) in the cluster (Figure 4.4), which may represent the termination point of this region.

Figure 4.4. Primary structure of the putative ESAT-6 gene cluster region 3 terminator. Stem-loop structure predicted from inspection of the sequence in the region downstream from Rv0292. Also shown is the free energy of secondary structure formation, ΔG.

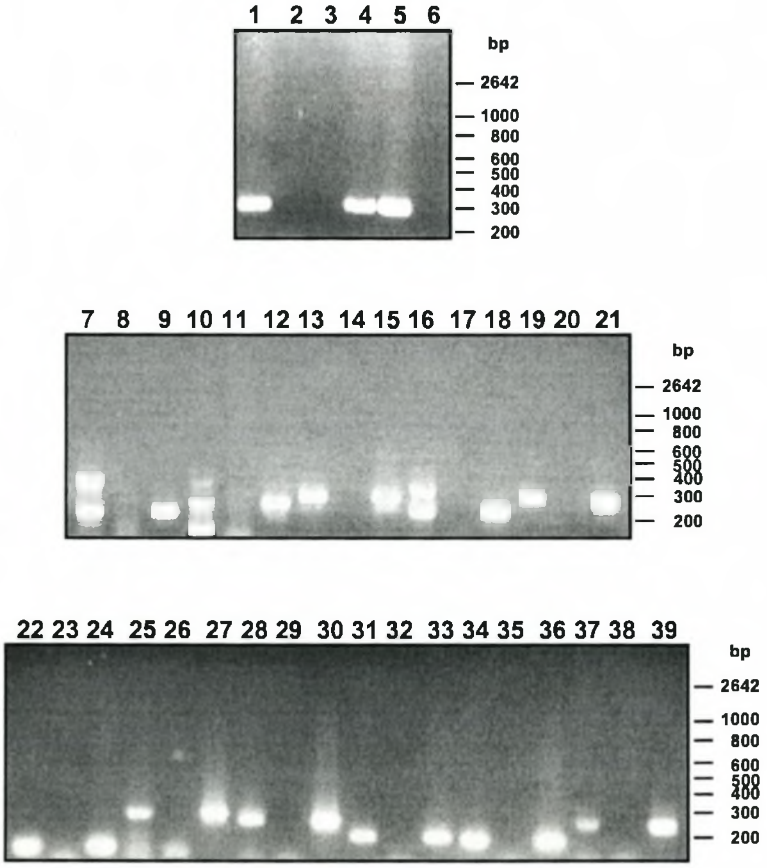


```
Rv0292
GTCGAACA TGA GGCCCTGCAGGAACACGGTCATCCGCCGCAGATA
GTCCAACTGGCTCACAT GCAGCaggtgGCTGC CGGGGAACCAGTG
                  putative terminator
```

4.3.2. Operon analysis

RT-PCR analysis of the intergenic regions between the adjacent open reading frames of the ESAT-6 gene cluster region 3 suggested that the whole region is expressed as one single polycistronic RNA of at least 14 743 bp (Figure 4.8). This result was not due to DNA contamination as all the necessary controls were present during analyses (see lanes 1-6, Figure 4.8). In three cases (T028283, T028384 and T028586), more than one amplification product was observed after the RT-PCR reaction (lanes 7, 10 and 16). These products were isolated and sequenced to confirm the amplification of the respective intergenic regions. In all three cases the presence of an amplicon corresponding to the intergenic region sequence was confirmed, with the other products being the result of non-specific priming (results not shown).

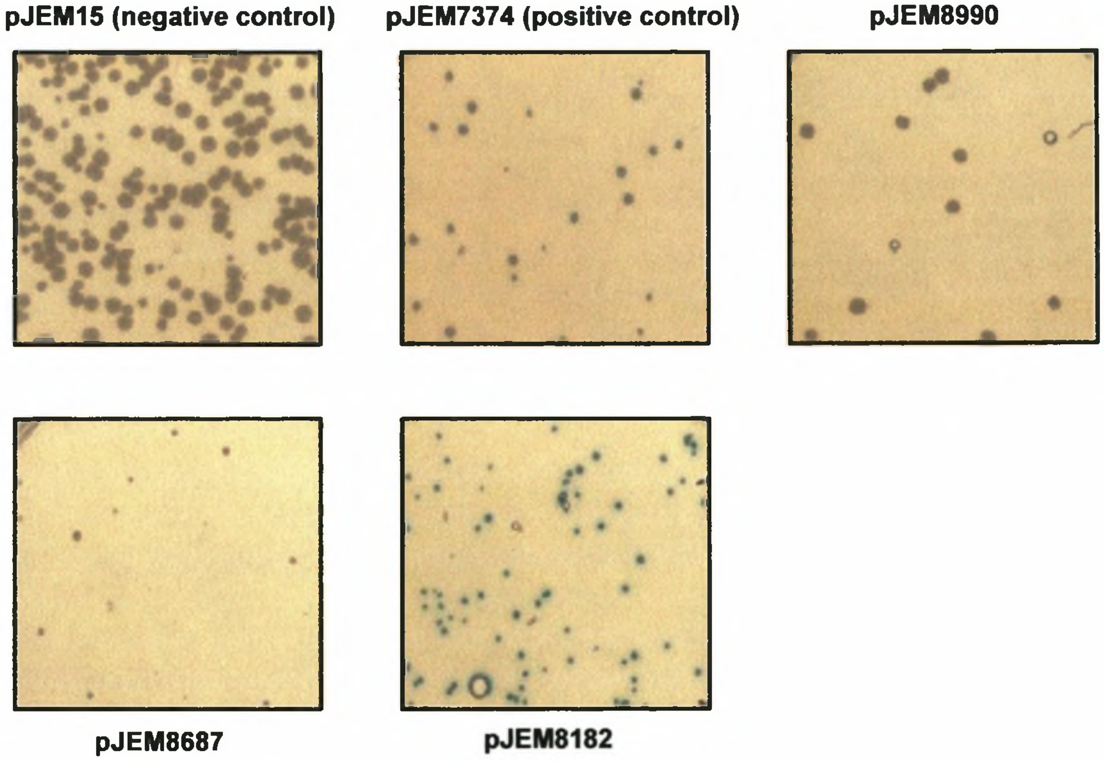
Figure 4.8. RT-PCR results of analysis of intergenic regions of the ESAT-6 gene cluster region 3. RT-PCR was done on *M. tuberculosis* RNA with primers spanning the intergenic regions of the ESAT-6 gene cluster region 3. Positive and negative controls for Reverse Transcriptase activity, DNA contamination, and DNA Polymerase activity were included and amplicons was separated on a 1.7% agarose gel. Lanes: 1, RT-PCR positive control; 2, H_2O negative control; 3, Inactivated reverse transcriptase negative control; 4, DNAseI positive control; 5, *M. tuberculosis* H37Rv DNA positive control; 6, *M. tuberculosis* H37Rv DNA + DNAseI negative control; 7, T028283 RT-PCR; 8, T028283 H_2O; 9, T028283 PCR; 10, T028384 RT-PCR; 11 T028384 H_2O; 12, T028384 PCR; 13, T028485 RT-PCR; 14, T028485 H_2O; 15, T028485 PCR; 16, T028586 RT-PCR; 17, T028586 H_2O; 18, T028586 PCR; 19, T028687 RT-PCR; 20, T028687 H_2O; 21, T028687 PCR; 22, T028788 RT-PCR; 23, T028788 H_2O; 24, T028788 PCR; 25, T028889 RT-PCR; 26, T028889 H_2O; 27, T028889 PCR; 28, T028990 RT-PCR; 29, T028990 H_2O; 30, T028990 PCR; 31, T029091 RT-PCR; 32, T029091 H_2O; 33, T029091 PCR; 34, T029192 RT-PCR; 35, T029192 H_2O; 36, T029192 PCR; 37, T387374 RT-PCR; 38, T387374 H_2O; 39, T387374 PCR.



4.3.3. Promoter analysis

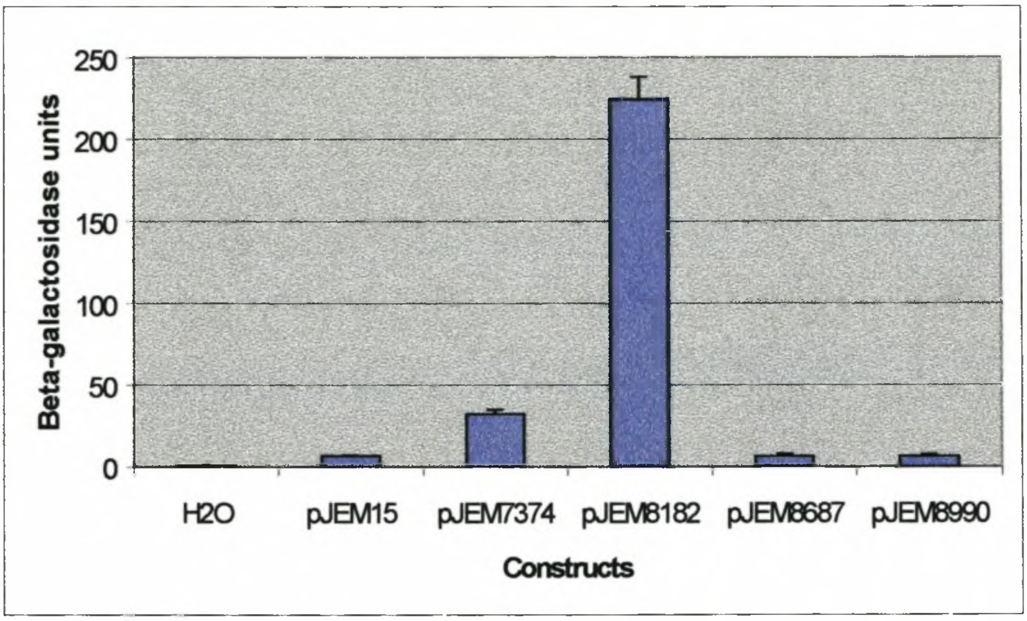
To identify the promoter elements driving the expression of the *M. tuberculosis* ESAT-6 gene cluster region 3 operon; we cloned selected intergenic regions (T028182, T028687, T028990, and T387374, see Table 4.2, Table 4.4, Figure 4.1 and Figure 4.2) into a β-galactosidase promoter-probe vector pJEM15. The intergenic region T028182 was selected for the analysis because of the identification of the putative P_{ESREG3} promoter in this region (described in Section 4.3.1). T028687 was selected because it is the largest intergenic region (48 bp, see Table 4.3) within this cluster. This region is also situated upstream of the *esat-6* gene family member Rv0287, and thus in the same position as both the *east-6* promoter from region 1 and the P_{AN} promoter from region 2 (Figure 4.2, Berthet *et al.*, 1998, Murray *et al.*, 1992, Genbank accession no AJ250015). T028990 was selected because it is the second largest intergenic region (46 bp, Table 4.3), and also the only other intergenic region large enough to have the potential to contain a promoter sequence. Lastly, the intergenic region T387374 was selected as a positive control for the presence of an ESAT-6 gene cluster region promoter, because it contains the previously identified *esat-6* promoter from ESAT-6 gene cluster region 1 (Berthet *et al.*, 1998). Promoter constructs were transformed into *E. coli* and plated on LB-Kanamycin plates containing X-Gal. As was observed by Timm *et al.* (1994), all transformants (including *E. coli* transformed with the promoterless vector pJEM15) were blue on LB-Kanamycin-X-Gal plates, indicating a higher copy number of the vector when inserted in *E. coli* (results not shown). This also confirms the conclusion reached by Timm and coworkers, which states that blue-white screening should be done directly in a mycobacterial host. When the same constructs were subsequently electroporated into *M. smegmatis* and plated onto Middlebrook 7H11-Kanamycin-X-Gal plates, white colonies were obtained in the case of pJEM15, pJEM8687 and pJEM8990, while light blue colonies were obtained with pJEM7374 and dark blue colonies with pJEM8182 (Figure 4.9). This indicates the presence of a strong promoter in construct pJEM8182, while no promoter activity could be detected in the intergenic regions of T028687 and T028990. The promoter activity of pJEM8182 also seemed to be much higher than that observed in the control construct pJEM7374.

Figure 4.9. Promoter constructs electroporated in *M. smegmatis* and plated on X-Gal plates. Promoter constructs were electroporated into *M. smegmatis* and plated onto Middlebrook 7H11-Kanamycin-X-Gal plates. White colonies were obtained in the case of the negative control pJEM15, and constructs pJEM8687 and pJEM8990, while light blue colonies were obtained with the positive control pJEM7374 and dark blue colonies with pJEM8182. This indicates the presence of a strong promoter in construct pJEM8182 (much stronger than that observed in the control construct pJEM7374), while no promoter activity could be detected in the intergenic regions of T028687 and T028990.



To quantitatively determine the amount of β-galactosidase activity, we used the previously described ONPG β-galactosidase spectrophotometric assay (Pardee *et al.*, 1959, Figure 4.10). The results clearly demonstrate that the intergenic region T028182 contains a very strong promoter activity (more than 7 times stronger than the control *esat-6* promoter region, T387374). None of the other regions showed any promoter activity, having values exactly the same as the promoterless vector pJEM15. This result confirms what was observed on the X-Gal plates as well as the sequence analyses of the ESAT-6 gene cluster region 3.

Figure 4.10. β-Galactosidase activities of *M. smegmatis* clones containing different intergenic regions from the ESAT-6 gene cluster region 3. Results are means of three independent experiments and standard deviations are indicated.



Construct	Units of β-Galactosidase activity detected in sonicated cell extracts of transformed *M. smegmatis**
H_2O	0 ± 0
pJEM15	6 ± 0
pJEM7374	32 ± 3
pJEM8182	224 ± 14
pJEM8687	6 ± 1
pJEM8990	6 ± 1

*Results are the averages and standard deviations for three independent experiments.

4.4. Discussion

Although a great number of mycobacterial antigens have been identified to date, the function of most of these proteins remains unknown. The 23 members of the potently immunogenic ESAT-6 antigen family are an example of this. Elucidation of the intracellular functions of these antigens and its contribution to the overall pathogenicity and immunogenicity profile of the organism may help us to obtain a better understanding of the processes involved in causing disease. As a first step to determining the function, an in-depth analysis of the regulation of expression and the genetic milieu of the genes encoding these antigens must be performed. In this study the potential polycistronic nature of the previously identified ESAT-6 gene clusters was investigated (by focusing on the gene cluster region 3) as this may reveal distinct information on the regulation and expression of these important antigens and the operons that they are situated in and may provide some clues to their intracellular function.

Experimental evidence pointing to a close relationship between the genes situated in the ESAT-6 gene clusters has been provided by several independent studies. Berthet and coworkers (1998) have shown that the *esat-6* and *lhp* genes from the ESAT-6 gene cluster region 1 are transcribed as a single polycistronic RNA and thus constitute an operon. In addition, Wards and coworkers (2000) have found that a knockout of Rv3871 upstream of the *lhp* (Rv3874) and *esat-6* (Rv3875) genes in region 1 resulted in a mutant showing a similar loss of virulence in guinea pigs as a *lhp/esat-6* knockout. Interestingly, this mutant also caused a negative ESAT-6 skin test reaction result in the guinea pigs, providing evidence that the knockout might have had a polar effect on the downstream *esat-6* gene. Further evidence indicating a close relationship between different parts of the ESAT-6 gene cluster region 1 comes as a result of the deletion of a certain part of this region in *M. bovis* BCG (the RD1 deletion region). Although both the ESAT-6 gene cluster region 1 genes Rv3881c and Rv3883c (mycosin-1) are present in the genome of this organism neither of them are expressed in *M. bovis* BCG (Brown *et al.*, 2000, Mattow *et al.*, 2001). As the Rv3881c and mycosin-1 genes lie 2 and 4 ORF's downstream of the RD1 deletion region, it indicates that there are certain factors present within this deleted region that are needed for their expression. This is supported by the fact that the ESAT-6 gene cluster region 2 mycosin-2 gene (Rv3886c), lying only 3 ORF's downstream of mycosin-1, are expressed efficiently in BCG (Brown *et al.*, 2000). All this experimental evidence supports the genome sequence analysis, which suggests that these regions may be expressed as one or more operons.

The results of this study showed that the ESAT-6 gene cluster region 3 is expressed in its entirety as one single operon. Although we have not yet performed these experiments on the other ESAT-6 gene clusters, there is no reason to believe that these duplicate clusters are not also expressed as one or more operons. As a parts of region 1 and region 4, respectively, are encoded on the opposite strand, these two regions would definitely be transcribed as more than one operon. It is clear from the relative positions of the genes present within these clusters that gene rearrangements have taken place during or after the duplication events, but it is known that gene order does not necessarily influence expression and function of similar well-investigated operon-like structures, for example that of certain lantibiotic operons (Siezen *et al.*, 1996).

The absence of identifiable terminators in region 3 is in agreement with what is observed for the whole genome sequence of *M. tuberculosis*. Using terminator-identifying computer algorithms, only a few possible terminators were identified in intergenic regions of the whole genome sequence (data not shown). It thus seems as if the genome sequence of *M. tuberculosis* is deficient in ordinary terminators having the consensus *E. coli* transcriptional termination motifs (*E. coli* transcriptional terminators are described in detail by d'Aubenton Carafa *et al.*, 1990). The reason for the absence of identifiable terminator sequences may be the G+C richness of the genome of *M. tuberculosis* (65.9%, Cole *et al.*, 1998) as one of the major features of ordinary *E. coli* rho-independent terminators is the use of a very conserved stretch of thymidines at the 3' end of the terminator (d'Aubenton Carafa *et al.*, 1990). In the putative terminator sequence identified from this study (Figure 4.4), we could only find a 4 bp stretch of guanosines at the 3' end of the terminator. The difference due to the G+C richness of the genome affects codon usage and promoter recognition sites (Mulder *et al.*, 1997) and is also represented in other DNA and RNA structural motifs. An example of this is the positional base preference in codons where there is an increase observed in the third position G or C. Another example is the wide range of promoter sequences in the mycobacteria, which differ quite considerably with the consensus *E. coli* sequence, and are much more reliant on G and C bases. The transcriptional start site that is most frequently an A in the case of *E. coli*, is also more frequently a G (in 48% of cases) in the mycobacterial sequences (Timm *et al.*, 1999). All of this indicates a difference in the mycobacterial transcription and translation mechanisms in comparison to other bacteria, which are most probably also reflected in the terminator sequences and termination mechanisms. The absence of strong, clearly identifiable consensus terminator motifs in the mycobacteria may indicate a low level of specific termination of transcription, the presence of as yet unidentified mycobacterial-specific G+C rich terminator sequences, or that the mycobacteria may rely more on rho-type protein dependent transcriptional termination. Although no transcriptional terminator structures were identified inside the region 3 operon, Berthet and coworkers (1998) did find a structure similar to a Rho-independent transcription terminator 40bp downstream from the stop codon of the *esat-6* gene in ESAT-6 gene cluster region 1. These authors speculated that there would be no more additional genes downstream of the *esat-6* gene that could form part of the *esat-6/cfp-10* operon (Berthet *et al.*, 1998). It is well known that many gene clusters, for example those involved in the biosynthesis of lantibiotics, consist of several transcription units (Sahl and Bierbaum, 1998). This allows for the high level transcription of the mRNA of certain genes required at higher levels, while the presence of a weak terminator structure allows only low levels of readthrough to the modification, secretion and processing enzymes.

To further elucidate the regulation of expression of the ESAT-6 gene clusters, we identified the promoter involved in the expression of ESAT-6 gene cluster region 3 and named it P_{ESREG3}. This promoter was initially identified by genome sequence analysis, and was confirmed to be the (only) promoter of the ESAT-6 gene cluster region 3 by cloning the region into a mycobacterial promoter probe vector. In this study we have not investigated the possibility that promoters may be present in intracistronic regions. The presence of promoters within mycobacterial genes has not yet been shown (W. Bourn, personal communication), although Strohl (1992) found that out of 139 streptomycete (an actinomycete relative of the mycobacteria) promoters studied, 6 were situated within open reading frames. The P_{ESREG3} promoter is expressed very strongly in *M. smegmatis*, at a seven times higher level than the region 1 *esat-6* promoter identified by Berthet *et al.* (1998). The sequence of P_{ESREG3} is closest in homology to the *M. paratuberculosis* P_{AN} promoter from ESAT-6 gene cluster region 2, previously identified by Murray and coworkers (1992). The promoter probe analysis done in this study did not reveal any promoter activity in the intergenic region upstream of the *esat-6* gene in the ESAT-6 gene cluster region 3, the same position in which the *esat-6* promoter of region 1 and the P_{AN} promoter from region 2 are situated. No homology could also be found between these two promoters and the upstream sequences of the *esat-6* genes of the other two ESAT-6 gene cluster regions (data not shown).

Although no other upstream promoters for region 1 and 2 have been identified yet, it is tempting to speculate that the *esat-6* and P_{AN} promoters are secondary promoters and that primary promoters for the expression of the whole region 1 and 2 is situated upstream in the same position as the region 3 P_{ESREG3} promoter. These promoters would be most likely found situated in the large intergenic regions (91 bp and 160 bp) before Rv3865 and Rv3896c respectively. It must be noted that, although there is no doubt that there is a promoter element present in the intergenic region before the *esat-6/cfp-10* operon in ESAT-6 gene cluster region 1 (the *esat-6* promoter), a multiple sequence alignment similar to the one in Figure 4.7 revealed that the -35 and -10 regions hypothesized by Berthet and coworkers (1998) are not conserved between different mycobacteria (Figure 4.11). Also, the -35 region proposed by these authors is situated inside the gene sequence of the upstream PPE gene. It may thus be more likely that the correct position of this promoter is situated within the conserved region indicated in Figure 4.11.

In conclusion, we have shown that at least one of the gene clusters encoding the very important T-cell antigen ESAT-6 family is expressed as a single polycistronic RNA and the 11 genes situated in this cluster thus form one single operon. We have also identified the promoter for this operon, P_{ESREG3}, and characterized its activity. The identification of this promoter may provide clues to the mechanisms involved in the 64% downregulation in the attenuated *M. tuberculosis* strain H37Ra of the ESAT-6 family member named Rv0288 (TB10.4 or CFP-7, Rindi *et al.*, 1999). It may even ultimately reveal the reason for the attenuation of the organism. The verification of the polycistronic nature of the ESAT-6 gene clusters as well as the identification of the promoter P_{ESREG3} is an important step in the elucidation of the function and regulation of members of the ESAT-6 family as well as its biosynthetic gene clusters. This study contributes to the slowly growing field of research aimed at the understanding of the mycobacterial transcriptional machinery. This could ultimately provide insights into the mechanisms involved in the regulation of antigen expression and pathogenicity.

Figure 4.11. Multiple sequence alignment of the ESAT-6 gene cluster region 1 *esat-6* promoter regions from different mycobacteria. C-terminal sequence of upstream gene Rv3873 is shown in red letters and start codon of Rv3874 gene in blue. Promoter regions proposed by Berthet *et al.* (1998) is highlighted: -35 in blue, -10 in green, transcriptional start sites in grey and ribosome binding site in yellow.

```
M. smegmatis      AGGATGAAGACGACGAGTGGTGA     ACACTCGCACAAGAACTTTCCGGT GCACTCGCCGGAA
                  | || ||  |  || ||||||||     || |     ||  ||||| ||||  || |  |||||||
M. leprae         ATGACGACCAGCACAACTGGTGA   ACACCCAAAATCACAAACTTCCCGGC GCCCGGGCCGGAAA
                        |   |    ||||||||    |    |  |  ||| ||||||||||  ||||||||||||
M. tuberculosis   GGACGAAGAGGACG ACTGGTGAGCTCCCGTAATGACAACAGACTTCCCGGCCACCCGGGCCGGAA


                  GACTCGCCATGGAATTGGTGAGGA CACAGGGAAATAAGGGGAAAT   CCA  ATG
                  |||| ||||      ||| ||||| ||    |||| |  |   ||    |||  |||
                  GACTTGCCAGCATC TGGCGAGTAACAGCAAGAAAGAGAGTAGAA    CCAACATG
                  ||||||||| |||  ||||||| ||       |||||||| |       ||| ||||
                  GACTTGCCAACATTTTGGCGAGGAA    GGTAAAGAGAGAAAGTAGT CCAGCATG
```

CHAPTER FIVE

MULTI-COMPONENT PROTEIN TRANSPORT SYSTEM

"He smiled and said 'Sir, does your mother know that you are out?'"

Misadventures at Margate – Rev. R.H. Barham (1788-1845)

NOTE: The results presented in the following chapter will be submitted for peer review and publication as: "**The ESAT-6 gene cluster of *Mycobacterium tuberculosis* forms a multi-component protein transport system for the secretion of members of the immunologically important ESAT-6 and CFP-10 multigene families.** Gey van Pittius, N.C., Daugelat, S., Warren, R.M., Kaufmann, S.H.E., and Van Helden, P.D."

5.1. Introduction

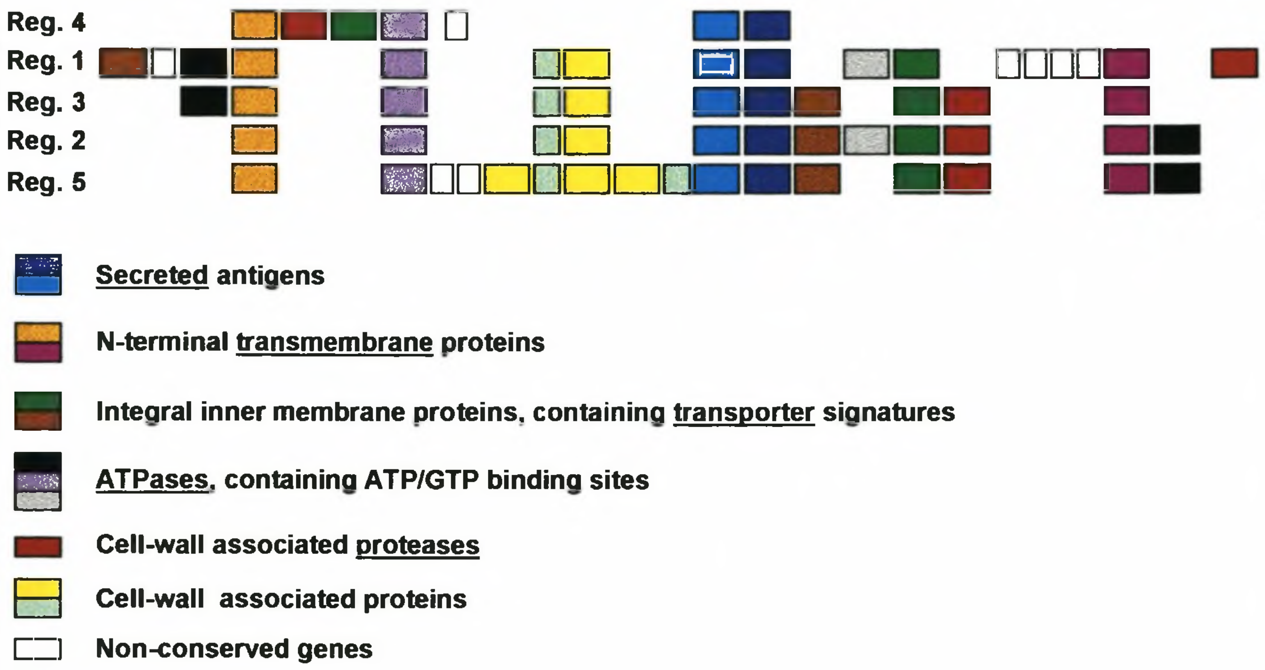
Approximately 20% of proteins synthesized by bacteria are trafficked to the cell membrane and extracellular environment and most of these exogenous proteins are transported via the general secretory pathway (Pugsley, 1993). This pathway makes use of the Sec protein-translocation system (Mori and Ito, 2001) as a first step to translocate the proteins across the cytoplasmic membrane. In Gram-positive bacteria, these translocated proteins are directly released into the extracellular milieu, while in Gram-negative bacteria a second transport mechanism, termed a terminal branch, is required to translocate the protein across the outer membrane (Pugsley, 1993). Protein secretion systems have been grouped into five major pathways, which have recently been the subject of a number of excellent reviews and will thus not be covered in this paper. These pathways are named the type I (signal sequence independent ATP-binding cassette [ABC] transporter pathway, Linton and Higgins, 1998), II (main terminal branch of the general secretion pathway, Sandkvist, 2001), III (contact-dependent secretion pathway, Plano *et al.*, 2001), IV (conjugal transfer system pathway, Christie, 2001), and V (autotransporter pathway, Jacob-Dubuisson *et al.*, 2001) secretion systems.

Although the mycobacteria are classified as Gram-positive organisms, they exhibit a highly complex cell wall structure displaying an exceptionally low permeability due to of the presence of an unusual layer of lipid (mycolate esters) (Brennan and Nikaido, 1995, Daffé and Draper, 1998, Barry, 2001a). All members of the genus *Mycobacterium*, examined so far, appear to have similar cell wall structures, although the permeability of these walls varies widely between species (Barry and Mdluli, 1996). Only a small number of studies have been done on secretion in the mycobacteria and little is known about how transport occurs through the complex cell wall and membrane (Braunstein and Belisle, 2000). The genome of *M. tuberculosis* contains all the genes necessary for an efficient Sec translocation system and although a large number of *M. tuberculosis* secreted proteins make use of the *sec*-dependent secretion pathway (predicted to be more than 700 proteins, Braunstein and Belisle, 2000), there are certain important proteins found in the culture filtrates that do not contain the ordinary N-terminal secretion signals (Daffé and Etienne, 1999, Gomez *et al.*, 2000, Braunstein and Belisle, 2000). This observation has led to the conclusion that *M. tuberculosis* is able to transport proteins independently of the general secretory pathway (Harth and Horwitz, 1997). These proteins, which include superoxide dismutase, alanine dehydrogenase, glutamine synthetase, alcohol dehydrogenase, thioredoxine (Daffé and Etienne, 1999) and the multiple members of the ESAT-6 family (Skjøt *et al*., 2001), are thus secreted by other, still unknown mechanisms.

Only as recently as 1997 did Barsom and coworkers describe the first protein-dependent ATP-binding cassette (ABC) transport system in mycobacteria, consisting of four closely-spaced open reading frames adjacent to the *mpr* gene (encoding a mycobacteriophage resistance protein) of *M. smegmatis*. The sequencing of the whole genome of *M. tuberculosis* (Cole *et al*., 1998) revealed a number of other genes potentially encoding the typical subunits of the ABC transporters (nucleotide binding domains, membrane-spanning domains and substrate binding proteins), leading to the identification of at least 26 complete and 11 incomplete ABC transporters (Braibant *et al*., 2000). ABC transporters transport various molecules (which includes ions, amino acids, peptides, proteins, antibiotics, polysaccharides, etc.) into (importers) and out of (exporters) bacterial cells. In prokaryotes these transporters are encoded by different genes organized in one or more operons, which are clustered in one DNA region on the chromosome. The proteins carrying the nucleotide binding domains of the transporters provide the energy to the active transport process by binding and hydrolyzing ATP (Braibant *et al*., 2000). These proteins contain conserved motifs namely the Walker$_A$ (PS00017 ATP/GTP-binding site motif A (P-loop)) and Walker$_B$ motifs (which together form an ATP binding pocket), as well as an ABC transporter family signature situated between the Walker motifs. The proteins containing the membrane-spanning domains consist of four to eight transmembrane α-helices that form a channel through which the substrate is transported. These proteins contain one conserved motif named the EAA loop motif (or binding protein-dependent transport systems inner membrane component signature). The last component of the ABC transporter is called the substrate binding protein and is usually more commonly found in importers (transporters that transport proteins into the cell).

The structure and organization of the ESAT-6 gene clusters have been described in detail previously (Gey van Pittius *et al*., 2001). A study of the potential functions of the proteins encoded by these clusters shows that most of these proteins have a potential to function in a protein-dependent ATP-binding cassette active transport system (Figure 5.1, Addendum 3A). These genes encode proteins containing Walker motif ATP binding sites, transporter family signatures as well as binding protein dependent transport systems inner membrane component signatures in combination with 11 transmembrane α-helices that could form a large channel through which the substrate could be transported.

Figure 5.1. Organization of the genes and potential functions of the encoded proteins situated within the ESAT-6 gene cluster regions.


Reg. 4
Reg. 1
Reg. 3
Reg. 2
Reg. 5
Secreted antigens
N-terminal transmembrane proteins
Integral inner membrane proteins, containing transporter signatures
ATPases, containing ATP/GTP binding sites
Cell-wall associated proteases
Cell-wall associated proteins
Non-conserved genes

It has been proposed previously that these gene clusters are responsible for the secretion of the ESAT-6 protein family members, explaining the absence of any ordinary *sec*-dependent secretion signals in the amino acid sequences of members of this family (Tekaia *et al.*, 1999, Gey van Pittius *et al.*, 2001). Thus, each of the duplicated ESAT-6 gene clusters may be a regulon that encodes components of a unique ABC transporter type structure that has not been identified previously in the mycobacteria. In the present study we adopted a novel approach to attempt to answer the question of whether the proteins encoded by the ESAT-6 gene clusters (Gey van Pittius *et al.*, 2001) function together as a mycobacterial membrane-bound complex involved in protein-dependent transport and if so, whether this transport system is responsible for the active secretion of the ESAT-6 protein family members.

5.2. Materials and Methods

5.2.1. *DNA and protein sequence analyses*

All DNA and protein sequence information for *M. tuberculosis* H37Rv as well as *M. smegmatis* mc^{2}155 was obtained from publicly available finished and unfinished genome sequence databases at the Pasteur Institute and the Institute for Genomic Research (TIGR) websites (http://genolist.pasteur.fr/TubercuList/, for *M. tuberculosis* H37Rv and http://www.tigr.org/cgi-bin/BlastSearch/blast.cgi?organism=m_smegmatis for *M. smegmatis*), as well as from previously published sequence analyses (Tekaia *et al.*, 1999, Gey van Pittius *et al.*, 2001).

5.2.2. *Bacterial strains*

Escherichia coli JM109 was used as a host for propagation of cosmids and plasmids. *Mycobacterium smegmatis* mc^{2}155 (Snapper *et al.*, 1990) was used as a heterologous mycobacterial host for transformation of plasmids and recombinant cosmid integration, as well as for the expression of genes and subsequent protein secretion studies.

5.2.3. *Media and culture conditions*

E. coli was grown on solid or in liquid Luria-Bertani (LB) medium as described by Sambrook *et al.* (1989). For RNA extraction, *M. smegmatis* was grown at 37°C for 2 days with shaking (200 rpm) in Middlebrook 7H9 broth (Difco) supplemented with filter-sterile ADC (0.5% BSA, 0.2% glucose, 0.015% catalase) and containing 0.05% Tween 80 (Sigma). For transformant selection on solid media, *M. smegmatis* was grown on Middlebrook 7H11 agar supplemented with filter-sterile OADC (0.005% oleic acid, 0.5% BSA, 0.2% glucose, 0.02% catalase, 0.085% NaCl) and containing 0.05% Tween 80 (Sigma). To obtain culture filtrate proteins in detergent/protein-free culture media for secretion analyses, different recombinant *M. smegmatis* clones were grown at 37°C in 150 ml Kirchner's broth (3 g/l Na_2HPO_4, 4 g/l KH_2PO_4, 1.07 g $MgSO_4.7H_2O$, 2.5 g/l Tri-sodium citrate, 20% glycerol, 5 g/l asparagine) to an optical density of 0.3 at 600 nm (early log phase). Hygromycin (50 μg/ml, Roche), Kanamycin (50 μg/ml, Roche) and/or Ampicillin (50 μg/ml, Roche) were added to bacterial cultures when antibiotic selection was required. Whenever mycobacteria were double transformed (integrating cosmid and plasmid), the cultures were grown in both Hygromycin and Kanamycin to select for both transforming structures.

5.2.4. RNA preparation

All procedures involving RNA were performed under the strictest RNase-free conditions. Total RNA was extracted from *M. smegmatis* cultures essentially as described in Chapter 4.2.5. Briefly, cells were pelleted, Trizol (GibcoBRL) was immediately added and transferred to FastPrep tubes containing silicone beads (Bio101). Cells were ribolyzed and after centrifugation the supernatant was removed from the beads/cell debris and transferred to a 1.5 ml microfuge tube containing 500 μl chloroform. The tube was vortexed, and placed on ice for 2 minutes with intermittent vortexing, after which it was centrifuged again and the aqueous phase was transferred to a 1.5 ml microfuge tube containing 500 μl isopropanol. After an overnight incubation at -20°C, the RNA was pelleted by centrifugation, washed with 70% ethanol and resuspended in nuclease-free H_2O (200 μl per 20 ml volume of original culture) containing DNAseI buffer, 1 U/μl RNasin and 10 mM DTT. This was incubated on ice for 30 minutes after which DNAseI was added at 0.1 U/μl and incubated at 37°C for a further 30 minutes. Stop solution was added (20 μl per 200 μl), and the RNA was purified using the QIAGEN RNeasy mini-kit, as described by the manufacturer. The integrity of the RNA was assayed on a 1 % TAE agarose gel. RNA stocks were stored in single-use aliquots at -20°C containing 1 U/μl RNasin and 10 mM DTT. RNA was confirmed to be totally free of any DNA contamination using PCR analysis excluding a reverse transcriptase step.

5.2.5. RT-PCR analysis

RT-PCR reactions were performed using the Roche Titan One-Tube RT-PCR System kit in combination with the Promega HotstarTaq PCR System, as described by the manufacturers. Two separate master mixes were prepared, and only mixed together immediately prior to cycling. Master mix 1 contained the RNA template, dNTP's (200 μM final concentration), primers (50 pmol each), DTT (10 mM final concentration), RNasin (1 U/μl) and water to a final volume of 25 μl. Master mix 2 contained 5X RT-PCR buffer (with $MgCl_2$ to a final concentration of 1.5 mM), Reverse Transcriptase enzyme mix (1 μl), HotstarTaq PCR enzyme mix (0.2 μl), Q-buffer (10 μl) and water to a final volume of 25 μl. Reverse transcription and cycling were performed without interruption in a Perkin Elmer

GeneAmp 2400 under the conditions indicated in Chapter 4 (Table 4.3). Results of RT-PCR reactions were visualized on a 2 % TAE agarose gel.

5.2.6. Primers

Table 5.1 contains a list of all oligonucleotide primers used in this study for RT-PCR, PCR for cosmid isolation confirmation as well as PCR for radioactively labeled probe amplification. Primers were chosen according to length, Tm, G+C content and length of the product, to efficiently amplify the selected gene region. To avoid problems with RT-PCR reactions due to length restrictions of products, all RT-PCR primers were chosen to result in a product less than 300 bp.

Table 5.1. List of RT-PCR and PCR oligonucleotide primers

Name of primer	Primer sequence (from 5' to 3')	Length of primer	Tm * [°C]	G+C (%)	Length of product	Application
Rv3866 f	cgtcatggtgcgcttcgt	18 bp	58	61.1	201 bp	PCR confirmation of cosmid isolation
Rv3866 r	gcggttgtgcattcggcta	19 bp	60	57.9		
MycP1 f	tgacgttgaccgcatagt	18 bp	54	50.0	230 bp	PCR confirmation of cosmid isolation
MycP1 r	ctgctctcgctacgtcag	18 bp	58	61.1		
S0288 f	ccagatcatgtacaactacccg	22 bp	66	50.0	240 bp	RT-PCR
S0288 r	gccatggtgttctgctcgt	19 bp	60	57.9		
Sesat f	gtatggaatttcgccggtatc	21 bp	62	47.6	213 bp	RT-PCR
Sesat r	ggtctgggcgaggttctgc	19 bp	64	68.4		
SrpoB f	tggcggcgatcaaggagt	18 bp	58	61.1	157 bp	RT-PCR
SrpoB r	tgcacgtcgcggacctcga	19 bp	64	68.4		
ESAT-6 f	agcagcagtggaatttcgc	19 bp	58	52.6	270 bp	Probe for cosmid isolation and PCR confirmation
ESAT-6 r	tcccagtgacgttgccttc	19 bp	60	57.9		

*Tm were calculated using the following formula: [4x (G+C)] + [2x (A+T)].

5.2.7. DNA manipulations for secretion analysis

All DNA manipulations were performed essentially as described by Sambrook *et al.* (1989). Details of the cosmid and plasmid constructs are presented in Table 5.2 and Figure 5.2. The *M. tuberculosis* H37Rv genomic DNA integrating cosmid library (Bange *et al.*, 1999) was a gift from F.-C. Bange (Medizinische Hochschule, Hannover, Germany). This library was constructed by cloning the

M. tuberculosis H37Rv Sau3A genomic DNA fragments (approximately 40 000 kb) into the BcII site of the vector pYUB412 (an *E.coli*-mycobacterial shuttle vector containing two selectable markers allowing for selection on Hygromycin- and Ampicillin-containing media). This vector also contains an integrase gene and integrates stably into the *attB* site in the genome of mycobacteria (Bange *et al.*, 1999).

Plasmid pMB154, an *E. coli*-mycobacterial shuttle expression vector containing the *M. tuberculosis* ESAT-6 gene cluster region 1-specific ESAT-6 protein gene (Rv3875) was originally constructed by M. Braunstein (Albert Einstein College of Medicine, New York, USA, unpublished data) and was a gift from S. Daugelat (Max-Planck-Institut für Infektionsbiologie, Berlin, Germany). This construct was used for the expression of recombinant HA-tagged ESAT-6 protein in mycobacteria. The C-terminal HA epitope included in this vector makes use of *M. tuberculosis*-specific codons. pMB154 was constructed by the insertion of the ESAT-6 gene in frame into the plasmid pSD21, which was used as a control plasmid and were also a gift from S. Daugelat.

Figure 5.2. Plasmid and cosmid constructs used in this study.

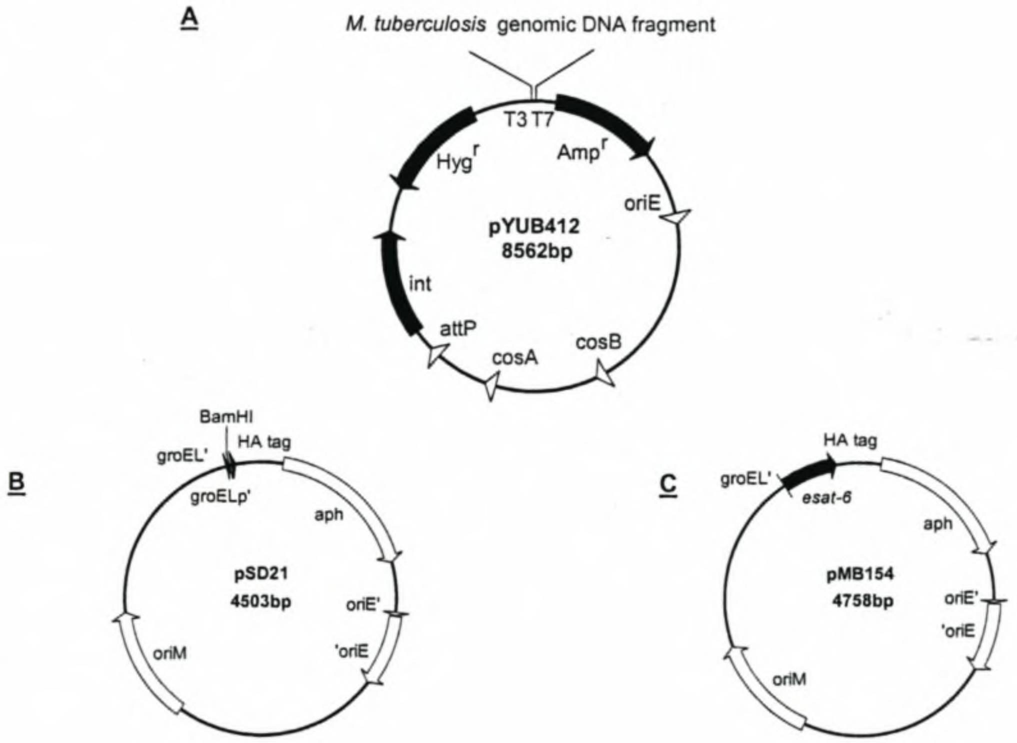


Table 5.2. Plasmids and cosmid constructs used in this study

Plasmids and cosmids	**Characteristics***	**Source or reference**
pYUB412-derived *M. tuberculosis* H37Rv genomic DNA cosmid library	Single copy integrating *E. coli*-mycobacterial shuttle vector library containing *M. tuberculosis* genomic DNA fragment inserts of approximately 40 000 bp, Hyg^r and Amp^r	F.-C. Bange (Bange *et al.*, 1999)
pSD21	*E. coli*-mycobacterial shuttle expression cloning vector for expression of C-terminal HA-tagged recombinant proteins, Kan^r	S. Daugelat
pMB154	*E. coli*-mycobacterial shuttle expression vector for expression of the C-terminally HA-tagged *M. tuberculosis* ESAT-6 gene cluster region 1-specific ESAT-6 protein, Kan^r	S. Daugelat (constructed by M. Braunstein)
cosmid H1(0)	Single copy integrating cosmid isolated from *M. tuberculosis* H37Rv genomic DNA library. Contains DNA fragment including the complete ESAT-6 gene cluster region 1, Hyg^r and Amp^r	This study

5.2.8. Isolation of cosmids containing selected genomic regions

Cosmids containing selected genomic regions were isolated from the *M. tuberculosis* H37Rv library by colony blotting, as described by Sambrook *et al.* (1989). Briefly, approximately 2000 clones of the *M. tuberculosis* H37Rv genomic DNA cosmid library were spread onto LB plates containing Ampicillin (50 µg/ml, Roche) and allowed to grow at 37°C for 16 h. Bacterial colonies were transferred to Hybond-N$^+$ membrane filters (Amersham) by placing the membrane onto the colonies and incubating for one minute. The membrane was then subjected to subsequent steps of denaturing, neutralization, and washing, after which the filters were baked for 2 hours at 80°C in a vacuum oven. After baking, the membranes were washed and prehybridized overnight at 42°C, after which it was subjected to hybridization at 42°C using a radioactively labeled probe. Washing, prehybridization and hybridization was performed essentially as described by Sambrook *et al.* (1989). During hybridization, membranes were probed using a [α-^{32}P] dCTP-labeled probe complementary to an internal region of the ESAT-6 gene sequence (Rv3875, for probing for ESAT-6 gene cluster region 1). Labeling of the radioactive probe was done by using the commercial Prime-It RmT Random Primer Labeling Kit (Stratagene) according to the manufacturer's instructions. Labeled probe was purified from unincorporated radioactive nucleotides using a G50M Sephadex desalting spin column. The positive clones were visualized by autoradiography, and corresponding colonies were subsequently picked (as the cosmid sizes are around 40 000 kb, positive clones were obtained with a frequency of ± 1:300). Positive clones were confirmed by PCR of the gene used as probe (ESAT-6 from region1), as well as the genes present in the first and last positions of the cluster (Rv3866 and Rv3883c). PCR was done using the HotStar Taq (Promega) PCR system according to the manufacturer's conditions. Finally, the isolated cosmids were sequenced on an automated sequencer using T3 and T7 sequencing primers to confirm correct DNA region as well as to determine exact start and stop of insert DNA.

5.2.9. Transformation of M. smegmatis

Transformation *of M. smegmatis* with the plasmid and cosmid constructs was done using electroporation, as described previously (Jacobs *et al.*, 1991). As the isolated integrating cosmid containing the selected *M. tuberculosis* H37Rv DNA sequence were integrated stably into the genome of *M. smegmatis* upon transformation, these recombinant cells could be retransformed with an episomally-replicating plasmid (pSD21 or pMB154) without loss of either construct. Transformants were selected on Hygromycin, Kanamycin or Hygromycin/Kanamycin-containing Middlebrook 7H11 agar plates, depending on whether transformation was done with cosmid, plasmid, or both. All work on recombinant *M. smegmatis* mc^{2}155 transformed with the single copy vector *M. tuberculosis* H37Rv genomic DNA cosmid library was performed under Biosafety Level II conditions, as was recommended by Bange *et al.* (1999).

5.2.10. Protein secretion analyses

Recombinant *M. smegmatis* clones were grown under conditions as specified in 5.2.3. For protein secretion assays in liquid media, mycobacterial culture cells were pelleted by centrifugation at 4500 x g and supernatant containing culture medium and secreted proteins were removed by aspiration. Cells were resuspended in 500 μl phosphate-buffered saline (PBS) and sonicated at 4.5 setting in a Misonix cup sonicator on ice for a total of 5 minutes (15-second bursts with 30-second intervals), whereafter insoluble debris was pelleted to obtain whole cell lysate proteins. Supernatants from both the culture filtrate and whole cell lysate were filter-sterilized by serial filtration through 1.0 μM, 0.45 μM and 0.22 μM filters, and culture filtrate proteins were concentrated using the 3 kDa cutoff Centriprep 3 centrifuge concentration system (Centricon). When grown to an OD_{600} = 0.3, 150 ml of culture filtrate produced approximately 0.5 - 1 mg of concentrated culture filtrate proteins.

Culture filtrate and whole cell lysate protein concentrations were accurately determined spectrophotometrically using the BioRad protein concentration determination assay, according to the manufacturer's instructions. Low molecular weight proteins was separated on a three-layered Tris/Tricine/SDS-PAGE gel (16% resolving gel, 10% spacer gel and 4% stack gel) according to the method of Schagger and von Jagow (1987). 10 μg of each sample was loaded into each well. *M. tuberculosis* H37Rv culture filtrate proteins was a gift from C. Pheiffer (University of Stellenbosch, Tygerberg, South Africa) and was used as a positive control of native ESAT-6 expression and secretion. Purified recombinant His-tagged ESAT-6 protein was used as a second positive control and was obtained from Dr. J.T. Belisle (Department of Microbiology, College of Veterinary Medicine and Biomedical Science, Colorado State University, USA). The protein was produced and provided through funds from the National Institutes of Health, National Institute of Allergy and Infectious Diseases, Contract No1-AI-75320, entitled "Tuberculosis Research Materials and Vaccine Testing."

Purified recombinant dimer ESAT-6 protein was used as a third positive control and was a gift from R. Skjøt and P. Andersen (Statens Serum Institut, Copenhagen, Denmark). Mouse anti-ESAT-6 monoclonal antibodies (HYB 76-8) were used at a dilution of 1/25 to detect the presence of *M. tuberculosis* ESAT-6 protein in Western blotting analyses, and was a gift from I. Rosenkrands (Statens Serum Institut, Copenhagen, Denmark). Mouse anti-HA monoclonal antibodies (HA.11, Clone 16B12, Covance) were used at a dilution of 1/2000 to detect HA-tagged recombinantly-expressed ESAT-6 protein (expressed from pMB154) in *M. smegmatis* fractions by Western blotting. Anti-HA antibodies have a high specificity to allow unambigious identification of the influenza hemagglutinin epitope (YPYDVPDYA). Horse radish peroxidase (HRPO)-conjugated goat anti-mouse antibodies (Caltag Laboratories) were used as secondary antibody in Western blotting at a concentration of 1/10 000.

5.3. Results

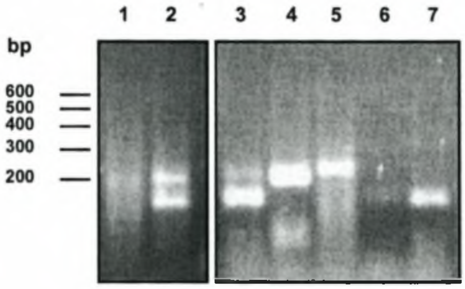
5.3.1. Sequence analyses

To determine the potential secretion function of the ESAT-6 gene clusters, a suitable mycobacterial host had to be identified. This host had to be a member of the genus *Mycobacterium*, as the cell wall structure of the mycobacteria is unique to the genus (Barry and Mdluli, 1996). Thus, another bacterial host with a different cell wall structure would not necessarily display the same secretion characteristics. The avirulent, fast-growing species *M. smegmatis* can be transformed with a high efficiency and has been used widely in the study of *M. tuberculosis* proteins (Snapper *et al.*, 1990). To ascertain whether *M. smegmatis* was a suitable host to study the ESAT-6 gene cluster secretion system, it was essential to determine whether the genome of this organism contains any copies of the ESAT-6 gene cluster regions which could influence the results obtained with the transfomed *M. tuberculosis*-specific regions. The results from the whole genome sequence analyses indicated that the genome of *M. smegmatis* contains three of the ESAT-6 gene clusters (region 1, 3 and 4), which confirmed previously described results (see Chapter 3, Gey van Pittius *et al.*, 2001). These gene clusters displayed the same gene organization as in *M. tuberculosis*, with a high percentage of homology shared between the gene sequences of the two organisms (data not shown).

5.3.2. RT-PCR analysis

To determine if the genes present in the ESAT-6 gene clusters of *M. smegmatis* are expressed in this organism, a RT-PCR analysis was done. This analysis was done on *M. smegmatis* RNA to determine expression of the genes *Sesat* (the *M. smegmatis* region 1 ESAT-6 orthologue MS3875) and *S0288* (the *M. smegmatis* region 3 TB10.4 orthologue MS0288), with the *M. smegmatis* copy of the *rpoB* gene (*SrpoB*) as positive control. The results show clearly that both of these genes are efficiently expressed in *M. smegmatis*, indicating that the gene cluster regions 1 and 3 are functional in this organism (Figure 5.3).

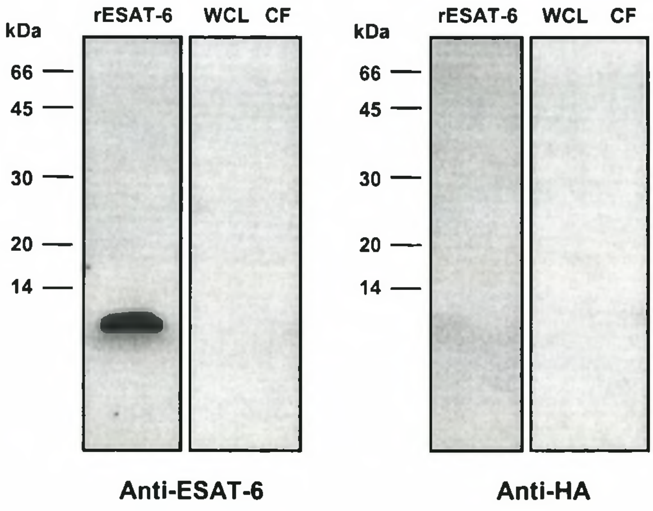
Figure 5.3. RT-PCR analysis of *M. smegmatis* ESAT-6 gene cluster region 1 and 3 expression. Analysis was done on *M. smegmatis* RNA to determine expression of the *M. smegmatis* region 1and 3 genes *Sesat* and *S0288* with the *M. smegmatis* copy of the *rpoB* gene (*SrpoB*) as positive control. Results indicate that both of these genes are efficiently expressed in *M. smegmatis*. Lanes: (1) *SrpoB* RT-PCR H_2O control for DNA contamination, (2) *M. smegmatis* DNA *SrpoB* positive control, (3) *SrpoB* RT-PCR, (4) *Sesat* RT-PCR, (5) *S0288* RT-PCR, (6) *SrpoB* RT inactivated control, (7) *SrpoB* DNAse positve control.



5.3.3. Cross-reactivity analysis of anti-ESAT-6 antibodies

In a final attempt to evaluate the suitability of *M. smegmatis* as a potential mycobacterial host for the secretion studies, a Western blot analysis was done with wild-type *M. smegmatis* culture filtrate proteins as well as whole cell lysate proteins. The Western blot was probed with the anti-ESAT-6 antibodies (HYB-76-8) to determine if these antibodies could also detect *M. smegmatis* ESAT-6. No ESAT-6 protein could be detected in either the whole cell lysate or the culture filtrate of the wild-type *M. smegmatis* (Figure 5.4, lanes WCL and CF). This was not due to the antibody not working as the recombinant His-tagged ESAT-6 protein (CSU) was detected efficiently by this antibody (Figure 5.4, lane rESAT-6). The results of this analysis indicate that the anti-ESAT-6 antibodies are *M. tuberculosis* ESAT-6-specific, with no cross reactivity being observed. The same analysis was done with the anti-HA antibodies, showing that these antibodies also did not cross react to any of the proteins found in the wild-type *M. smegmatis* whole cell lysates as well as culture filtrates and could not detect ESAT-6 protein that was not HA-tagged (Figure 5.4). This result provided evidence that it was possible to use *M. smegmatis* as a host to study the *M. tuberculosis* ESAT-6 gene cluster region, as the antibodies used in this study do not cross react with the native *M. smegmatis* expressed proteins.

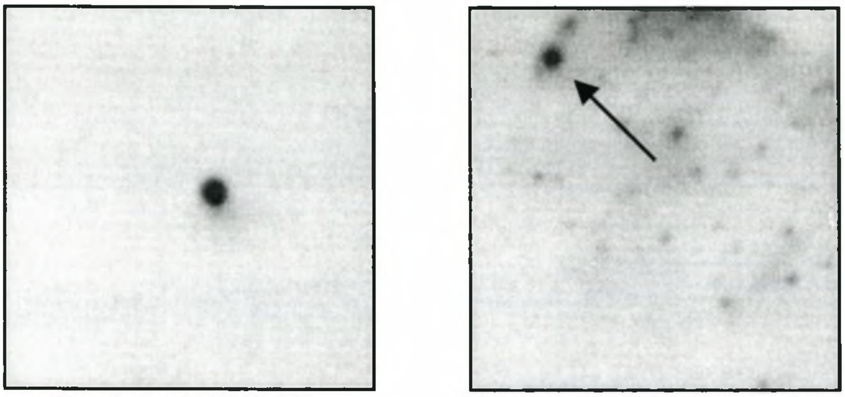
Figure 5.4. Cross reactivity analysis of anti-ESAT-6 antibodies. Western blot analysis of wild-type *M. smegmatis* whole cell lysate (WCL) as well as culture filtrate (CF) proteins was done by probing with the anti-ESAT-6 antibodies (HYB-76-8) and the anti-HA antibodies to determine if these antibodies could also detect *M. smegmatis* ESAT-6. No ESAT-6 protein could be detected in either the whole cell lysate or the culture filtrate of the wild-type *M. smegmatis* with the anti-ESAT-6 antibodies and the anti-HA antibodies also did not cross react to any of the proteins found in the wild-type *M. smegmatis* whole cell lysates as well as culture filtrates. Recombinant His-tagged ESAT-6 protein (rESAT-6) was used as a positive control.



5.3.3. Isolation of cosmids containing selected genomic regions

Cosmids containing *M. tuberculosis* H37Rv genomic DNA encoding ESAT-6 gene cluster region 1 were isolated using colony blotting (Figure 5.5).

Figure 5.5. Colony blotting of a *M. tuberculosis* H37Rv genomic DNA intergrating cosmid library. *M. tuberculosis* H37Rv genomic DNA cosmid library on LB plates were transferred to Hybond-N^+ membrane filters, and probed with a [α-^{32}P] dCTP-labeled probe complementary to an internal region of the ESAT-6 gene sequence (Rv3875). (A) *M. tuberculosis* H37Rv genomic DNA positive control membrane, (B) colony blot with positive clone indicated by an arrow.



***M. tuberculosis* H37Rv genomic DNA**

***M. tuberculosis* H37Rv genomic DNA library colony blot**

Isolated cosmids were confirmed to contain the specified region by PCR analysis of the first and last genes of the gene cluster (Figure 5.6). Final confirmation of the exact start and stop of the genome sequence fragment present in the cosmid was obtained by DNA sequencing using the T3 and T7 sequencing primers situated at the 5' and 3' ends of the insert. A cosmid containing the complete ESAT-6 gene cluster region 1 were isolated and named cosmid H1(0) (see Figure 5.7).

Figure 5.6. PCR analyses of the isolated cosmids. Isolated cosmid H1(0) were confirmed to contain the specified region by PCR analysis of the first (Rv3866) and last (Rv3883c) genes of the gene cluster region 1. Lanes (1) Rv3866 PCR H_2O control, (2) Rv3866 PCR positive control with *M. tuberculosis* H37Rv DNA, (3) Rv3866 PCR with cosmid H1(0), (4) Rv3883c PCR H_2O control, (5) Rv3883c PCR positive control with *M. tuberculosis* H37Rv DNA, (6) Rv3883c PCR with cosmid H1(0).

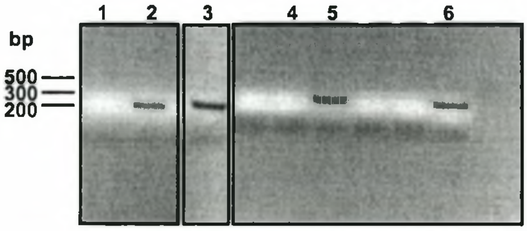
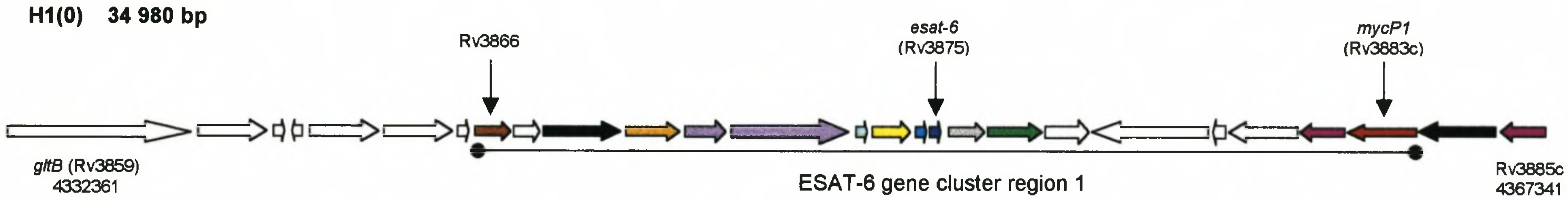


Figure 5.7. Genes present within isolated cosmid H1(0) insert. The *M. tuberculosis* H37Rv annotated ORF numbers and position in the whole genome sequence (as determined by T3 and T7 sequencing) is shown at the termini of the insert. Genes confirmed to be present by PCR are indicated by arrows. The ESAT-6 gene cluster region 1 is indicated by a horizontal line.



5.3.4. Secretion analyses

To determine if wild-type *M. smegmatis* are able to secrete recombinant *M. tuberculosis* ESAT-6 into the culture filtrate, pMB154 (expressing HA-tagged ESAT-6) was transformed into *M. smegmatis* mc^{2}155. Whole cell lysates and culture filtrate proteins were Western blotted and probed with anti-HA antibodies. The results of this analysis (shown in Figure 5.8, lanes 3 and 4, as well as in schematic format in Figure 5.9A) indicate that although very high levels of ESAT-6-HA are expressed in recombinant *M. smegmatis* whole cell lysate, no ESAT-6-HA protein could be detected in the culture filtrate. This was confirmed using the anti-ESAT-6 antibodies (Figure 5.8, lanes 3 and 4). In agreement to what was observed in the cross reactivity analysis, no *M. smegmatis*-specific ESAT-6 protein could be detected in wild-type *M. smegmatis* whole cell lysate or culture filtrate. This result confirms that the *M. smegmatis*-specific ESAT-6 gene cluster region 1 present within the genome of *M. smegmatis* is not able to secrete the *M. tuberculosis*-specific ESAT-6 protein, making *M. smegmatis* a suitable host for the analysis of the secretion of *M. tuberculosis* ESAT-6.

To analyse the secretion of native *M. tuberculosis* ESAT-6 protein in *M. smegmatis*, the cosmid H1(0), containing the complete ESAT-6 gene cluster region 1, was stably integrated into the genome of *M. smegmatis* mc^{2}155. Whole cell lysates and culture filtrate proteins were Western blotted and probed with anti-HA as well as anti-ESAT-6 antibodies. As would be expected, no protein could be detected using the anti-HA antibodies (Figure 5.8, lanes 5 and 6). However, using the anti-ESAT-6 antibodies, high levels of ESAT-6 protein were detected in the whole cell lysate as well as culture filtrate of H1(0) cosmid-transformed *M. smegmatis* (Figure 5.8, lanes 5 and 6, Figure 5.9B).

To determine whether the *M. smegmatis* genome-integrated *M. tuberculosis* ESAT-6 gene cluster region 1 encodes proteins which are able to cause the secretion of ESAT-6 from the cells in trans, pMB154 was transformed into *M. smegmatis* transformed with cosmid H1(0). This resulted in the efficient secretion of high levels of recombinant ESAT-6-HA-tagged protein (Figure 5.8, lanes 7 and 8). This result clearly demonstrates that ESAT-6 is secreted into the culture medium only in the presence of the ESAT-6 gene cluster region 1 from *M. tuberculosis* (Figure 5.9C), indicating that there are certain factors present within this genomic region that are essential for the efficient translocation of the ESAT-6 protein across the complex mycobacterial cell wall. It furthermore indicates that the orthologous region 1 in the genome of *M. smegmatis* is not able to perform the same function when an *M. tuberculosis*-specific ESAT-6 protein is used, indicating a high specificity of protein sequence recognition and secretion.

Figure 5.8. Secretion of *M. tuberculosis* H37Rv ESAT-6 in recombinant *M. smegmatis*. *M. smegmatis* was transformed with pMB154 (expressing HA-tagged ESAT-6) and the cosmid H1(0) (containing the complete ESAT-6 gene cluster region 1), respectively. As H1(0) integrated stably into the genome, this transformant could be double transformed with pMB154. Whole cell lysates and culture filtrate proteins of each of the three transformants were Western blotted and probed with anti-HA as well as anti-ESAT-6 antibodies. Wild-type *M. smegmatis* whole cell lysates and culture filtrates were used as a negative control and *M. tuberculosis* H37Rv culture filtrate were used as a positive control. Lanes (1) Wild-type *M. smegmatis* whole cell lysate, (2) Wild-type *M. smegmatis* culture filtrate, (3) pMB154 transformed *M. smegmatis* whole cell lysate, (4) pMB154 transformed *M. smegmatis* culture filtrate, (5) H1(0) transformed *M. smegmatis* whole cell lysate, (6) H1(0) transformed *M. smegmatis* culture filtrate, (7) H1(0)+pMB154 transformed *M. smegmatis* whole cell lysate, (8) H1(0)+pMB154 transformed *M. smegmatis* culture filtrate, (9) *M. tuberculosis* H37Rv culture filtrate.

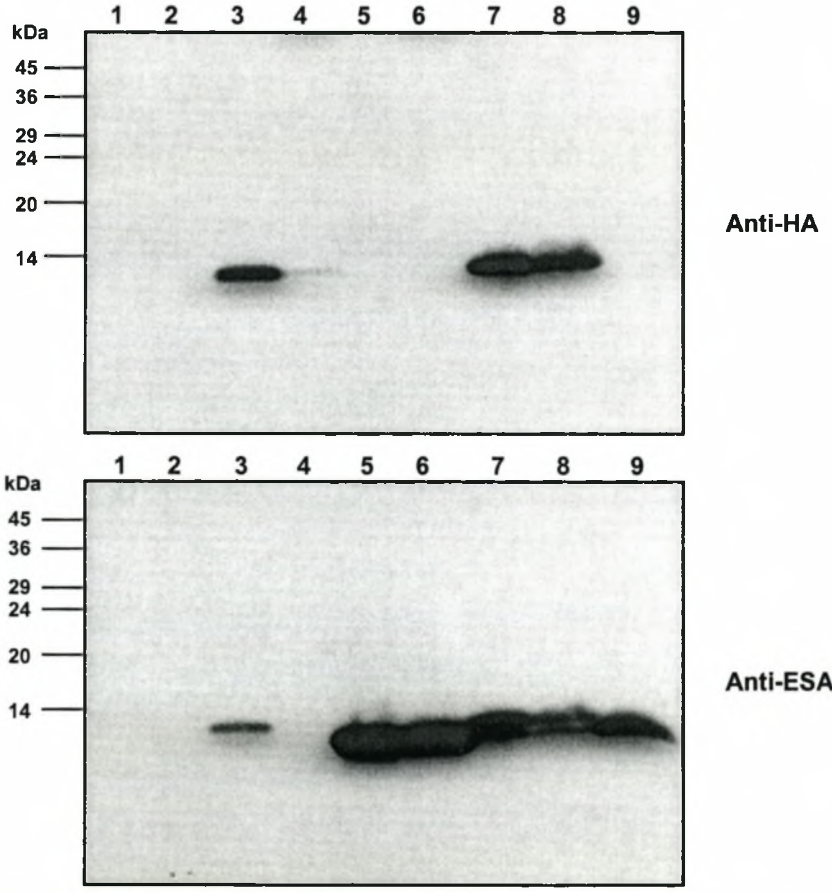
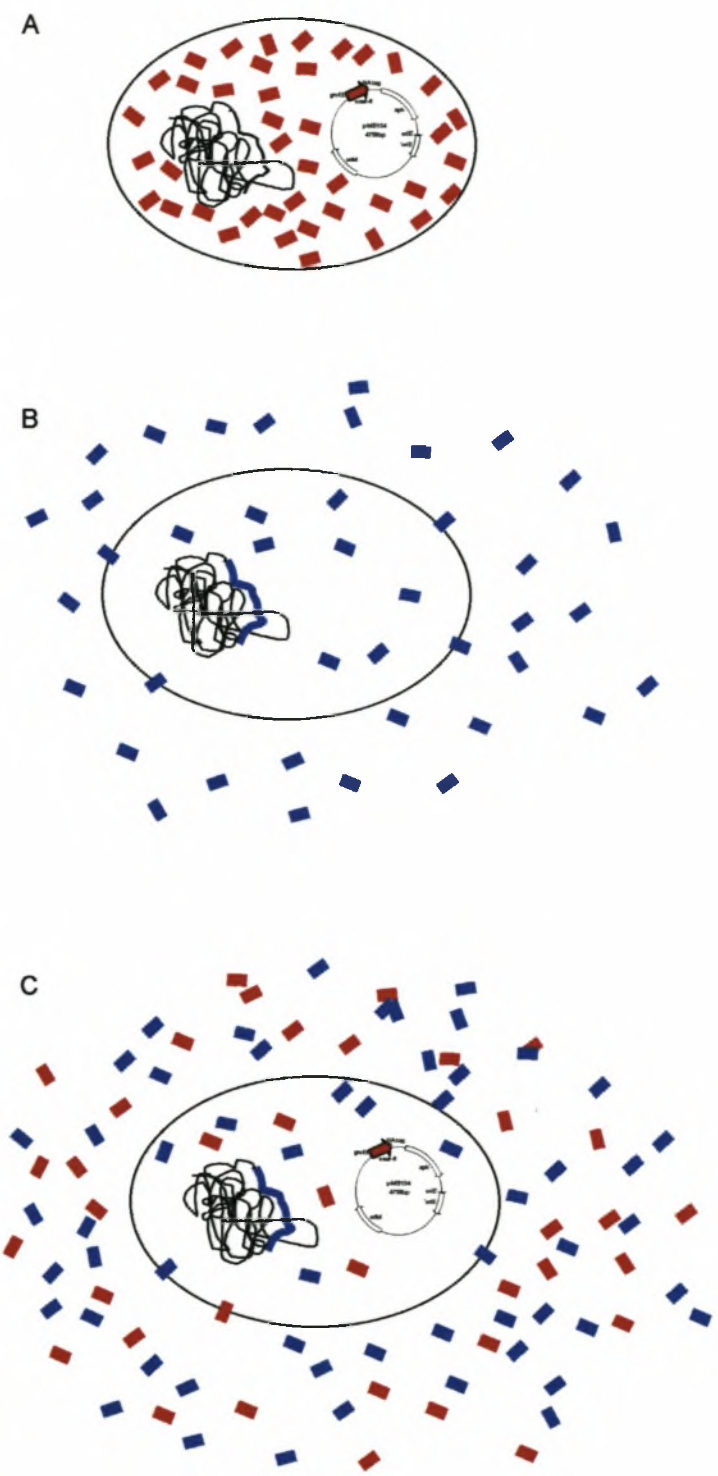


Figure 5.9. Schematic representation of the *M. tuberculosis* ESAT-6 gene cluster region 1-specific secretion of ESAT-6 protein in *M. smegmatis*.



A: *M. smegmatis* transformed with plasmid pMB154 and expressing C-terminally HA-tagged ESAT-6 protein. No ESAT-6-HA is secreted.

B: *M. smegmatis* containing stably integrated cosmid H1(0) with complete ESAT-6 gene cluster region 1 and expressing native ESAT-6 protein. High levels of ESAT-6 are secreted into the extracellular medium.

C: *M. smegmatis* containing stably integrated cosmid H1(0) with complete ESAT-6 gene cluster region 1 and expressing native ESAT-6 protein as well as transformed plasmid pMB154 expressing C-terminally HA-tagged ESAT-6 protein.. High levels of native ESAT-6 as well as recombinant HA-tagged ESAT-6 are secreted into the extracellular medium.

5.4. Discussion

Protein translocation systems are central to antigen secretion and the uptake of antibiotics, making the dearth of information available on these systems in the mycobacteria surprising. The advent of whole genome sequencing made the analysis of the coding potential of complete genomes possible. The availability of the genome sequence information also makes the identification of transporter systems much easier than before, as most of the constituent genes of these systems are encoded by one or more operons situated in the same DNA region. In this way, a wealth of information on the number of transport systems in the mycobacteria have become available since the completion of the genome sequence of *M. tuberculosis* in 1998 (Cole *et al.*, 1998).

We have previously studied the genomic organization and duplication of a cluster of putative transport associated genes surrounding the immunologically important ESAT-6 family proteins (Gey van Pittius *et al.*, 2001). An analysis of the expression of the ESAT-6 gene clusters have revealed that at least one of these clusters (region 3) is expressed as one single polycistronic RNA and thus forms an operon structure (Chapter 4). The co-expression of the genes situated within the ESAT-6 gene clusters indicated that there might be a functional relationship between the encoded proteins. The results of the present study revealed that the ESAT-6 gene cluster regions are in fact involved in the secretion of the ESAT-6 protein family members, as was demonstrated with the ESAT-6 protein Rv3875 and the ESAT-6 gene cluster region 1. This study indicated that the ESAT-6 protein is located intracellularly when expressed heterologously in *M. smegmatis* in the absence of the complete ESAT-6 gene cluster region 1, but is secreted to the extracellular milieu efficiently as soon as the whole region 1 is present. This result indicates that there are components present within the region that are able to allow the transport of the ESAT-6 protein through the membrane of the organism, proving previous hypotheses on the potential secretion function of the clusters (Tekaia *et al.*, 1999, Gey van Pittius *et al.*, 2001). This result also fits the proposed functions of the constituent genes of the clusters, as was described previously (Addendum 3A). The present study also clearly demonstrated that the secretion of the proteins is extremely sequence specific, as no interspecies secretion could be obtained although the regions were orthologs. From this result it may be hypothesized that the different regions within a specific organism may also only be able to secrete a specific sequence belonging to that region, and this is the subject of ongoing experiments.

The sequence specificity observed in the secretion system may explain the high level of sequence homology observed in the Mtb9.9 ESAT-6 subfamily (Alderson *et al.*, 2000, Gey van Pittius *et al.*, 2001, Addendum 3A). This subfamily consists of duplications of the ESAT-6 family members originating from the ESAT-6 gene cluster region 5. Members of the subfamily which do not from part of region 5 may thus be able to be recognized by the region-specific transport apparatus and may be secreted efficiently due to their extremely high level of sequence homology (Addendum 3A, Figure 3A.6). This hypothesis is supported by the fact that a study by Alderson and coworkers (2000) revealed that 83% of PPD$^+$ donors made a significant proliferative response to rMtb9.9A, indicating that it was an efficiently secreted protein. The gene for this Mtb9.9 family member is not situated in an ESAT-6 gene cluster, but is the homologue of Rv1793 (situated in the region 5 gene cluster) and is thus most probably secreted by the ESAT-6 gene cluster region 5 transporter apparatus.

Another aspect of the ESAT-6 secretion system that we are continuing to study, is the question of which of the genes in the clusters are essential for efficient secretion. As not all the cluster regions contain the same amount of genes, it is logical to assume that not all the genes would be necessary for the formation of an effective translocation apparatus. Wards and coworkers (2000) have found that a knockout of Rv3871 (situated 4 genes upstream of ESAT-6) resulted in a mutant showing a similar loss of virulence in guinea pigs as an *esat-6* knockout. Surprisingly, this mutant also did not sensitize the animals to an ESAT-6 skin test. Despite the fact that no information was given on the expression of ESAT-6 in this mutant, it is tempting to speculate that ESAT-6 was possibly expressed in the organism, but that the protein was not secreted due to the knockout of Rv3871. As Rv3871 belongs to a conserved gene family in the clusters which is hypothesized to function as ATPases, the knockout of this gene may disrupt the provision of energy to the active transport system and may thus cause the whole transport apparatus to become nonfunctional.

The presence of very high levels of ESAT-6 protein observed in the whole cell lysates in this study is in agreement with previous observations showing similar high amounts of ESAT-6 and CFP-

10 family members (Mtb9.9 subfamily, ESAT-6, CFP-10, TB10.4) inside the cells, with lower concentrations found in the culture filtrate (Sørensen *et al.*, 1995, Alderson *et al.*, 2000, Skjøt *et al.*, 2000). This observation could be an indication that either the *esat-6* genes are transcribed faster than it can be transported and thus accumulates inside the cell, or that the secretion mechanism only works optimally under *in vivo* conditions. Tekaia and coworkers (1999) suggested that the cytoplasmic accumulation observed is consistent with the existence of a dedicated secretion apparatus as intracellular stockpiling of proteins for secretion has been observed previously in other pathogens that possess type III secretion systems. Other explanations may be that the efficacy of translocation of these proteins may be less because of the absence of a consensus signal sequence (Sørensen *et al.*, 1995) or that the transport system may need some external factor to be fully activated.

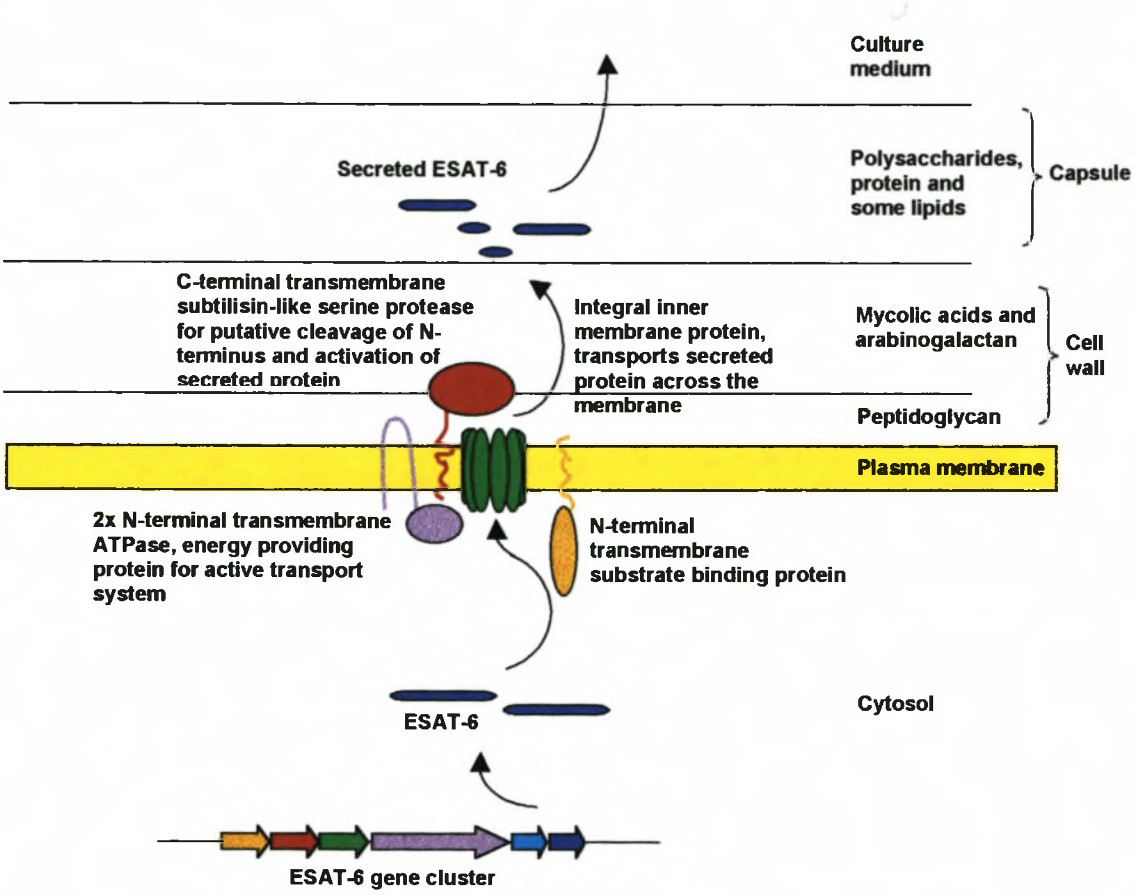
Although we concede that *M. smegmatis* may not be the most appropriate host for these studies due to the presence of three ESAT-6 gene cluster regions in the genome of this organism, we have been compelled to use this species because of the absence of another suitable bacterial host containing the same or similar cell wall structure. The feasibility of *M. smegmatis* as a model for the study of the pathogenesis of tuberculosis has been questioned in the past, with positive (Reyrat and Kahn, 2001) and negative opinions (Barry, 2001b). The results of our study indicates, though, that the use of *M. smegmatis* in the secretion analyses is justified, as the ESAT-6 gene clusters from this species seems to be sufficiently different from that of *M. tuberculosis* to not interfere with the analyses, while the cell wall structure is sufficiently similar to allow the transport mechanisms to operate efficiently. An alternative method to replace the use of complementation in the secretion studies would be to use the technique of gene knockouts. The presence of multiple copies of these clusters in the genomes of the mycobacteria has made the use of gene knockout studies unfeasible in the past. This is due to the fact that it was hypothesized that a gene from one region may be able to compensate for the loss of a gene from another region. The present study has indicated though, that the ESAT-6 gene clusters show very high individual specificity towards the dedicated secretion of specific protein sequences, so that the use of gene knockouts may still be an option to consider in future studies of the ESAT-6 secretion mechanisms. In the case where this technique is used, it would be important to consider that the disruption of one gene by knockout may have a polar effect on the whole region, which once again may influence secretion results. In addition to this, we can not exclude the possibility that the specificity of the ESAT-6 gene cluster region secretion may be due to the specificity of only a singular or a few of the genes present within these regions, and that knockouts of the other genes may be compensated for by other regions.

Although one of the limitations of this study is that we do not know how this apparatus is assembled, we have constructed a model based on the putative functions of the genes present in the ESAT-6 gene clusters for the secretion of the members of the ESAT-6 protein family. This model contains all of the components necessary to form a dedicated, multi-component, binding protein-dependant active transport system (Figure 5.10).

Such a system could consist of:

- a cytoplasmic substrate binding protein (possibly the conserved N-terminal transmembrane proteins),
- 1 or 2 reciprocally homologous integral inner membrane proteins that translocate the substrate across the membrane (possibly the putative transporter proteins),
- 1 or 2 peripheral membrane ATP-binding proteins that couple energy to the active transport system (possibly the ATP/GTP binding proteins),
- a dedicated protease to cleave the inactivating prepeptide and activate the transported protein or to authenticate the secreted proteins and to clear slowly folding or misfolded proteins from the vicinity of the translocation complex (possibly the subtilisin-like serine proteases) and
- a secreted protein transported without the ordinary *sec*-dependant secretion pathway (possibly the ESAT-6 and CFP-10 family members).

Figure 5.10. Schematic representation of the proposed model of ESAT-6 secretion.


Culture medium
Secreted ESAT-6
Polysaccharides, protein and some lipids
Capsule
C-terminal transmembrane subtilisin-like serine protease for putative cleavage of N-terminus and activation of secreted protein
Integral inner membrane protein, transports secreted protein across the membrane
Mycolic acids and arabinogalactan
Cell wall
Peptidoglycan
Plasma membrane
2x N-terminal transmembrane ATPase, energy providing protein for active transport system
N-terminal transmembrane substrate binding protein
ESAT-6
Cytosol
ESAT-6 gene cluster

It is unknown whether the ESAT-6 gene cluster transport system is also able to import the ESAT-6 molecules back into the organism, although it is possible as this is commonly observed in other secretions systems such as those of the lantibiotics (e.g. nisin, Sahl and Bierbaum, 1998). If this is the case, these transporters may well be used as carriers for antimycobacterial drugs that are designed according to the ESAT-6 protein structure.

In conclusion, we have developed a novel method of looking at secretion mechanisms in the mycobacteria. Our results have shown that secretion of members of the ESAT-6 protein family is dependent on the presence of the ESAT-6 gene cluster regions. Furthermore, we have shown that this secretion is region-specific, to such an extent that even orthologous regions between different mycobacteria are unable to cross secrete ESAT-6. We have also constructed a possible model for the secretion apparatus, based on putative functions of proteins encoded by the ESAT-6 gene cluster regions. This investigation has important implications for the study of dedicated mycobacterial transport and secretion mechanisms, as well as for the understanding of secretion of important T-cell antigens of the mycobacteria. This could lead to the development of efficient strategies to either terminate or enhance secretion of these antigens, thereby influencing the immunogenicity of the pathogens, which may ultimately have an impact on vaccine design and development.

ADDENDA TO CHAPTER FIVE

ADDENDUM 5A

POTENTIAL QUORUM SENSING FUNCTION

"...shall your city call us lord, in that behalf which we have challenged it? Or shall we give the signal to our rage and stalk in blood to our possession?"

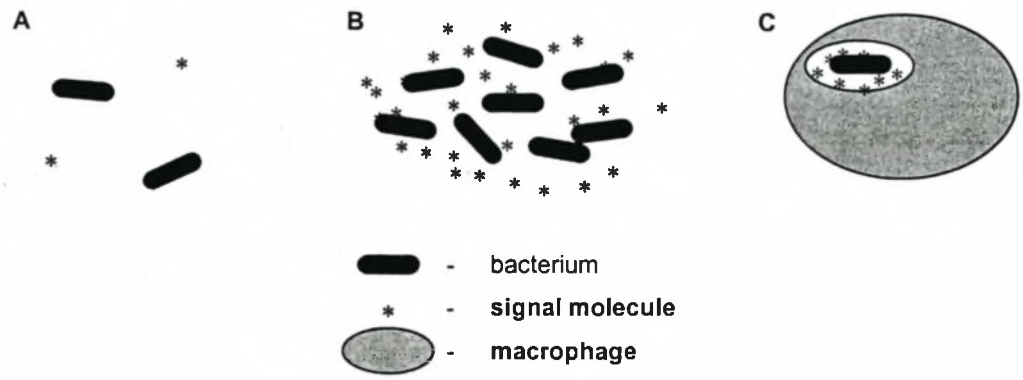
King John (Act II Scene I) - Shakespeare

NOTE: The views presented in the following chapter will be submitted as an hypothesis for peer review and publication as: "**Are the immunologically important mycobacterial ESAT-6 gene clusters involved in cell-cell signaling?**, Gey van Pittius, N.C., Warren, R.M., Siezen, R.J., and Van Helden, P.D."

5A.1. Introduction

Intracellular communication or quorum sensing is an important mechanism for the control of essential processes in the growth of microorganisms. In Gram-positive bacteria quorum sensing is a cell-cell communication mechanism which depends on the secretion of signaling molecules, mostly peptides or modified peptides (Dunny and Leonard, 1997). Quorum sensing has one important function, which is the sensing and monitoring of bacterial density in the immediate environment, and the reaction on the information obtained by regulation of gene expression. When a single bacterium or starter bacterial colony releases these molecules, the concentrations of the signals are very low (see Figure 5A.1.A). However, as soon as these bacteria accumulate, the concentration of the molecules increases, leading to the sensing and subsequent activation or repression of different target genes (Figure 5A.1.B, De Kievit and Iglewski, 2000). These target genes may include virulence factors, gene transfer proteins as well as antibiotic- and bacteriocin(lantibiotic)-production proteins (Dunny and Leonard, 1997). In a situation where a single bacterial cell is phagocytosed by a macrophage (which is predominantly the case with the mycobacterial cell during infection), the bacterium is encapsulated inside a small space within the phagosome (Figure 5A.1.C). In this scenario a single bacterium would be able to release enough signaling molecules on its own, to cause a high concentration of these molecules present in the small space between the cell wall of the organism and the vacuolar membrane (Dunny and Leonard, 1997). As this critical threshold concentration is reached fairly quickly, specific gene expression could be induced to manage the new environment. In this way the bacterium is able to monitor its environment, to sense that it is now intracellular and is then able switch on genes that may include genes involved in intracellular growth, resistance to killing, resistance to acidification of the phagosome etc. In an example of this mechanism, Sperandio *et al.* (1999) showed that quorum sensing controls the expression of operons encoding the type III secretion system of the LEE pathogenicity island in both enterohemorrhagic *Escherichia coli* and enteropathogenic *E. coli*. These authors speculate that the discovery of quorum sensing mechanisms in pathogenic bacteria is an important factor in the search for the inhibition of intracellular communication between bacteria and thus also the inhibition of subsequent expression of virulence determinants *in vivo*.

Figure 5A.1. Environments influencing the regulation of gene expression by quorum sensing. (A) single bacteria secreting signaling molecules could determine the absence of other bacteria in the immediate environment, (B) higher numbers of bacteria cause a high concentration of signaling molecules causing the expression of genes to compensate for the decrease in nutrients and space, (C) intracellular bacteria within vacuole of macrophage could determine intracellular residence by high concentration of signaling molecules and induce the expression of genes to compensate for the new intracellular environment. Adapted from Dunny and Leonard (1997).

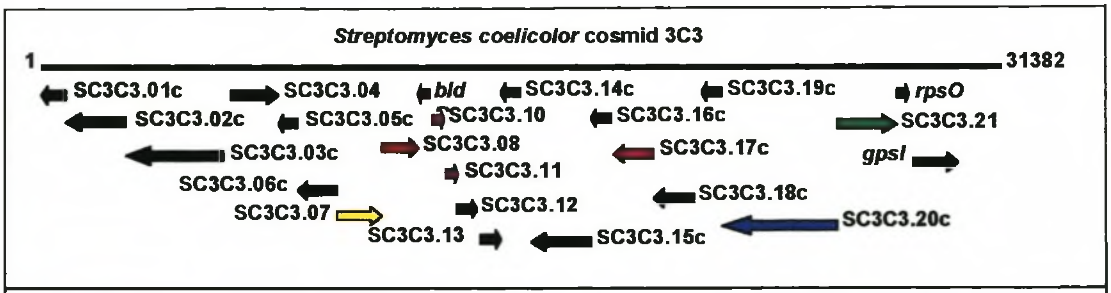


In this review on the potential function of the ESAT-6 gene clusters, we provide evidence from the literature showing that these clusters share many characteristics with quorum sensing operons of Gram-positive bacteria. We hypothesize that the ESAT-6 gene clusters may represent mycobacterial-specific quorum sensing operons that may be involved in the determination of intracellular residence.

5A.2. Comparison of the ESAT-6 gene clusters with cell-signaling and lantibiotic operons

As discussed in previous chapters, we have shown that the ESAT-6 gene clusters (Chapter 3) are organized as operon structures (Chapter 4) and encode proteins that are involved in the secretion of the small T-cell antigens from the *esat-6* gene family (Chapter 5). We have also shown the presence of copies of these gene clusters in the genomes of all other mycobacteria tested (Chapter 3), as well as the presence of an orthologue of the cluster region 4 in the genome of the saprophytic bacterium *S. coelicolor* (Figure 5A.2).

Figure 5A.2. Schematic representation of the genes present within the *S. coelicolor* genome sequence cosmid clone 3C3. Orthologues of the ESAT-6 gene cluster genes are indicated in colour, with a reference table at the bottom of the figure. Positions, relative sizes of genes as well as the direction of transcription are indicated by blocked arrows.



🟨 = SC3C3.07, homologous to Rv3450c - N-terminal transmembrane protein
🟥 = SC3C3.08 and 17c, homologous to Rv3449 - Subtilisin-like serine protease
🟩 = SC3C3.20c, homologous to Rv3448 - Integral membrane protein (putative transporter)
🟦 = SC3C3.21, homologous to Rv3447c - ATPase (3 x ATP/GTP binding sites)
🟪 = SC3C3.10 and 11, homologous to Rv3874 and Rv3875 - ESAT-6/CFP-10 secreted T-cell antigens

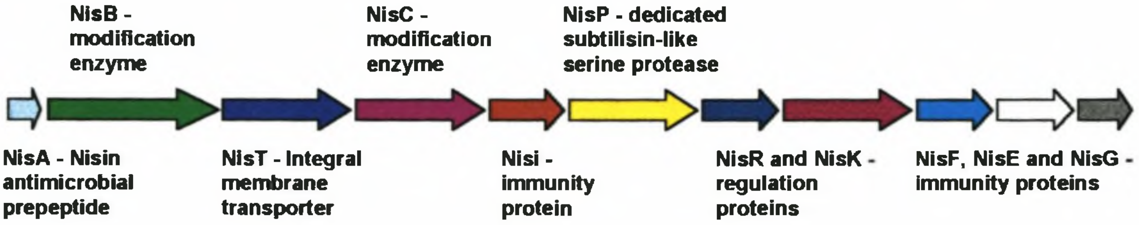
The presence of an orthologue of the ESAT-6 gene cluster in *S. coelicolor* may provide some clues to the potential function of this gene cluster. Most of the genes present within the region in *S. coelicolor*, are involved with some peptide transport and secretion function, based on their sequence homology to known transporter proteins. One of the most interesting genes from this region is the gene *bldB* (SC3C3.09, Figure 5A.2). *bldB* encodes a small protein that is required for morphogenesis, antibiotic production, catabolite control and signal production required for cell-cell communication in *S. coelicolor* (Harasym *et al.*, 1990, Pope *et al.*, 1998). These are quite diverse activities for a single protein. Pope and coworkers (1998) attempted to explain this interesting observation from the fact that certain *bld* mutants are able to erect aerial hyphae only when grown in close proximity to wild-type *S. coellcolor.* The fact that these mutants were able to erect the hyphae indicates that they are indeed able to make the structures associated with sporulation, but that they are unable to initiate the differentiation process on their own, because of an inability to send and/or receive the required extracellular signals. It seems that these mutants are thus unable to sense the conditions in its immediate environment (in other words it lost its quorum sensing abilities). As said previously, bldB is required for morphogenesis, antibiotic production, catabolite control and signal production required for cell-cell communication (Harasym *et al.*, 1990, Pope *et al.*, 1998). As carbon utilization, initiation of differentiation and morphogenesis and antibiotic production are all processes controlled by the mechanism of quorum sensing, we speculate that the whole locus surrounding *bldB* may be involved in the process of intracellular signaling and signal molecule secretion.

Pope *et al.* (1998) also deduced from the structure of the bldB protein that it is a putative transcription factor that might regulate a distinct group of developmental genes. It has also been postulated that the products of the *bld* genes are either directly or indirectly involved in the secretion or uptake of extracellular signaling molecules (Willey *et al.*, 1993). The *bldK* locus encodes for proteins homologous to the subunits of the ATP-binding cassette (ABC) membrane-spanning transporters (Nodwell *et al.*, 1996). This oligopeptide transporter was shown to be responsible for the import of an extracellular signaling molecule involved in the production of aerial mycelia (Nodwell *et al.* 1998). Similarly, the ESAT-6 gene cluster region ortholog in *S. coelicolor* also seems to be involved in active transport and the numerous transport-associated proteins encoded by this region could form the subunits for an ABC-type membrane spanning transporter. If this locus is involved in cell-cell communication as with the other *bld* mutants, it is highly plausible that the ESAT-6 gene clusters in the mycobacteria are also cell-signaling loci. We have shown that the ESAT-6 gene clusters in the mycobacteria are involved in the secretion of the ESAT-6 protein family members (Chapter 5). We thus speculate that as a potential transcription factor, BldB could be involved in the regulation of this whole region and that the orthologues for ESAT-6 and CFP-10 (SC3C3.10 and SC3C3.11) are secreted as extracellular signaling molecules.

In agreement with this hypothesis, Mahairas and coworkers have shown that the *M. bovis* BCG RD1 deletion region (situated inside the ESAT-6 gene cluster region 1) might have an influence on regulation of expression of a number of proteins (Mahairas *et al.*, 1996). The reintroduction of RD1 to *M. bovis* BCG appears to strongly repress the expression of at least 10 proteins and downregulates the expression of many additional cellular proteins (Mahairas *et al.*, 1996). Indications are that several of these regulated proteins may be heat shock or stress proteins. If this is true, it might indicate a function in the same manner as the extracellular signaling loci of *S. coelicolor*, which in effect controls the expression of numerous genes important in differentiation, and could implicate signaling events during changes of the extracellular environment and subsequent expression of stress-related proteins. Mahairas and coworkers speculated that the disruption of this stress response might affect the ability to adapt and survive in the host and to cause disease, leading to the observed decrease in virulence caused by the deletion of RD1 (Maharias *et al.*, 1996) and the genes situated within it (Wards *et al.*, 2000). These ESAT-6 gene clusters seem to be key genetic loci in mycobacterial (and clearly much wider) genomics, and therefore it may make sense to use the *S. coelicolor* homologous region as a starting point to further investigate its function.

In the Gram-positive bacteria, the majority of signaling molecules requires a specialized export mechanism and the structural gene is mostly situated within an operon containing export- and modification-encoding genes, as is observed with the ESAT-6 gene clusters (Chapter 5). In addition, all peptide or peptide-derived signaling molecules are acquired through the posttranslational processing of a larger precursor peptide. One of the commonly used examples of quorum sensing mechanisms in Gram-positive bacteria, is the cell-cell signaling mechanism and regulation of nisin lantibiotic biosynthesis in the organism *Lactococcus lactis* (Dunny and Leonard, 1997). Several species of Gram-positive bacteria secrete small, gene-encoded antimicrobial peptides (bacteriocins) that, like the members of the ESAT-6 protein family, also lack ordinary *sec*-dependant secretion signals (Sahl and Bierbaum, 1998). The lantibiotics forms a unique class of the bacteriocins and contain unusual amino acids and lanthionine rings that are introduced by post-translational modifications (Montville and Chen, 1998). Interestingly, the biosynthesis, processing and transport of these peptides are also accomplished by between 3 and 12 genes organized in operons. This includes genes encoding the small antibiotic prepeptide, modification enzymes, dedicated subtilisin-like serine proteases and ABC transporters (containing ATPase activity for active transport). Figure 5.8 gives a schematic representation of one of these lantibiotic gene clusters, the nisin lantibiotic gene cluster.

Figure 5.8. The nisin lantibiotic gene cluster.


NisB - modification enzyme
NisC - modification enzyme
NisP - dedicated subtilisin-like serine protease
NisA - Nisin antimicrobial prepeptide
NisT - Integral membrane transporter
Nisi - immunity protein
NisR and NisK - regulation proteins
NisF, NisE and NisG - immunity proteins

It is well known that some lantibiotics, for example nisin, indeed also function as signaling molecules that can, once they are activated by cleavage and extracellularly secreted, sense the population density in their environment and induce their own biosynthesis (also in surrounding cells) by binding to specific receptors. Only very low levels of transcription of the genes in the biosynthetic gene cluster of nisin occur during exponential growth, so that there is a very gradual accumulation of the peptides in the surrounding culture filtrate. This accumulation takes place up to a level where the concentration is high enough (an indication of cell density) to signal the activation of the promoters and the initiation of transcription of massive levels of mRNA from the genes in the cluster (Sahl and Bierbaum, 1998).

Although they do not seem to contain any modification enzymes as found in the lantibiotic clusters, it is clear that the mycobacterial ESAT-6 gene clusters contain some of the secretion and processing features of lantibiotic operons. It is interesting to note that Sørensen *et al.* (1995) and several other authors observed different forms of the ESAT-6 protein with unexplainable size and pI differences on 2D PAGE analyses. These differences may be due to posttranslational modifications. As the functions of most of the proteins forming part of the ESAT-6 gene clusters are still unknown, it is possible than one or more of them may function as modification enzymes, although classical lantibiotic-type modification enzyme motifs could not be detected in any of them (N.C. Gey van Pittius, Unpublished results).

The presence of the mycosins in the ESAT-6 gene clusters is another feature shared with the lantibiotics clusters, as these clusters are the only other known examples of subtilisin-like membrane-anchored serine proteases forming part of a gene cluster with secretion-associated proteins. The subtilisin enzyme domains of the mycosins in the ESAT-6 gene clusters are separated from their membrane anchors by a short linker or spacer (see Addendum 2A), a feature that is also only found in the nisin-lantibiotic activating subtilisin-like serine protease NisP (Siezen *et al.*, 1996). These lantibiotic subtilases have the exclusive function to cleave the N-terminal extensions of the secreted lantibiotic prepeptide that keeps it inactive, in order to activate the lantibiotic after secretion. Evidence does exist for the processing of ESAT-6 and TB11 (another ESAT-6 family member, Rv1038c, Rosenkrands *et al.*, 2000a, Rosenkrands *et al.*, 2000b), to lower molecular weight products (Peter Andersen and Karin Weldingh, personal communication), as indicated in Figure 5.9. Skjøt and coworkers (2000) also saw a similar processing after obtaining the N-terminal sequence of the purified culture filtrate protein TB10.4 (Rv0288) that belongs to the ESAT-6 family and is located in the ESAT-6 gene cluster region 3. They observed that the amino acid sequence starts from residue 13 onwards (Figure 5.9), and speculated that it may be because of an alternative start site or a partial cleavage of the protein in the culture filtrate.

Figure 5.9. Evidence for N-terminal cleavage of members of the ESAT-6 protein family.

```
                                          Cleavage site
                                                ↓
Rv3875 (ESAT-6): 95 aa     MTEQQWNFAGIEAAA----SAIQGNVTSIHSLLDEGKQSLTKLAAAWG

Rv1038c (TB11): 98 aa     MASRFMTDPHAMRDM----AGRFEVHAQTVEDEARRMWASAQNISGAG

Rv0288 (TB10.4): 96 aa    MSQIMYNYPAML-------GHAGDMAGYAGTLQSLGAEIAVEQAALQS


                       GSGSEAYQGVQQKWDATATELNNALQNLARTISEAGQAMASTEGNVTGMFA

                       WSGMAEATSLDTMTQMNQAFRNIVNMLHGVRDGLVRDANNYEQQEQASQQILSS

                       AWQGDTGITYQAWQAQWNQAMEDLVRAYHAMSSTHEANTMAMMARDTAEAAKWGG
```

Sørensen *et al.* (1995) described multiple forms of the ESAT-6 protein present when analyzing short term-culture filtrates on 2D-E gels. They observed the protein to be focused at two pI's and found three versions of it of sizes differing between 6 and 4kDa at each pI. This could further indicate that there is some N-terminal processing of the proteins resulting in smaller forms. These observations suggest that the members of the ESAT-6 family are activated during secretion. It also provides a possible role for the mycosins, similar to that of the activation subtilases of the lantibiotics. The fact that uncleaved ESAT-6 proteins are observed in culture filtrates, could be explained by the unnatural conditions under which the bacteria are grown in the laboratory. It is possible that the ESAT-6 family members need to be delayed long enough during the secretion process for the cleavage of the N-terminus by the transporter apparatus-associated subtilase. This feature would only be obtainable *in vivo* and would not be possible under the continual stirring or shaking of *in vitro* culture growth conditions.

Thus, if the ESAT-6 gene clusters are not involved in cell-cell signaling *per se* as discussed above, it might be interesting to investigate the possibility that these clusters evolved from early lantibiotic-like clusters in saprophytic mycobacteria and that they might even still show some antimicrobial activity. In fact, Brandt *et al.* (1996) have shown that only low levels of transcription of ESAT-6 are observed during growth *in vitro*, and have speculated that the fact that it is such a potent target *in vivo* could indicate that there is an upregulation of expression of this gene during growth within the macrophage. This upregulation may thus also be the result of a sensing mechanism that recognizes the new intracellular environment.

Lantibiotics, like nisin, act as antimicrobial agents through a number of mechanisms, but primarily through the formation of pores in the membrane of the susceptible organism (Montville and Chen, 1998, reviewed by Sahl and Bierbaum, 1998 and McAuliffe *et al.*, 2000). Nisin forms these poration complexes in the membrane through a multi-step process of binding, insertion and pore formation. In tuberculosis infections, mycobacterial growth damages the vacuolar membrane, leading to leakage (Andersen, 1997). Teitelbaum *et al.* (1999) recently demonstrated that only viable *M. bovis* BCG organisms are able to create phagosomal membrane permeability, suggesting that live mycobacteria may release molecules that create pores in the vesicular membrane. It was shown that formalin-killed BCG and nonpathogenic *M. smegmatis* were unable to accomplish the same. This is important, as we know that mycobacteria reside within a phagosome within the macrophage where it is difficult to access host nutrients (Teitelbaum *et al.*, 1999). The authors speculate that the mycobacteria may biosynthesize and secrete pore-forming molecules that could facilitate bi-directional, size-restricted transport of nutrients and antigens through the phagosomal membrane. It was shown previously that although BCG was able to facilitate presentation of antigens to T-cells in an MHC class I-restricted manner, this process was several-fold less efficient than that observed in virulent *M. tuberculosis* (Mazzaccaro *et al.*, 1996). It could be hypothesized that BCG may thus contain less or inefficient copies of the pore-forming molecules, resulting in inefficient access of mycobacterial antigens to the MHC molecules. If the small, secreted ESAT-6 family members function in a lantibiotic manner (owing to their similarity to the nisin lantibiotic biosynthesis system), these proteins might be the secreted proteins involved in the formation of pores in the host vacuolar membrane, and could be responsible for the differences in phagosomal permeability phenotype observed between the species of mycobacteria.

The fact that there are multiple copies of the ESAT-6 gene clusters found in *M. tuberculosis* (Gey van Pittius *et al.*, 2001) may indicate an intricate cellular sensing network that would greatly benefit or even be required by such a successful intracellular pathogen. It may also be possible that the functions of some of these clusters have diverged and that they have not all retained the same biological function as putative sensing molecules. The effects of ESAT-6 gene cluster region 1 deletions could not be complemented by any of the other regions (Mahairas *et al.*, 1996, Wards *et al.*, 2000), indicating that each of these regions perform a specific function in the biology of the organism.

In conclusion, we have shown from the available literature that the mycobacterial ESAT-6 gene clusters contain a number of features of quorum sensing and lantibiotic operons. We hypothesize that members of the ESAT-6 family may be secreted as signaling molecules and are involved in the regulation of expression of genes during intracellular residence of the bacterium within the macrophage.

CHAPTER SIX

PE AND PPE EXPANSION

" .... a theory trying to unify a vast and difficult field with innumerable details is certainly nothing static; it is a fleeting moment in an eternal flux."

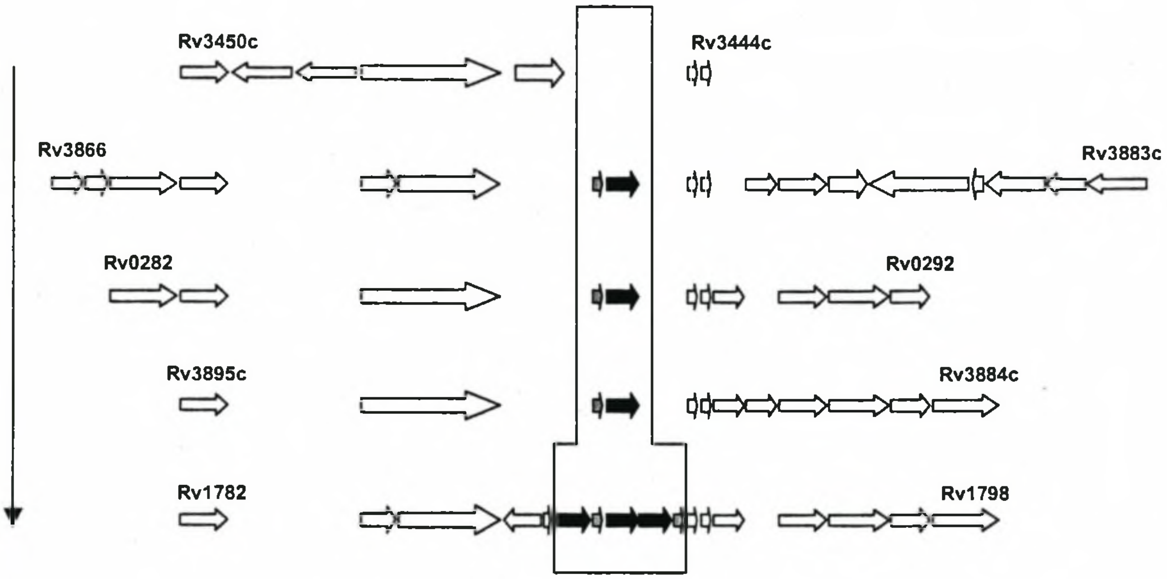
The theory of the gene – R. Goldschmidt (1951)

NOTE: The results presented in the following chapter will be submitted for peer review and publication as: "**The evolutionary history of the expansion of the *Mycobacterium tuberculosis* PE and PPE multigene families and its association with the duplication of the ESAT-6 gene cluster,** Gey van Pittius, N.C., Sampson, S.L., Lee, H., Warren, R.M., and Van Helden, P.D."

6.1. Introduction

The genome of *Mycobacterium tuberculosis* contains five duplicate copies of the immunologically important ESAT-6 gene clusters. Each gene cluster encodes proteins involved in energy provision for active transport, membrane pore formation and protease processing and could thus assemble to form a dedicated biosynthesis, transport and putative processing system for the secretion of the potent T-cell antigens belonging to the ESAT-6 protein family (Chapter 5). Interestingly, in addition to these genes, there are two families of genes present within these clusters which seem to be anomalous, named the PE and PPE gene families (Figure 6.1).

Figure 6.1. Genomic organization of the *Mycobacterium tuberculosis* ESAT-6 gene clusters. The vertical arrow indicates the direction of duplication, with the ESAT-6 gene cluster regions numbered region 4, 1, 3, 2 and lastly 5 in descending order. The positions of the PE (small arrow in grey) and PPE (larger arrow in black) genes are blocked.



The PE and PPE protein families are large multigene families (99 and 68 members respectively) of unknown function (Cole *et al.*, 1998). They are glycine-rich protein families comprising about 10 % of the coding potential of the genome of *M. tuberculosis* (Cole *et al.*, 1998).

The PE family is characterized by the presence of a proline-glutamic-acid (PE) motif at positions 8 and 9 in a very conserved N-terminal domain of around 110 amino acids (Gordon *et al.*, 1999b). Similarly, the PPE family also contains a highly conserved N-terminal domain of around 180 amino acids, with a proline-proline-glutamic acid (PPE) motif at positions 7 - 9 (Cole *et al.*, 1998). The C-terminal domains of both these protein families are of variable size and sequence and contain repeat sequences of different copy numbers in a number of cases (Gordon *et al.*, 1999b).

Both these families are divided into subgroups according to the homology and presence of motifs in their C-terminal domains (Gordon *et al.*, 1999b). The polymorphic GC-rich sequence (PGRS) subgroup of the PE family is the largest subgroup and contains proteins with multiple tandem repeats of a glycine-glycine-alanine or a glycine-glycine-asparagine motif in the C-terminal domain. The other subgroup consists of proteins with C-terminal domains of low homology. The PPE family can be divided into three subgroups (Gordon *et al.*, 1999b) of which the major polymorphic tandem repeat (MPTR) subgroup is the largest. The proteins of this subgroup contain multiple repeats of the motif AsnXGlyXGlyXAsnXGly encoded by a consensus repeat sequence GCCGGTGTTG, seperated by 5 bp spacers (Cole and Barrell, 1998). The PPE-SVP subgroup is characterized by the motif GlyXXSerValProXXTrp at position 350 in the amino acid sequence and the last subgroup consists of proteins with a low percentage of homology at the C-terminus (Gordon *et al.*, 1999b).

Until recently, no evidence has been available for the subcellular localization of the members of the PE and PPE proteins, although an early paper by Doran and coworkers (1992) suggested that the members of the PPE-MPTR family were likely to be cell wall associated. Recently, though, it has been shown that certain PE-PGRS proteins are cell-surface constituents which influence the interactions of the organism with other cells (Brennan *et al.*, 2001). In addition to this, the PPE-MPTR protein Rv1917c was also found to be situated in the cell wall and is at least partly exposed on the cell surface (S. Sampson, submitted for publication).

Although the 167 members of the PE and PPE gene families are of unknown function, it has been suggested that the proteins encoded by these gene families may inhibit antigen processing or may be involved in antigenic variation due to the highly polymorphic nature of their C-terminal domains (Cole *et al.*, 1998, Cole, 1999, Gordon *et al.*, 1999b). In agreement with this, sequence variation has been observed between the orthologues of the PE and PPE protein families in an *in silico* analysis of the genomes of *M. tuberculosis* H37Rv and *M. bovis* (Cole *et al.*, 1998, Gordon *et al.*, 2001). Extensive variation of a subset of PPE genes in clinical isolates of *M. tuberculosis* have also been observed recently (S. Sampson, submitted for publication). Other clues to the putative functions of the members of these families also exist. For example, Rodriguez and colleagues (1999) have found that the PPE gene Rv2123 is upregulated under low iron conditions, leading to the hypothesis that it may encode a siderophore involved in iron uptake. Abou-Zeid *et al.* (1991) described a 55 kDa fibronectin binding protein, which was later found to be a member of the PE-PGRS subfamily and related to Rv1759c (Cole *et al.*, 1998, Espitia *et al.*, 1999). Rv1759c was in turn also found to be able to bind fibronectin (Espitia *et al.*, 1999). Furthermore, it was recently shown that two members of the PGRS subfamily from *M. marinum* are essential for replication in macrophages as well as persistence in granulomas (Ramakrishnan *et al.*, 2000). Additional recent data have also suggested that the members of the PPE gene family may be involved in disease pathogenesis, as a transposon mutant of the gene Rv3018c was attenuated for growth in macrophages (Camacho *et al.*, 1999). The fact that these genes encode for about 4% of the total protein species in the organism (if all genes are expressed), indicates that they most probably fulfill an important function or functions in the organism.

The duplication order of the ESAT-6 gene clusters within the genome of *M. tuberculosis* has been predicted by systematic phylogenetic analyses of the constituent genes (Gey van Pittius *et al.*, 2001). This duplication order was shown to extend from the ancestral region named region 4 (Rv3444c-Rv3450c) to region 1 (Rv3866-Rv3883c), 3 (Rv0282-Rv0292), 2 (Rv3884c-Rv3895c), and lastly region 5 (Rv1782-Rv1798)(Figure 6.1 and Figure 3.8B). The absence of a pair of PE and PPE proteins within the ancestral region 4, indicates that these genes may have been integrated into the first duplicate of this region (region 1), and have subsequently been successfully co-duplicated together with the rest of the genes within the regions. Supporting evidence for the published duplication order of the ESAT-6 gene clusters also comes from the fact that the last duplicate (region 5) includes multiple separate duplications of the PE and PPE genes within the region (Figure 6.1).

This study investigates the duplication characteristics of the PE and PPE gene families situated in and outside of the immunologically important ESAT-6 gene clusters, using a combination of phylogenetic analyses, DNA hybridization as well as comparative genomics (between the genomes of the pathogenic slow-growing mycobacterium *M. tuberculosis* and the fast-growing, non-pathogenic *M. smegmatis*). This investigation attempts to answer the question of why these PE and PPE proteins are situated, as well as tolerated within the ESAT-6 gene clusters, as well as whether the duplication of the genes into the ESAT-6 gene clusters lend some kind of advantage to the family as a whole. We envisage that this data will provide a better understanding of the factors involved in the massive expansion of the PE and PPE families and the contribution of the relationship to the ESAT-6 gene clusters in the evolutionary history.

6.2. Materials and Methods

6.2.1. Genome sequence data and analyses

Annotations, descriptions and protein sequences of individual genes belonging to the PE and PPE families were obtained from the publicly available finished and unfinished genome sequence databases for *M. tuberculosis* H37Rv (http://genolist.pasteur.fr/TubercuList/) as well as *M. smegmatis* mc^{2}155 (http://www.tigr.org/cgi-bin/BlastSearch/blast.cgi?organism=m_smegmatis).

The phylogenetic reconstruction of the evolutionary relationships between the members of the PE and PPE protein families was done by analyses of four separate datasets. The first two datasets downloaded from the *M. tuberculosis* H37Rv database included the protein sequences of all the members of the PE and PPE protein families that are present within the ESAT-6 gene clusters, respectively.

For the third dataset, the protein sequences of the sixty-eight members of the PPE family were downloaded from the *M. tuberculosis* H37Rv database. Ten of the predicted PPE proteins did not contain the characteristic N-terminal PPE motif, but in five of these (Rv0305c, Rv3425, Rv3426, Rv3429, Rv3892c) it was found that one of the two proline residues in the conserved motif have been substituted and they could be reliably aligned to the rest of the family members due to a high percentage of sequence homology. The other five proteins (Rv0304c, Rv0354c, Rv2353c, Rv3021c and Rv3738c) were excluded from the analysis as it was found that their upstream regions were disrupted by either IS*6110* insertion or apparent frameshift mutations.

For the fourth dataset, the protein sequences of the ninety-nine members of the PE family were downloaded from the *M. tuberculosis* H37Rv database. One of the members of the predicted PE family (Rv3020c) was found to have been annotated incorrectly by Cole *et al.* (1998) (Gey van Pittius *et al.*, 2001, see Addendum 3A). Two members of the predicted PE proteins (Rv3539 and Rv2126c) could not be reliably aligned due to a loss of the N-terminal conserved regions, and were excluded from further analyses. Six members (Rv0833, Rv1089, Rv2098c, Rv3344c, Rv3512, and Rv3653), which also did not have conserved N-termini, were shown to actually be situated adjacent to a gene encoding for the N-terminus (Rv0832, Rv1088, Rv2099c, Rv3345c, Rv3511, and Rv3652). Closer inspection of this organization suggested that these genes were actually one gene that was split by stop codon formation during frameshifting. Thus, each pair of genes from this group were combined and included in the analyses.

6.2.2. Multiple sequence alignments

Due to the highly polymorphic nature of the C-terminal part of the PE as well as the PPE proteins, only the conserved N-terminal domains of 100 aa and 180 aa, respectively were used to construct the multiple sequence alignments. Multiple sequence alignments of the protein sequences of the ninety-six PE and sixty-three PPE proteins were done using ClustalW 1.5 on the WWW server at the European Bioinformatics Institute website (http://www2.ebi.ac.uk/clustalw/; Thompson *et al.*, 1994). The alignments were manually checked for errors and refined where appropriate.

6.2.3. Phylogenetic trees

Neighbour-joining phylogenetic analyses were done using the program PAUP 4.0b8 (Swofford, 1998), and 1000 subsets were generated for Bootstrapping resampling of the data. Confidence intervals for the internal topology of the trees were obtained from the resampling analyses and only nodes occurring in over 50% of the trees were assumed to be significant (Felsenstein, 1985). All branches with a zero branch length were collapsed. Based on the evolutionary order defined for the ESAT-6 gene clusters (Gey van Pittius *et al.*, 2001), we have used the ancestral PE and PPE genes present within ESAT-6 gene cluster region 1 (Rv3872 and Rv3873, respectively) as the outgroups. The consensus trees of the above were calculated using the majority rule formula and were drawn using the program Treeview 1.5 (Page, 1996).

6.2.4. Comparative genomics analyses

BLAST similarity searches (Altschul *et al.*, 1990) with the tblastn algorithm, were done using the WU-BLAST version 2.0 (http://blast.wustl.edu/) server in the database search service of the TIGR website, to identify orthologues of the *M. tuberculosis* PE and PPE genes from the *M. smegmatis* mc^{2}155 whole genome sequence. To confirm the identity of the resulting sequences, potential surrounding open reading frames were identified in the immediate vicinity of the genes in the *M.*

smegmatis genome sequence. These open reading frames were subsequently examined to determine if they correspond to the genes surrounding the *M. tuberculosis* PE and PPE genes, thereby confirming the identity of the orthologue.

6.2.5. Primers and Probes

The primers from the gene sequences of the PPE gene family used in this study are listed in Table 6.1. PPE 5' terminal probes were generated for genomic DNA hybridization from the selected primers by separate PCR amplification of regions from the genes Rv1787 (primers ppe-15 and ppe-16), Rv2123 (primers 2123F and 2123R), Rv3018c (primers 3018cF and 3018cR) and Rv3429 (primers 3425F and 3429R), respectively. The respective probes were named 15/16, 2123, 3018, and 3429.

Table 6.1. List of oligonucleotide primers used for dot blot hybridization probe generation

Name of primer	Primer sequence (from 5' to 3')	Length of primer	Tm * [°C]	G+C (%)	Application
ppe-15	tgg act tcg ggg cgt tac	18 bp	58	61.1	*Amplification of 499 bp 5' terminal*
ppe-16	aac gga atc aac cgc gac	18 bp	56	55.5	*region from Rv1787 and Rv1790*
2123F	atg tgg ttc gca gtt ccg c	19 bp	60	57.9	*Amplification of 227 bp 5' terminal*
2123R	gtt agc caa tac cgg aac gg	20 bp	62	55.0	*region from Rv2123*
3018cF	att cgg cgc tgc taa gtg c	19 bp	60	57.9	*Amplification of 160 bp 5' terminal*
3018cR	aac tca gca ctg gga ccc tg	20 bp	64	60.0	*region from Rv3018c and Rv3021c*
3425F	cat cca atg ata cca gcg gag	21 bp	64	52.4	*Amplification of 148 bp 5' terminal*
3429R	gct cgc cga gcc tgt cgg	18 bp	64	77.8	*region from Rv3429*

*Tm were calculated using the following formula: [4x (G+C)] + [2x (A+T)].

6.2.6. Dot blot analyses

Dot blot analyses were done by blotting a small amount of genomic DNA isolated from different mycobacterial species onto a membrane and probing this using ECL-labelled probes as listed in section 6.2.5. The genomic DNA probed in this analysis was isolated from the mycobacterial species listed in Table 6.2.

Table 6.2. Mycobacterial species used to obtain genomic DNA for dot blot analyses

	Mycobacterial species	Slow/fast growing	ATCC number
1	*M. aichiense*	Fast	ATCC 27280
2	*M. asiaticum*	Slow	ATCC 25276
3	*M. aurum*	Fast	ATCC 23366
4	*M. avium*	Slow	ATCC 25291
5	*M. celatum*	Slow	ATCC 51131
6	*M. celatum*	Slow	ATCC 51130
7	*M. chitae*	Fast	ATCC 19627
8	*M. fallax*	Fast	ATCC 35219
9	*M. fortuitum*	Fast	ATCC 6841
10	*M. fortuitum*	Fast	ATCC 49403
11	*M. fortuitum*	Fast	ATCC 49404
12	*M. genavense*	Slow	ATCC 51233
13	*M. gilvum*	Fast	ATCC 43909
14	*M. gordonae*	Slow	ATCC 14470
15	*M. haemophilum*	Slow	ATCC 29548
16	*M. intracellulare*	Slow	ATCC 13950
17	*M. kansasii*	Slow	ATCC 12478
18	*M. malmoense*	Slow	ATCC 29571
19	*M. marinum*	Slow	ATCC 927
20	*M. mucogenicum*	Fast	ATCC 49650
21	*M. neoaurum*	Fast	ATCC 25795
22	*M. nonchromogenicum*	Slow	ATCC 19530
23	*M. parafortuitum*	Fast	ATCC 19686
24	*M. peregrinum*	Fast	ATCC 14467
25	*M. phlei*	Fast	ATCC 11758
26	*M. scrofulaceum*	Slow	ATCC 19981
27	*M. senegalense*	Fast	ATCC 35796

28	*M. simiae*	Slow	ATCC 25275
29	*M. smegmatis*	Fast	ATCC 19420
30	*M. terrae*	Slow	ATCC 15755
31	*M. thermoresistibile*	Fast	ATCC 19527
32	*M. triviale*	Slow	ATCC 23292
33	*M. tuberculosis* H37Rv	Slow	ATCC 25618
34	*M. tuberculosis* K (Korean clinical strain)	Slow	N/A
35	*M. ulcerans*	Slow	ATCC 19423
36	*M. vaccae*	Fast	ATCC 15483
37	*M. xenopi*	Slow	ATCC 19250

N/A - not applicable

6.3. Results

6.3.1. Phylogeny of the PE and PPE protein families

The phylogenetic trees constructed from the results of the analyses of all the members of the PE and PPE families present in the ESAT-6 gene cluster regions (Figure 6.2) showed topologies similar to the phylogenetic trees obtained for all the other gene families situated in the clusters (Chapter 3, Figure 3.8 and 3.9). This confirms that the PE and PPE genes were duplicated together with the ESAT-6 gene clusters after initial insertion, rather than being inserted during multiple events. These results also confirms the previously determined duplication order of the gene clusters (Gey van Pittius *et al.*, 2001).

The phylogenetic tree constructed from the ninety-six PE protein family N-terminal sequences (and rooted to the PE outgroup from ESAT-6 gene cluster region 1, Rv3872) showed an evolutionary topology similar to the phylogenetic tree constructed from the sixty-three PPE sequences (rooted to the PPE outgroup, Rv3873, Figure 6.3 and Figure 6.4). Each tree is characterized by five distinct corresponding sublineages (indicated by Roman numericals in Figure 6.3 and 6.4). Three of these sublineages correspond to the PE-PGRS, PPE-SVP and PPE-MPTR subgroups, respectively. These results confirm the subgroupings of the PE and PPE families proposed previously (Cole *et al.*, 1998, Gordon *et al.*, 1999b). As the tree topologies correspond to each other, it also suggests a possible evolutionary history for the gene families. Interestingly, this evolutionary history corresponds to the evolutionary history determined for the ESAT-6 gene clusters, with duplication events expanding from region 1 to 3, 2 and lastly region 5. The topology of the phylogenetic trees suggests that the PE-PGRS as well as the PPE-MPTR subgroups are the result of the most recent evolutionary events and have evolved from the subgroups that include the ESAT-6 gene cluster region 5 PE and PPE genes, respectively (Figure 6.3 and 6.4, sublineage 4). This is supported by the observation that some members of the PPE sublineage 4 (PPE-SVP subgroup) contains isolated MPTR repeats, suggesting the existence of a common progenitor gene from which the PPE-MPTR subgroup expanded (S. Sampson, unpublished results).

Figure 6.2. Phylogeny of the PE and PPE protein families present within the ESAT-6 gene clusters. The phylogenetic tree of the PE family members are indicated on the left, with the PPE phylogenetic tree situated on the right.

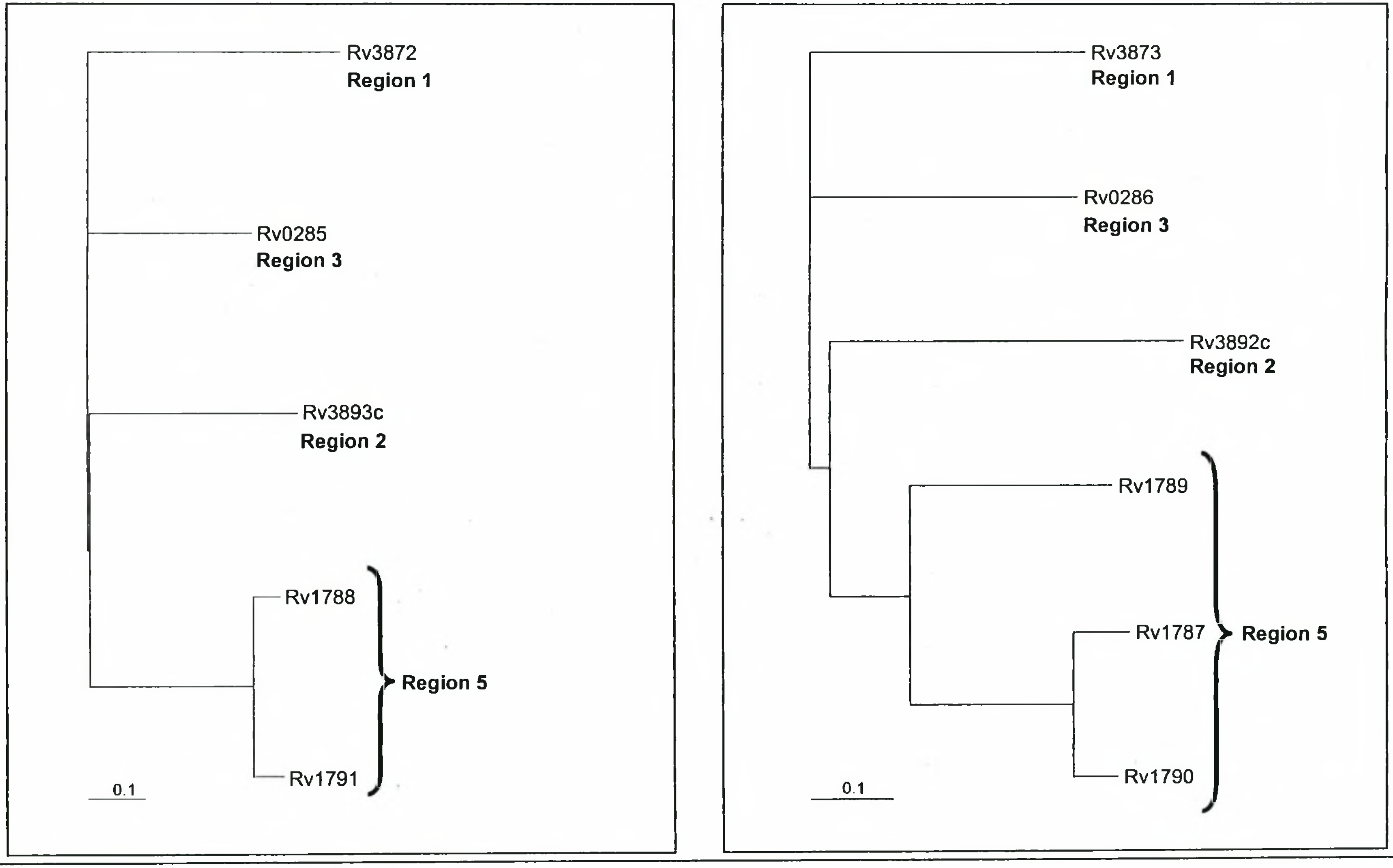


Figure 6.3. Phylogenetic reconstruction of the evolutionary relationships between the members of the PPE protein family. The phylogenetic tree was constructed from the phylogenetic analyses done on the 180 aa N-terminal domains of the PPE proteins. The tree was rooted to the outgroup, which was chosen as Rv3873. This gene has been shown previously to be the first PPE insertion into the ESAT-6 gene clusters (region 1). The gene highlighted in purple is present in ESAT-6 gene cluster region 1, genes highlighted in green are present in or have been previously shown to be duplicated from ESAT-6 gene cluster region 3 (Gey van Pittius *et al.*, 2001), gene highlighted in blue is present in ESAT-6 gene cluster region 2, genes highlighted in red are present in or have been previously shown to be duplicated from ESAT-6 gene cluster region 5 (Gey van Pittius *et al.*, 2001) and genes highlighted in yellow are members of the MPTR subgroup of the PPE family. Arrows indicate genes identified to be present within the *M. smegmatis* genome sequence. Five sublineages (including the PPE-SVP and PPE-MPTR subgroups) are indicated by Roman numericals.

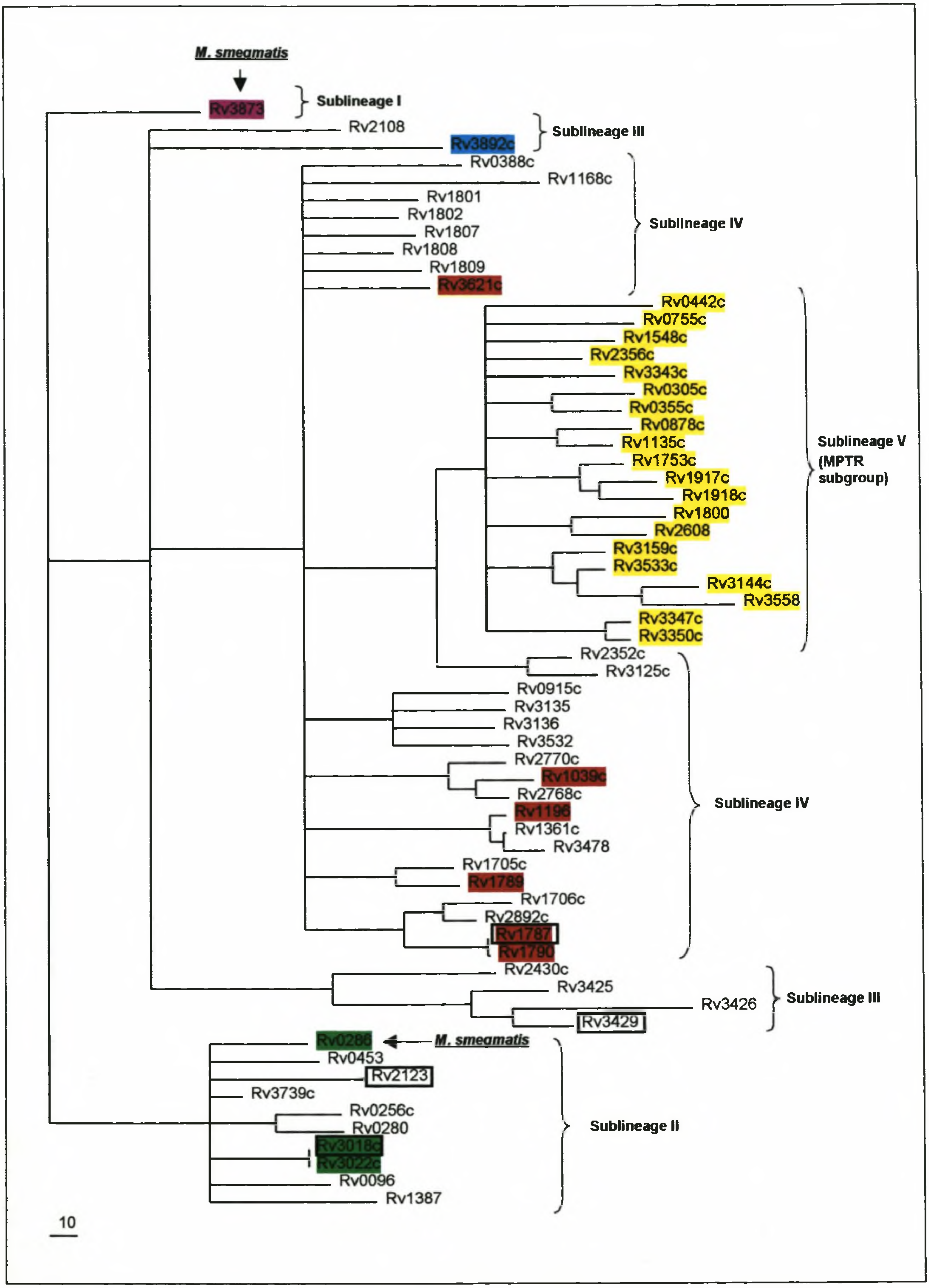
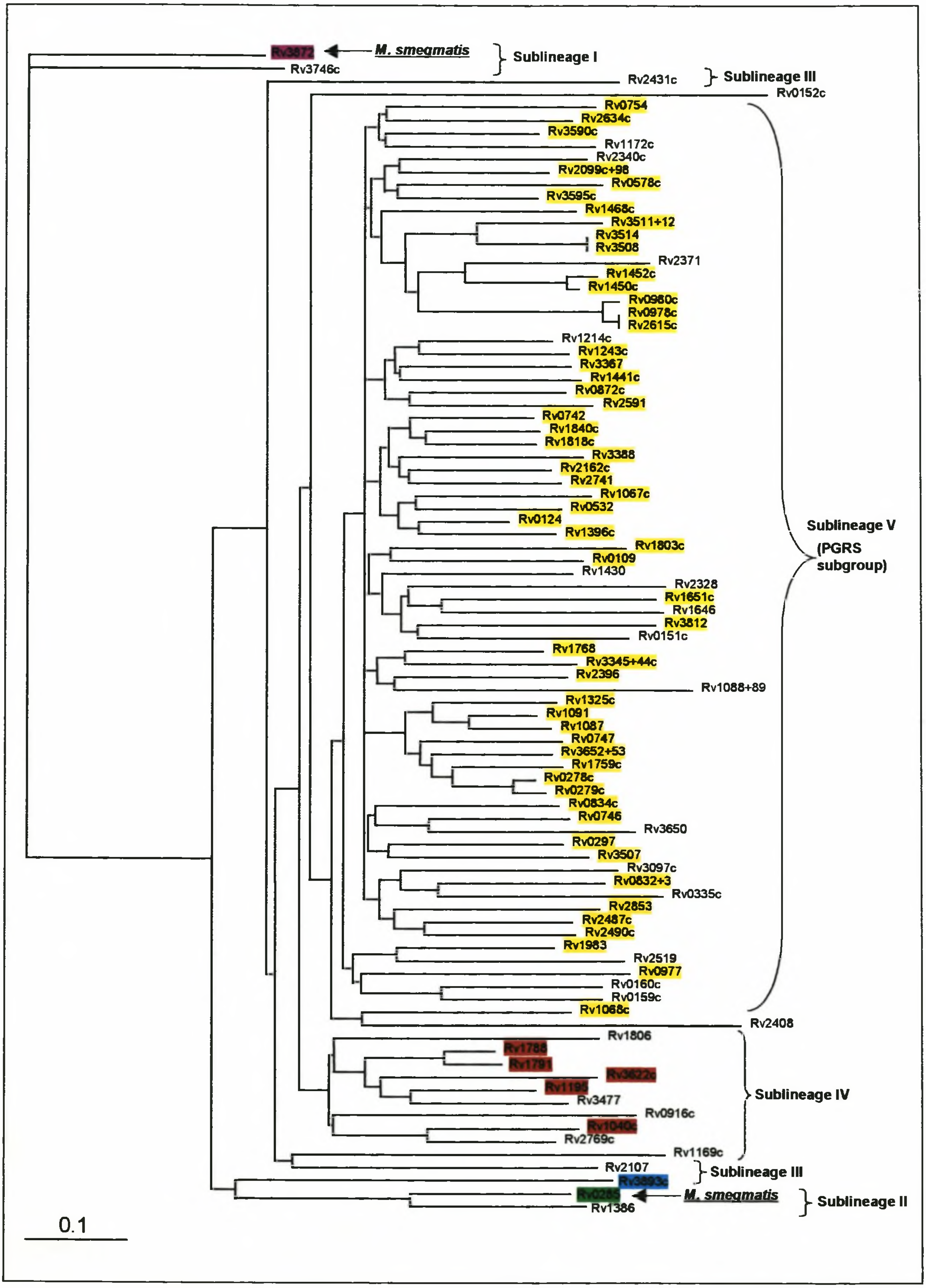

M. smegmatis
Rv3873
Sublineage I
Rv2108
Rv3892c
Sublineage III
Rv0388c
Rv1168c
Rv1801
Rv1802
Rv1807
Rv1808
Rv1809
Rv3621c
Sublineage IV
Rv0442c
Rv0755c
Rv1548c
Rv2356c
Rv3343c
Rv0305c
Rv0355c
Rv0878c
Rv1135c
Rv1753c
Rv1917c
Rv1918c
Rv1800
Rv2608
Rv3159c
Rv3533c
Rv3144c
Rv3558
Rv3347c
Rv3350c
Sublineage V
(MPTR
subgroup)
Rv2352c
Rv3125c
Rv0915c
Rv3135
Rv3136
Rv3532
Rv2770c
Rv1039c
Rv2768c
Rv1196
Rv1361c
Rv3478
Rv1705c
Rv1789
Rv1706c
Rv2892c
Rv1787
Rv1790
Sublineage IV
Rv2430c
Rv3425
Rv3426
Rv3429
Sublineage III
Rv0286
M. smegmatis
Rv0453
Rv2123
Rv3739c
Rv0256c
Rv0280
Rv3018c
Rv3022c
Rv0096
Rv1387
Sublineage II
10

Figure 6.4. Phylogenetic reconstruction of the evolutionary relationships between the members of the PE protein family. The tree was rooted to the outgroup, which was chosen as Rv3872. This gene has been previously shown to be the first PE insertion into the ESAT-6 gene clusters (region 1). The gene highlighted in purple is present in ESAT-6 gene cluster region 1, the gene highlighted in green is present in ESAT-6 gene cluster region 3, the gene highlighted in blue is present in ESAT-6 gene cluster region 2, genes highlighted in red are present in or have been previously shown to be duplicated from ESAT-6 gene cluster region 5 (Gey van Pittius *et al.*, 2001) and genes highlighted in yellow are members of the PGRS subgroup of the PE family. Arrows indicate genes identified to be present within the *M. smegmatis* genome sequence. Five sublineages (including the PE-PGRS subgroup) are indicated by Roman numericals.


Rv3872
M. smegmatis
Sublineage I
Rv3746c
Rv2431c
Sublineage III
Rv0152c
Rv0754
Rv2634c
Rv3590c
Rv1172c
Rv2340c
Rv2099c+98
Rv0578c
Rv3595c
Rv1468c
Rv3511+12
Rv3514
Rv3508
Rv2371
Rv1452c
Rv1450c
Rv0980c
Rv0978c
Rv2615c
Rv1214c
Rv1243c
Rv3367
Rv1441c
Rv0872c
Rv2591
Rv0742
Rv1840c
Rv1818c
Rv3388
Rv2162c
Rv2741
Rv1067c
Rv0532
Rv0124
Rv1396c
Rv1803c
Rv0109
Rv1430
Rv2328
Rv1651c
Rv1646
Rv3812
Rv0151c
Rv1768
Rv3345+44c
Rv2396
Rv1088+89
Rv1325c
Rv1091
Rv1087
Rv0747
Rv3652+53
Rv1759c
Rv0278c
Rv0279c
Rv0834c
Rv0746
Rv3650
Rv0297
Rv3507
Rv3097c
Rv0832+3
Rv0335c
Rv2853
Rv2487c
Rv2490c
Rv1983
Rv2519
Rv0977
Rv0160c
Rv0159c
Rv1068c
Sublineage V
(PGRS
subgroup)
Rv2408
Rv1806
Rv1788
Rv1791
Rv3622c
Rv1195
Rv3477
Rv0916c
Rv1040c
Rv2769c
Sublineage IV
Rv1169c
Rv2107
Sublineage III
Rv3893c
Rv0285
M. smegmatis
Sublineage II
Rv1386
0.1

The observation that the highly polymorphic PGRS and MPTR subfamilies seem to have expanded from the sublineage including the ESAT-6 gene cluster region 5 genes, is potentially significant, as the genes from this cluster seem to be highly prone to duplication. This is evident from the fact that it is the only one of the five ESAT-6 gene clusters which contains multiple copies of the PE and PPE genes situated inside the cluster (Figure 6.1). Furthermore, this ESAT-6 gene cluster is also the parent of a number of secondary duplications containing only the genes for PE, PPE, ESAT-6 and CFP-10 (Chapter 3, Figure 3.2). It is thus clear that this region plays an important role in the propagation of both the ESAT-6/CFP-10 and the PE/PPE genes. We also speculate that this duplication propensity of the region 5 genes may have resulted in the expansion of the PGRS and MPTR subfamilies.

Closer inspection of the positions of the PE and PPE genes in the *M. tuberculosis* genome sequence revealed that in a number of cases a copy of each of these families was found situated adjacent to each other (Table 6.3). By examining the evolutionary positions of these genes on the PE and PPE phylogenetic trees, it was found that these genes are also always situated in the same sublineage on the trees, indicating that they were co-duplicated. Furthermore, the order of their positions are also always conserved with the PE protein always found situated upstream of the PPE gene. These paired genes are found in all the sublineages except in the highly polymorphic PGRS and MPTR subfamilies (sublineage 5). In this sublineage, the genes were in most cases found situated on their own within a specific genomic location. Thus, although it is clear that a change have taken place in the duplication characteristics with the expansion of the PGRS and MPTR subfamilies, the cause and significance of this organization is unclear.

Table 6.3. Paired genes present in both the PE and PPE multigene families.

Sub-lineage	Paired genes (PE)	(PPE)	Associated ESAT-6 gene cluster region
I	Rv3872	Rv3873	Situated in ESAT-6 gene cluster region 1
I or II	Rv3746c	Rv3739c	Associated with ESAT-6 gene cluster region 1 or 3
II	Rv0285	Rv0286	Situated in ESAT-6 gene cluster region 3
II	Rv1386	Rv1387	Associated with ESAT-6 gene cluster region 3
III	Rv3893c	Rv3892c	Situated in ESAT-6 gene cluster region 2
III	Rv2107	Rv2108	Associated with ESAT-6 gene cluster region 2
III	Rv2431c	Rv2430c	Associated with ESAT-6 gene cluster region 2
III or IV	Rv1169c	Rv1168c	Associated with ESAT-6 gene cluster region 2 or 5
IV	Rv1788 / 91	Rv1787 / 89 / 90	Situated in ESAT-6 gene cluster region 5
IV	Rv3622c	Rv3621c	Duplicated from ESAT-6 gene cluster region 5
IV	Rv1195	Rv1196	Duplicated from ESAT-6 gene cluster region 5
IV	Rv1040c	Rv1039c	Duplicated from ESAT-6 gene cluster region 5
IV	Rv1806	Rv1801 / 2 / 7 / 8 / 9	Associated with ESAT-6 gene cluster region 5
IV	Rv3477	Rv3478	Associated with ESAT-6 gene cluster region 5
IV	Rv2769c	Rv2768c / 70c	Associated with ESAT-6 gene cluster region 5
IV	Rv0916c	Rv0915c	Associated with ESAT-6 gene cluster region 5

6.3.3. Comparative genomics

Previously performed comparative genomic analyses indicated that the genome of the non-pathogenic, fast-growing mycobacterium *M. smegmatis* only contains three of the five ESAT-6 gene cluster regions (region 4, 1 and 3), with region 2 and 5 being absent (Gey van Pittius *et al.*, 2001). Although regions 2 and 5 may have been deleted from the genome of this organism, it is much more likely that they were not duplicated because these regions were determined to be the last two duplicates of the ESAT-6 gene cluster evolution (Gey van Pittius *et al.*, 2001). This hypothesis is supported by the fact that the genome of *M. smegmatis* is approximately 1.7 times larger than that of *M. tuberculosis* (Reyrat and Kahn, 2001), and thus does not display the same reductive properties observed in the genome of *M. leprae* (which was confirmed to have lost ESAT-6 gene cluster region 2 and 4 by deletion, Cole *et al.*, 2001).

The phylogenetic analyses done on both the PE as well as the PPE protein families supported a single evolutionary distribution similar to the duplications of the ESAT-6 gene clusters, with region 1 being duplicated to region 3, 2 and lastly 5 (section 6.3.2). It also seems clear from the phylogenetic analyses that the PGRS and MPTR subgroups are the last duplications of the two PE and PPE families respectively, and that they have each originated for the ESAT-6 gene cluster region 5 duplicates. If the hypothesis that regions 2 and 5 were not duplicated in the genome of *M. smegmatis* is true, it would also be logical then to presume that any genes that were duplicated from these regions in *M. tuberculosis* (in other words the PGRS and MPTR gene families as well as any of the other genes associated with region 2 and 5) would not be present in the genome of *M. smegmatis*. Closer inspection of the genome sequence of *M. smegmatis* revealed only two copies of the PE and PPE protein families respectively (indicated in Figure 6.4 and 6.5). These genes are the Rv3872/3 orthologues from ESAT-6 gene cluster region 1 (70% and 55% similarity to the *M. tuberculosis* H37Rv proteins respectively), and the Rv0285/6 orthologues from ESAT-6 gene cluster region 3 (87% and 64% similarity to the *M. tuberculosis* H37Rv proteins respectively). None of the other members of the PE or PPE protein families could be detected within the *M. smegmatis* genome, including any of the PE-PGRS or PPE-MPTR genes (supporting the hypothesis that the genes from these subgroups were duplicated from the ESAT-6 gene cluster region 5).

6.3.4. Dot blot analyses

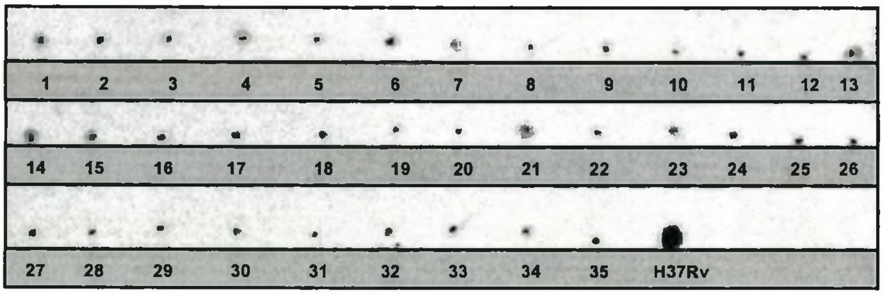
To confirm the results obtained *in silico* with the genome sequence of *M. smegmatis*, as well as to determine the distribution of the PPE protein family in the genomes of various members of the genus *Mycobacterium*, dot blot analyses were done using selected PPE gene probes. The results showed that the genes Rv2123 and Rv3429 (present in sublineages II and III) are not present within the genomes of the mycobacteria other than tuberculosis (MOTTS), indicating that these genes are recent *M. tuberculosis*-specific duplications (Figure 6.6A and B). The sublineage II probe 3018 (for Rv3018c, a duplication of the ESAT-6 gene cluster region 3 PPE) hybridized to the genomic DNA of all the mycobacterial species tested (Figure 6.6C). As it was determined in section 6.3.3 that the *M.*

smegmatis genome does not contain this gene, the sequence of this probe was BLAST-searched against the *M. smegmatis* genome sequence, and resulted in the identification of only the orthologue of Rv0286. The gene sequence of Rv3018c was aligned with the sequence of Rv0286, resulting in a two sequence alignment showing a significant percentage of homology between these two sequences (Figure 6.7). This indicates that the 3018 probe most probably hybridized to the orthologues of the gene Rv0286 in the other mycobacterial species.

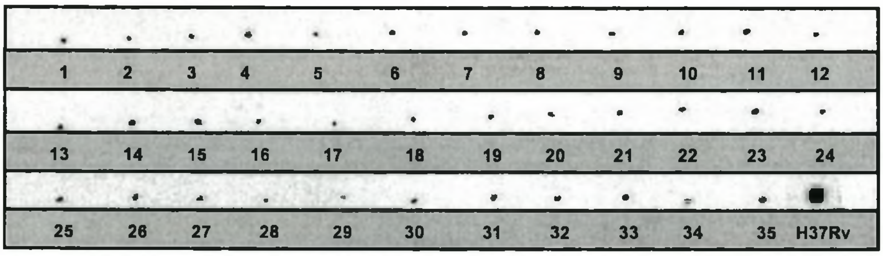
The dot blot result obtained with probe 15/16 (Rv1787, Figure 6.6D) confirmed the results from the comparative genomics on the *M. smegmatis* genome sequence by not hybridizing to the genomic DNA of *M. smegmatis* (Figure 6.6D, number 19). Furthermore, it was found that the gene Rv1787 is not present in the genomes of any of the fast-growing mycobacterial species (see Figure 6.8), but it is present within the genomes of all the slow-growing mycobacterial species tested. The only exception for this is *M. nonchromogenicum*, which might have undergone a deletion of this region. This specific member of the PPE family is thus able to distinguish between slow-growing and fast-growing mycobacteria. As this gene is situated in the ESAT-6 gene cluster region 5, it may be possible that the whole region 5 is absent only in the fast-growing mycobacteria as was observed in the case of *M. smegmatis*.

Figure 6.5. Dot blot analyses of genomic DNA of different members of the genus *Mycobacterium* probed with different members of the PPE gene family. Genomic DNA of different species were blotted onto the membrane in the following order: (1) *M. asiatIcum*, (2) *M. avium*, (3) *M. fortuitum* ATCC 6841, (4) *M. fortuitum* ATCC 49403, (5) *M. fortuitum* ATCC 49404, (6) *M. gordonae*, (7) *M. intracellulare*, (8) *M. kansasii*, (9) *M. malmoense*, (10) *M. nonchromogenicum*, (11) *M. phlei*, (12) *M. scrofulaceum*, (13) *M. terrae*, (14) *M. triviale*, (15) *M. celatum* ATCC 51131, (16) *M. celatum* ATCC 51130, (17) *M. marinum*, (18) *M. peregrinum*, (19) *M. smegmatis*, (20) *M. genavense*, (21) *M. xenopi*, (22) *M. haemophilum*, (23) *M. simiae*, (24) *M. ulcerans*, (25) *M. vaccae*, (26) *M. aichiense*, (27) *M. aurum*, (28) *M. gilvum*, (29) *M. neoaurum*, (30) *M. senegalense*, (31) *M. parafortuitum*, (32) *M. chitae*, (33) *M. fallax*, (34) *M. thermoresistibile*, (35) *M. mucogenicum*, (H37Rv) *M. tuberculosis* H37Rv, (K) *M. tuberculosis* K (Korean clinical strain); A, Dot blot probed with probe 2123 (Rv2123); B, Dot blot probed with probe 3429 (Rv3429); C, Dot blot probed with probe 3018 (Rv3018c); D, Dot blot probed with probe 15/16 (Rv1787).

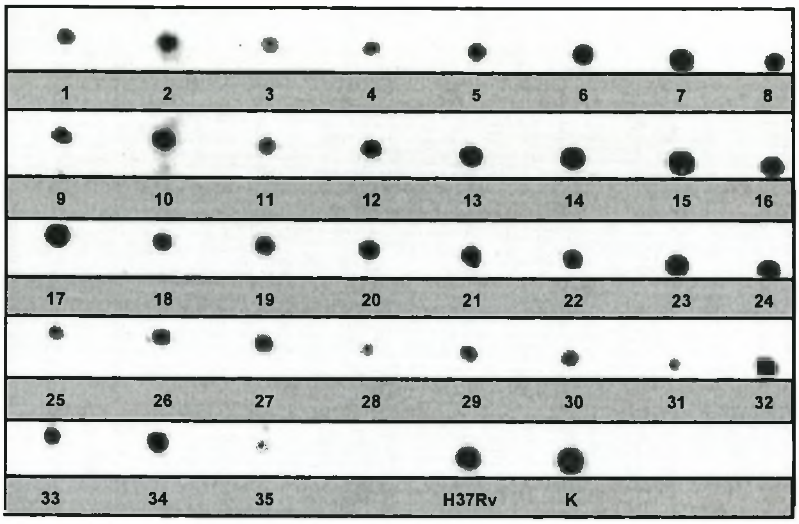
(A) Rv2123



(B) Rv3429



(C) Rv3018c



(D) Rv1787

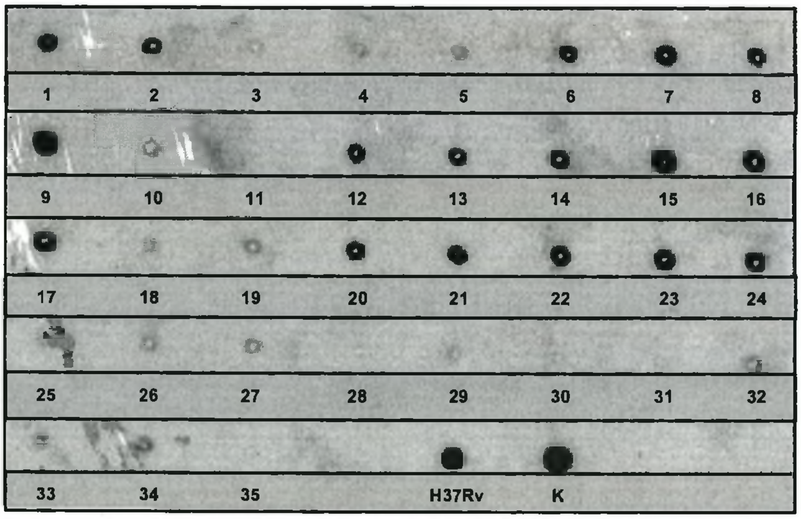


Figure 6.6. Partial two gene sequence alignments of the *M. tuberculosis* H37Rv genes Rv0286 and Rv3018c, as well as the *M. tuberculosis* H37Rv gene Rv3018c and the *M. smegmatis* Rv0286 orthologue. Alignment A shows the high percentage of homology shared between the N-terminal parts of the *M. tuberculosis* H37Rv genes Rv0286 and Rv3018c. Alignment B shows the high percentage of homology shared between the N-terminal parts of the *M. tuberculosis* H37Rv gene Rv3018c and the *M. smegmatis* gene orthologue for Rv0286. Area covered by probe 3018 sequence is highlighted in yellow.

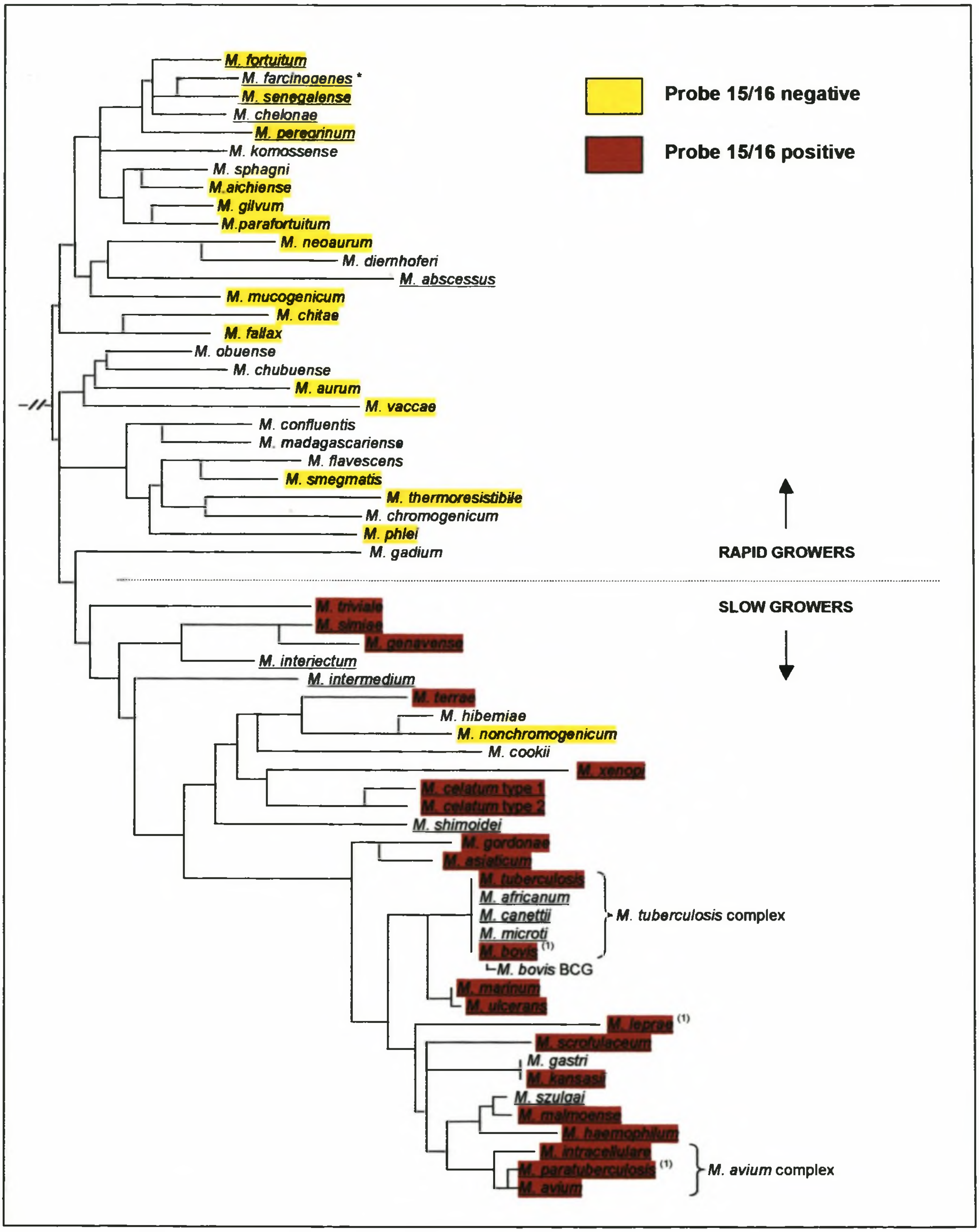
A

```
Rv3018c_      GTGACGGCGCCGGTGTGGTTGGCGTCGCCGCCGGAGGTGCATTCGGCGCTGCTAAGTGCT 60
Rv0286_       ATGGCCGCGCCCATCTGGATGGCTTCGCCGCCGGAGGTACATTCGGCGTTGCTTAGCAAT 60
               ** * *****  * *** **** ************** ********* **** **   *

Rv3018c_      GGTCCGGGGCCGGGTTCGTTGCAGGCGGCCGCGGCGGGGTGGAGCGCGTTAAGCGCCGAG 120
Rv0286_       GGTCCGGGCCCGGGTTCGCTAGTGGCGGCTGCCACGGCCTGGAGCCAGCTGAGTGCCGAG 120
              ******** ********* *   ****** **  ***  ******  * * ** ******

Rv3018c_      TACGCCGCTGTGGCGCAAGAGTTGAGCGTGGTGGTGGCCGCGGTGGGGGCCGGGGTGTGG 180
Rv0286_       TATGCCTCGACGGCAGCAGAACTCAGTGGGCTACTGGGGGCGGTACCTGGTTGGGCATGG 180
              ** *** *   ***   ***  * ** * * *  ***  *****    *   ***  ***

Rv3018c_      CAGGGTCCCAGTGCTGAGTTGTTTGTGGCCGCCTATGTGCCGTATGTGGCGTGGTTGGTG 240
Rv0286_       CAGGGGCCCAGCGCGGAGTGGTACGTGGCCGCGCATTTGCCATATGTGGCGTGGCTGACG 240
              ***** ***** ** **** **  ********  ** **** ************ **  *

Rv3018c_      CAGGCCAGTGCGGATAGCGCGGCGGCGGCCGGTGAGCATGAGGCCGCGGCGGCTGGCTAT 300
Rv0286_       CAGGCCAGTGCGGATGCCGCAGGAGCAGCGGCCCAGCACGAGGCCGCCGCGGCGGCCTAC 300
              ***************  *** *  ** ** *   **** ******** ***** * ***

Rv3018c_      GTTTGTGCGTTGGCGGAGATGCCGACGTTGCCGGAGTTGGCGGCCAACCACCTCACGCAT 360
Rv0286_       ACCACTGCCTTGGCAGCCATGCCGACATTAGCGGAGTTGGCCGCCAACCACGTGATTCAC 360
                   *** ***** *  ******** **  ********** ********* * *  **

Rv3018c_      GCGGTGTTGGTGGCGACGAATTTCTTTGGGATCAACACGATCCCGATCGCGCTCAACGAG 420
Rv0286_       ACCGTGTTGGTGGCGACGAATTTCTTTGGGATCAACACGATTCCCATCACGCTCAATGAG 420
               * ************************************** ** *** ******* ***
```

B

```
Rv3018c_      GTGACGGCGCCGGTGTGGTTGGCGTCGCCGCCGGAGGTGCATTCGGCGCTGCTAAGTGCT 60
smeg          ATGACGGCCCCTATCTGGATGGCTCTGCCGCCCGAGGTGCACTCGTCGCTGCTGTCCAGC 60
               ******* **  * *** ****   ****** ******** *** *******

Rv3018c_      GGTCCGGGGCCGGGTTCGTTGCAGGCGGCCGCGGCGGGGTGGAGCGCGTTAAGCGCCGAG 120
smeg          GGCCCAGGCCCCGGGTCGCTGCTGGCCGCCGCGGGGGCGTGGCAGTCGCTCAGCGCCGAA 120
              ** ** ** ** ** *** *** *** ******* ** ****    ** * ********

Rv3018c_      TACGCCGCTGTGGCGCAAGAGTTGAGCGTGGTGGTGGCCGCGGTGGGGGCCGGGGTGTGG 180
smeg          TACGCCGCGGCGGCAGCCGAACTCACGAGTGTGCTGAGCGCGGTGCAGGCCGGCTCGTGG 180
              ******** * ***    **  * *     *** **  *******  ******   ****

Rv3018c_      CAGGGTCCCAGTGCTGAGTTGTTTGTGGCCGCCTATGTGCCGTATGTGGCGTGGTTGGTG 240
smeg          GAAGGTCCGAGTTCCGAGCAGTATGTCGCGGCCCACGCGCCGTATCTGCAGTGGCTCGCG 240
               * ***** *** * ***  ** *** ** *** * * ******* **  **** * * *

Rv3018c_      CAGGCCAGTGCGGATAGCGCGGCGGCGGCCGGTGAGCATGAGGCCGCGGCGGCTGGCTAT 300
smeg          CAGCAGAGCGCCAACAGCGCGGCCGCGGCCGTCCAGCACGAGACCGCGGCCGCGGCGTAC 300
              ***   ** **  * ******** *******   **** *** ******* ** *  **

Rv3018c_      GTTTGTGCGTTGGCGGAGATGCCGACGTTGCCGGAGTTGGCGGCCAACCACCTCACGCAT 360
smeg          TCCACGGCACTGGCCACGATGCCGACCATGGCCGAACTGGCGCTCAACCACACCATGCAC 360
                    **  ****   *********  ** * **  *****  *******  ** ***

Rv3018c_      GCGGTGTTGGTGGCGACGAATTTCTTTGGGATCAACACGATCCCGATCGCGCTCAACGAG 420
smeg          GGTGTGCTCGTGGCCACGAACTTCTTCGGGATCAACACGATCCCGATCGCGCTCAACGAG 420
              *  *** * ***** ***** ***** *********************************
```

Figure 6.7. Presence and absence of Rv1787 in members of the genus *Mycobacterium*. Presence and absence of gene Rv1787 was examined by dot blot analysis using probe 15/16. Gene Rv1787 is absent in species indicated in yellow and present in species indicated in red. The division between fast and slow-growing species are indicated by a dotted line. Underlined species are pathogens. * = *M. farcinogenes* is a slow growing mycobacterium. [(1)] = Rv1787 was confirmed to be present in the genome of this organism by whole genome sequencing and not by dot blot analyses. The taxonomical relationships between members of the genus *Mycobacterium* are based on 16S rRNA gene sequence information and were adapted from data published by Pitulle *et al*. (1992), Shinnick and Good (1994) and Springer *et al*. (1996).


🟨 Probe 15/16 negative
🟥 Probe 15/16 positive
M. fortuitum
M. farcinogenes *
M. senegalense
M. chelonae
M. peregrinum
M. komossense
M. sphagni
M.aichiense
M. gilvum
M.parafortuitum
M. neoaurum
M. diernhoferi
M. abscessus
M. mucogenicum
M. chitae
M. fallax
M. obuense
M. chubuense
M. aurum
M. vaccae
M. confluentis
M. madagascariense
M. flavescens
M. smegmatis
M. thermoresistibile
M. chromogenicum
M. phlei
M. gadium
RAPID GROWERS
SLOW GROWERS
M. triviale
M. simiae
M. genavense
M. interjectum
M. intermedium
M. terrae
M. hibemiae
M. nonchromogenicum
M. cookii
M. xenopi
M. celatum type 1
M. celatum type 2
M. shimoidei
M. gordonae
M. asiaticum
M. tuberculosis
M. africanum
M. canettii
M. microti
M. bovis (1)
M. bovis BCG
M. tuberculosis complex
M. marinum
M. ulcerans
M. leprae (1)
M. scrofulaceum
M. gastri
M. kansasii
M. szulgai
M. malmoense
M. haemophilum
M. intracellulare
M. paratuberculosis (1)
M. avium
M. avium complex

6.4. Discussion

The PE and PPE protein families are large families (99 and 68 members respectively) of repetitive sequence-containing genes of largely unknown function. These genes are found distributed throughout the genome of *M. tuberculosis* and copies are also found situated in the ESAT-6 gene cluster regions (Cole *et al.*, 1998, Gey van Pittius *et al.*, 2001). In contrast to their wide distribution in the genome, little is known about the function of these two protein families. Although nearly all proposed functions remains at most speculative, the most widely accepted hypothesis suggest that these proteins could function as a source of antigenic variability, due to their highly polymorphic C-terminal domains (Cole *et al.*, 1998, Gordon *et al.*, 1999b). Other putative functions include a function as storage proteins for the rare amino acid asparagine (Cole *et al.*, 1999), the inhibition of antigen processing (Cole *et al.*, 1998), an involvement in iron uptake (Rv2123, Rodriguez *et al.*, 1999) and fibronectin binding (Abou-Zeid *et al.*, 1991, Rv1759c - Espitia *et al.*, 1999). Furthermore, it was recently shown that two members of the PGRS subfamily from *M. marinum* (the Rv3812 and Rv1651c orthologues) are essential for replication in host macrophages as well as persistence in granulomas (Ramakrishnan *et al.*, 2000). Additional recent data from Camacho and colleagues (1999) have also suggested that the members of the PPE gene family may be involved in disease pathogenesis, as a transposon mutant of the gene Rv3018c was attenuated for growth in macrophages.

It is unknown why the insertions of the PE and PPE genes within the ESAT-6 gene clusters have occurred and why they are tolerated. It may be possible that the high levels of expression of the ESAT-6 gene clusters (Chapter 4) and hypothesized upregulation under *in vivo* conditions (Addendum 5A) are providing an advantage to the organism by the co-expression of these genes. The presence of the PE and PPE genes within the ESAT-6 gene clusters prompted the investigation of the evolutionary history of these large gene families to determine whether the duplication order of the ESAT-6 gene clusters could be connected to the duplications of the PE and PPE families. It was thought that this may provide clues to the potential functions of these large gene families and may explain the major expansion thereof. The overall topology of the PE phylogenetic tree in this study is highly similar to the tree predicted by Espitia *et al.* (1999), although 19 sequences had been excluded from their calculations. The absence of these sequences, which include the PE proteins belonging to the ESAT-6 gene cluster regions 1 (Rv3872), 2 (Rv3893c) and 3 (Rv0285) leaves a major gap in the study of the evolutionary expansion of this family.

Phylogenetic reconstruction of the evolutionary history of the PE and PPE gene families suggested that these genes have been initially inserted into the ESAT-6 gene cluster region 1 after the first duplication of the clusters, and have been subsequently duplicated together with the regions. The results further indicated that the expansion of the PE and PPE gene families has occurred in unison with the duplications of the ESAT-6 gene cluster regions. After each main duplication event involving an ESAT-6 gene cluster region, a number of secondary subduplications of the PE and PPE genes (in some cases associated with a copy of the ESAT-6 and CFP-10 genes, see Chapter 3, Figure 3.2) have occurred from the ESAT-6 gene cluster regions. This phenomenon seems to have culminated in the duplication of the ESAT-6 gene cluster region 5, from which not only a large number of PE and PPE genes were secondarily multiplied to the rest of the genome, but also within which multiple secondary duplication events of these families have occurred. In addition to this, the evolutionary history predicted by the phylogenetic trees suggests that the highly duplicated PGRS as well as MPTR subfamilies have also been duplicated from region 5. It thus seems as if the PE and PPE genes present within region 5 have an enhanced ability for duplication, which also allowed the expansion of these genes into the highly polymorphic PGRS and MPTR subfamilies. These sequences have become extremely mobile, almost analogous to transposon elements.

It has been suggested previously that it seems as if the occurrence of numerous PE (and PE-PGRS) proteins is restricted to members of the *M. tuberculosis* complex and a few other pathogenic mycobacterial species (Brennan *et al.*, 2001). This was supported by the *in silico* comparative genomics analyses results in the present study which demonstrated the absence of the multiple duplications of this family and the PPE family from the genome of *M. smegmatis*. It was previously found that the genomes of members of the corynebacteria contain one copy of the ESAT-6 gene clusters (region 4, Gey van Pittius *et al.*, 2001), but no copies of the PE or PPE genes could be identified in these organisms (N.C. Gey van Pittius, unpublished results). It is clear from the evolutionary relationships between the members of the genus *Mycobacterium* determined by 16S rRNA sequence information that the *M. smegmatis* species is quite distant from the pathogenic *M.*

tuberculosis complex, and seem to represent an evolutionary much earlier species. It thus seems as if the PE and PPE gene insertion into the second duplication of the ESAT-6 gene clusters (region 1) occurred early on in the evolution of the mycobacteria, after their divergence from the corynebacteria, and that the multiple duplications of these families have coincided with the evolution of the genus, reaching a maximum in the slow-growing pathogenic mycobacteral species. The dot blot results presented in this study confirm the above hypothesis. The genes Rv2123 and Rv3429 were found to be *M. tuberculosis* specific, confirming their recent duplication. In addition to this, the probe 3018 (most probably hybridizing to the Rv0286 orthologues) gave a positive hybridization result to all the species tested, indicating that it is an ancient duplication present in the earliest mycobacteria. The absence of the gene Rv1787 from the genome of *M. smegmatis* came as no surprise, as it has been shown to be absent by the comparative genomic analyses. What was fascinating was the observation that the probe for this gene only hybridized to the members of the slow-growing mycobacteria, indicating a clear evolutionary division between these two phenotypically separated groups. It thus seems as if the duplication of this region (which have been shown in this study to be prone to secondary duplications and the potential origin of the PGRS and MPTR families) had some influence on the growth characteristics of the organisms. To our knowledge no genetic factors have been identified which differentiate between these two divisions of the genus to such a clear extent, making this a novel and highly applicable finding. Whether this differentiation ability is restricted to Rv1787 (the PPE protein in the ESAT-6 gene cluster region 5), or the whole subgroup of PPE proteins surrounding the gene Rv1787 on the phylogenetic tree is at this stage uncertain. In addition to this, it must still be established whether the other genes present within the ESAT-6 gene cluster region 5 also have this differentiation ability. Our results could thus not distinguish whether the differentiation is due to the presence of the PPE genes specifically or the presence of the ESAT-6 gene cluster (region 5) as a whole, and this is the subject of ongoing studies. These results did indicate that the duplications of the PPE and PE gene families and/or the ESAT-6 gene clusters are in some way involved in the differences observed between fast-growing and slow growing mycobacteria.

In conclusion, we aimed to investigate the evolutionary distribution of the PE and PPE gene families in relation to their observed presence within four of the five ESAT-6 gene clusters. We have shown that the expansion of the PE and PPE families are linked to the duplications of the ESAT-6 gene clusters. We have also shown that this association has led to the absence of the multiple duplications of the PE and PPE families in the fast-growing mycobacterium *M. smegmatis*. This includes the members of the multigene PGRS and MPTR subgroups, which are hypothesized to be involved in antigenic variation by virtue of their hypervariable C-terminal domains (Cole *et al*., 1998, Cole, 1999, Gordon *et al*., 1999b). We have also showed that the PPE gene present in ESAT-6 gene cluster region 5 is able to distinguish between mycobacteria belonging to the slow-growing or fast-growing species. This result is highly significant with regard to the exploitation of this sequence as a potential epidemiological tool to differentiate between fast- and slow-growing species. It is also an important step in advancing the study of the differences between the members of these two divisions. This research contributes to the development of an understanding of the PE and PPE gene families, in terms of stability, absence/presence of the PE and PPE genes within the genomes of various mycobacteria, their assosiation with the ESAT-6 gene clusters and links to growth rate and cell wall structure.

NOTE: *Ms. S. Sampson (Department of Medical Biochemistry, Faculty of Health Sciences, University of Stellenbosch, Tygerberg, South Africa) and Dr. H. Lee (Dept. of Microbiology, Korean Institute of Tuberculosis, Seoul, South Korea) are thanked for their invaluable contributions towards the PPE phylogeny and dot blot analyses, respectively.*

CHAPTER SEVEN

DISCUSSION AND FUTURE DIRECTIONS

"Imbeciles! . . .Tuberculosis! Everybody knows the true remedy, which would be the paying of sufficient wages, and the tearing down of the filthy tenements into which the laborers are packed - those who are the most useful and the most unfortunate among our population! But needless to say, no one wants that remedy, so we go round begging the workingmen not to spit on the sidewalks. "

Damaged Goods - Upton Sinclair

7.1. The mycosin proteases

Exported proteases are commonly associated with virulence in bacterial pathogens, yet there is a paucity of information regarding their role in *Mycobacterium tuberculosis* (Goguen *et al.*, 1995, Brown *et al.*, 2000). *Mycobacterium tuberculosis* possesses over 70 protease genes, which includes the closely related family of five subtilisin-like serine protease genes (the *mycP* genes) identified in the present study. These genes make up the largest protease family in *M. tuberculosis*, and encode transmembrane subtilases, termed the mycosins. Using Southern blotting, multiple copies of the *mycP* genes were shown to also be present in the genomes of other mycobacterial species, and all the mycosins were found to be constitutively expressed in *M. tuberculosis*. Mycosin-1 was found to not be expressed in the attenuated vaccine strain *M. bovis* BCG (*bacille de Calmette et Guérin*) although the gene is present in the genome of this organism. Closer inspection revealed that the gene *mycP1* is situated 3700 bp (four ORF's) from the RD1 deletion region in the genome of *M. bovis* BCG (Mahairas *et al.*, 1996), indicating that the deletion of this region may have removed regulatory sequences required for the expression of the gene.

Subcellular localization of selected members of the mycosins showed that the proteins are secreted, membrane bound, cell wall-associated proteases that are shed by an unknown mechanism (actively or passively) from the cell wall during growth of *M. tuberculosis* under *in vitro* and *in vivo* conditions. In support of this, it was also shown previously that at least one of the mycosins is able to elicit delayed-type hypersensitivity (DTH) reactions only in guinea pigs immunized with live mycobacteria, indicating the release of protein only during active growth of the organism (Romain *et al.*, 1993). As DTH is known to be a strictly T-cell dependant immune reaction, the possibility that the mycosins may also be recognized by the T-cell population mediating the cell mediated immune response (CMI) reaction was investigated. Whole blood assay results indicated that the extracellularly located mycosin proteases are able to elicit low levels of T-cell dependent cellular proliferation, with concomitant production of relatively high levels of IFN-γ. An interesting observation of the primary study was the indication that the mycosins may be specifically recognized only in healthy individuals, which points to the possibility that these proteins may be involved in protective immunity. Further results indicated that the mycobacterial mycosins are predominantly recognized by

Mantoux positive individuals, thereby being able to stimulate the CMI response. These mycosins are thus interesting for their potential use as components for future subunit vaccines against tuberculosis. However, it is acknowledged that the sample bases of both the primary and secondary experiments were not optimal, and that it will be necessary to increase the number of subjects in future investigations to determine the full extent of antigenicity of the members of the mycosin family.

A number of protease activity assays were performed to determine the substrate specificty, activity, and conditions of activity of the mycosins, and thereby to obtain insight into their functions *in vivo*. Despite all efforts, no protease activity could be detected. The most likely reason for not observing protease activity for the recombinant mycosins (which could include incorrect folding of the fusion protein, the absence of or incorrect processing of the prepromycosins, substrate specificity, pH, temperature, cofactors etc.), was provided by the fact that the mycosins revealed characteristics shared only by the lantibiotic peptidases and the proprotein convertases (Siezen and Leunissen, 1997). These subtilases are in highly specialized families of proprotein processing proteases, suggesting that the activity of the mycosins may also be highly substrate specific. Almost all of the known proteases that have been examined previously were identified from an observed, yet unknown protease activity, thereby immediately providing a usable substrate (see for example Butler *et al.*, 1996). The fact that the mycosins were identified from their gene sequences was a limitation in this study, as it made the identification of a possible substrate, as well as optimal conditions for activity, virtually impossible. It is clear from this study that the substrate of the mycosins has to be identified before any further activity analyses could be done and the potential role of the mycosins in disease pathology could be evaluated. To obtain clues to a potential substrate and the possible function of these proteases in the physiology of *M. tuberculosis*, the genetic environment of the mycosin genes was studied in the then newly released whole genome sequence of *M. tuberculosis* (Cole *et al.*, 1998). This revealed the fascinating fact that these mycosin genes were not duplicated alone in the genome of the organism, but were actually found situated in a cluster of between 6 and 12 genes, which were duplicated five times in the genome. The identification of this gene cluster led to the hypothesis that these genes may encode proteins that function together. Thus, it was important to study the gene clusters in which the mycosins were situated, as the function of the mycosin proteases may be closely linked to the functions of the other genes in these clusters.

7.2. The ESAT-6 gene clusters

The most interesting observation in the study of the gene cluster regions was the fact that each of these regions contained members of the previously identified, immunologically important ESAT-6 T-cell antigen family (Andersen *et al.*, 1995, Sørensen *et al.*, 1995, Berthet *et al.*, 1998, Van Pinxteren *et al.*, 2000), leading to these regions being named the ESAT-6 gene cluster regions. A comparative genomics approach was implemented to establish the relationship between the multiple copies of the ESAT-6 gene cluster as well as to determine the evolutionary history of the cluster duplication. The results demonstrated that the ESAT-6 gene cluster is of ancient origin, with the progenitor cluster, region 4, also being present in members of other genera (for example *Corynebacterium*). It also demonstrated that this cluster seem to be a feature of the high G+C gram-positive bacteria as it is present in and restricted to the genomes of other members of the Firmicutes such as *Corynebacterium diphtheriae* and *Streptomyces coelicolor.* Furthermore, it was shown to be duplicated multiple times only in *Mycobacterium tuberculosis* and other mycobacteria. It is thus tempting to speculate that the multiple copies of the ESAT-6 gene cluster present only in the genomes of the mycobacteria, may point to an important function for these clusters in the mycobacterial physiology that differentiates this genus from the other genera of the Firmicutes.

The observation that live mycobacteria generate a much more efficient protective immunity than killed bacteria (Andersen, 1997), but that both sensitize animals for a DTH reaction, has been linked to the fact that live bacteria secrete peptides early in infection that are needed to recruit protective T-cells. The key antigenic determinants of the culture filtrate was shown to be smaller than the 10kDa fraction (Boesen et al., 1995), in which multiple copies of the small immunodominant ESAT-6 protein family members were identified (Sørensen *et al.*, 1995, Berthet *et al.*, 1998, Alderson *et al.*, 2000). The fact that ESAT-6 elicits a high level of interferon-gamma from memory effector cells during the first phase of a protective immune response is important because mycobacterial diseases are generally characterized by strong Th1 responses and high levels of interferon-gamma (Andersen, 1997). However, it is not known what advantage the bacterium obtains from the secretion of the multiple immunodominant copies of the ESAT-6 proteins. It may be possible that the induction of a massive host immune response, resulting in intense inflammation, tissue destruction, caseous necrosis and the formation of cavitory lesions may facilitate the spread of disease by the release of the bacteria from the damaged airways of susceptible individuals unable to contain the infection.

Although the function of the ESAT-6 gene clusters and their role in the growth and pathogenicity of the mycobacteria is still unknown, data originating from a number of studies have indicated that the genes that form part of these clusters may play a very important role in the growth of these organisms and in tuberculosis infection. For example, Wards and coworkers (2000) showed that an *esat-6/cfp-10* knockout mutant of *M. bovis* was less virulent than its parent if gross pathology, histopathology and mycobacterial culture of tissues were taken into account. A comparable decrease in virulence was also shown in a knockout of Rv3871 (another gene present in ESAT-6 gene cluster region 1) by the same authors. Similarly, Mahairas and colleagues (1996) and others (Behr *et al.*, 1999, Brosch *et al.*, 2000a) have implicated the RD1 deletion region (containing most of the genes of ESAT-6 gene cluster region 1) in the attenuation of *M. bovis* BCG, supporting the importance of this region with regard to virulence. Further clues are provided by the fact that these genes are expressed under oxygen tension similar to that observed in intracellular residence (Imboden *et al.*, 1998), and the fact that the genome of *M. leprae*, which is commonly though to contain the minimal gene set for a pathogenic mycobacterium (Vissa and Brennan, 2001), contains at least two and maybe even three functional copies of the ESAT-6 gene clusters. The ESAT-6 family member Rv0288 (belonging to gene cluster region3) is also highly downregulated in the avirulent strain *M. tuberculosis* H37Ra (Rindi *et al.*, 1999). The multiplicity of these gene clusters, their immunological significance as well as the links to pathogenicity suggest that they have an important function in the mycobacteria and are worth further investigation.

During this study, the topical subject of diagnosis of *M. tuberculosis* infection using the members of the ESAT-6 protein family (specifically ESAT-6 and CFP-10) was re-evaluated. The ESAT-6 and CFP-10 proteins have been evaluated over the past few years as diagnostic agents (see for example Van Pinxteren *et al.*, 2000), as there is a constant search for new and effective diagnostic tests for the determination of *M. tuberculosis* infection due to the non-specificity of the current tuberculin test (Andersen *et al.*, 2000). Both of these proteins have shown promise as a tool to differentiate between BCG vaccination, *M. avium* infection and *M. tuberculosis* infection (Ravn *et al.*, 1999, Colangeli *et al.*, 2000, Arend *et al.*, 2000b), but contrary to common belief (Harboe *et al.*, 1996, Elhay *et al.*, 1998, Ravn *et al.*, 1999, Arend *et al.*, 2000a, Arend *et al.*, 2000b, Van Pinxteren *et al.*, 2000, Andersen *et al.*, 2000, Arend *et al.*, 2001a, Arend *et al.*, 2001b), this study showed that these proteins are not *M. tuberculosis* specific (Gey van Pittius *et al.*, 2001). As the genes for these proteins are also present in fast growing environmental mycobacterial species (with a high percentage of protein homology), the presence of environmental mycobacteria may interfere with diagnostic tests based on these antigens (Gey van Pittius *et al.*, 2001). This should be especially evident in developing countries where environmental mycobacteria is present in large amounts (Vekemans *et al.*, 2001).

The ESAT-6 gene family consists of members of the ESAT-6 and CFP-10 subfamilies, and encode small proteins, which are potent T-cell antigens of *M. tuberculosis*, but which are secreted without ordinary sec-dependent secretion system signals (Andersen *et al.*, 1995, Sørensen *et al.*, 1995). It has thus been widely accepted in the literature that these proteins must have some other, still unknown mechanism of transport to procure the release into the extracellular environment (Tekaia *et al.*, 1999). An in depth bioinformatics analysis of the other genes present within the ESAT-6 gene clusters revealed that all of these genes may encode proteins that are directed to the cell wall/membrane where they are potentially involved in binding protein dependent transport systems subcomponent functions (energy provision, pore formation, etc). Thus, the proteins encoded by the ESAT-6 gene clusters may be able to function together to provide an active transport system, and it was hypothesized that this system may be able to transport the sec-independently secreted members of the ESAT-6 T-cell antigen family across the mycobacterial membrane.

In support of the abovementioned hypothesis, the results presented in this study have shown that at least one of the ESAT-6 gene clusters is expressed as a single polycistronic RNA and the 11 genes situated in this cluster thus form one single operon. Furthermore, the promoter driving the expression of this operon, P_{ESREG3}, was identified and its activity characterized. The verification of the polycistronic nature of the ESAT-6 gene clusters as well as the identification of the promoter P_{ESREG3} is an important step in the elucidation of the function and regulation of members of the ESAT-6 family as well as its putative biosynthetic gene clusters. Currently, there is only a limited amount of information available on the mycobacterial transcriptional machinery (Mulder *et al.*, 1997), making this study important as a basis for investigating the mechanisms of antigen expression in *M. tuberculosis*.

During the present study, a novel method of examining secretion mechanisms in the mycobacteria was developed, making use of double transformation after cosmid integration. The results have shown that secretion of members of the ESAT-6 protein family is dependent on the presence of the ESAT-6 gene cluster regions, as was illustrated in the case of ESAT-6 and ESAT-6 gene cluster region 1. Furthermore, this secretion seems to be region-specific, to such an extent that even orthologous regions between different mycobacteria are unable to cross-secrete ESAT-6. A possible model for the secretion apparatus was also constructed, based on the putative functions of the proteins encoded by the ESAT-6 gene cluster regions. The results obtained from this investigation provides an important basis for the study of dedicated mycobacterial transport and secretion mechanisms, as well as for the understanding of T-cell antigen secretion in the mycobacteria. As the mycobacterial ESAT-6 gene clusters contain a number of features of quorum sensing and lantibiotic biosynthetic, transport and processing operons (Sahl and Bierbaum, 1998), it is hypothesized that the members of the ESAT-6 family may be secreted as signaling molecules and are possibly involved in the regulation of expression of genes during intracellular residence of the bacterium.

7.3. The history of the PE/PPE gene family expansion

Each of the ESAT-6 gene cluster regions encode proteins involved in energy provision and membrane pore formation, and have been shown in the present study to be involved in the transport of the potent T-cell antigens belonging to the ESAT-6 protein family. In addition, there are two families of genes present within the ESAT-6 gene clusters (named the PE and PPE gene families), which seem to be anomalous as they do not appear to be involved in the transport system, are not found in region 4, and members of these two families are found distributed throughout the genome of *M. tuberculosis* (Cole *et al.*, 1998). To gain an insight into the evolutionary history of these genes, the PE and PPE gene families were investigated with the aim of determining the ancestral genes from both families, as well as to determine whether the presence of copies of these genes within the ESAT-6 gene clusters have any significance with regard to the expansion of both families. The results presented in this study have shown that the expansion of the PE and PPE families is linked to the duplications of the ESAT-6 gene clusters. Furthermore, it is clear that this association has led to the absence of the multiple duplications of the PE and PPE families in the fast-growing mycobacterium *M. smegmatis*. This includes the members of the multigene PGRS and MPTR subgroups, which are hypothesized to be involved in antigenic variation by virtue of their hypervariable C-terminal domains (Cole *et al.*, 1998, Cole, 1999, Gordon *et al.*, 1999b). A surprising result was obtained with the dot blot analysis of the PPE gene present in ESAT-6 gene cluster region 5, which demonstrated that this gene was not present in the genomes of any of the fast-growing mycobacterial species tested and is clearly able to distinguish between mycobacteria belonging to the slow-growing or fast-growing species. This indicates that the PPE/ESAT-6 gene cluster region 5 may be involved in some function which differentiates these two groups of mycobacteria. This research contributes to the development of an understanding of the PE and PPE gene families, in terms of stability, absence/presence of the PE and PPE genes within the genomes of various mycobacteria, their assosiation with the ESAT-6 gene clusters and links to growth rate and cell wall structure.

7.4. Future directions

The results of this investigation satisfied most of the aims set out at the inception of this study. However, a number of new questions have also arisen, which may present excellent challenges for future investigations.

The recommendations for future studies is as follows:

- cloning, expression and antibody generation with mycosin-4 and -5,
- more extensive T-cell assays with larger subject groups to determine the full extent of the mycosin family antigenicity,
- the identification of a substrate for the mycosins, possibly the members of the ESAT-6 protein family,
- subsequent protease activity assays on the substrate to determine optimal activity parameters - (initial studies may focus on investigating whether the mycosins are able to cleave the members of the ESAT-6 protein family,
- *in silico* modelling of substrate binding site,
- *in silico* investigation of the presence of the ESAT-6 gene clusters in the newly sequenced genomes of other members of the high G+C Gram-positives, namely *Thermobifida fusca* and *Clavibacter michiganensis* as well as other members of the genus *Mycobacterium*, namely *Mycobacterium ulcerans, Mycobacterium miroti* and *Mycobacterium marinum*,
- investigation to determine the influence of gene knockouts described by Wards *et al.* (2000) on expression and secretion of ESAT-6,
- investigation to determine the influence of gene knockouts of the singular ESAT-6 gene clusters in members of the genus *Corynebacterium* and/ or *Streptomyces*,
- cloning, expression and purification of *M. smegmatis* ESAT-6 proteins to investigate whether host cellular immune response are able to distinguish between the ESAT-6 and CFP-10 proteins secreted from environmental mycobacteria and *M. tuberculosis*,
- identification of operon structures in other ESAT-6 gene cluster regions
- identification of other primary or secondary promoters present in the ESAT-6 gene clusters,

- investigation of the regulation of expression of the gene clusters,
- gene deletion in cosmids to determine which genes are essential for the ESAT-6 secretion system to function efficiently,
- investigation to determine whether different regions in a certain species are able to cross-secrete members of the ESAT-6 protein family,
- investigation to determine whether addition of ESAT-6 protein to cultures has an influence on growth characteristics and signalling between organisms,
- extensive Southern blotting analyses to determine the presence of different PE and PPE genes and the ESAT-6 gene clusters in different species of mycobacteria.

CHAPTER EIGHT

CONCLUSION

"But most important, the ancient foe of man, known as consumption, the great white plague, tuberculosis, or by whatever other name, is on the way to being reduced to a minor ailment of man. The future appears bright indeed, and the complete eradication of the disease is in sight."

The Conquest of Tuberculosis – S. A. Waksman (1964)

In this study, a closely related family of five subtilisin-like serine protease genes (the *mycP* genes), encoding transmembrane subtilases (termed the mycosins) was identified. These genes were cloned and characterized with the aim of elucidating their potential function in *M. tuberculosis* and to investigate the possibility that they may be involved in the virulence mechanisms of the organism. Subcellular localization of selected members of the mycosins revealed that these proteins are secreted, membrane bound, cell wall-associated subtilisin-like serine proteases that may be shed from the cell wall during growth of *M. tuberculosis* under *in vitro* and *in vivo* conditions. Whole blood assay results indicated that the extracellularly located mycosin proteases are able to elicit low levels of T-cell dependent cellular proliferation, with concomitant production of relatively high levels of IFN-γ, and may be involved in protective immunity. The ability of the mycosins to stimulate the CMI response presents a potential opportunity for them to be evaluated as components for future subunit vaccines against tuberculosis. No protease activity could be attributed to the mycosins, but protein sequence analyses revealed that these mycobacterial subtilases share characteristics with the lantibiotic peptidases and the eukaryotic proprotein convertases. This indicates that they may be involved in the specific activation of secreted proteins and that substrate specificity may thus be extremely crucial. To obtain further clues into the possible function of these enzymes, their immediate genetic environment was studied, which revealed that the mycosins actually form part of a gene cluster of between 6 and 12 genes, which are duplicated five times in the genome of *M. tuberculosis*. This gene cluster also contains members of the previously identified immunologically important ESAT-6 T-cell antigen family and have thus been named the ESAT-6 gene clusters. In addition to this, the gene clusters contain a number of genes potentially involved in different aspects of protein transport. Comparative genomics analyses revealed that the presence of the ESAT-6 gene cluster seems to be a characteristic shared by all high G+C gram-positive bacteria and that multiple duplications of this cluster have occurred and are maintained only within the genomes of members of the genus *Mycobacterium*. One of the ESAT-6 gene clusters was shown to be expressed as a single polycistronic RNA, forming an operon structure. The promoter for this operon, P_{ESREG3}, was also identified and its activity characterized. This led to the hypothesis that these operons may encode proteins that function together for the active transport and processing of the members of the sec-independently secreted ESAT-6 T-cell antigen family. Subsequent secretion analyses results have shown that secretion of members of the ESAT-6 protein family is dependent on the presence of the

ESAT-6 gene cluster regions, confirming the putative transport associated functions of the ESAT-6 gene cluster-encoded proteins. The mycobacterial ESAT-6 gene clusters contain a number of features of quorum sensing and lantibiotic operons, and an extensive review of the literature has led to the hypothesis that the members of the ESAT-6 family may be secreted as signaling molecules and may be involved in the regulation of expression of genes during intracellular residence of the bacterium within the macrophage. Finally, the results of the investigation of the evolutionary history of the PE and PPE gene families have shown that the expansion of these families are linked to the duplications of the ESAT-6 gene clusters, and that the highly polymorphic PGRS and MPTR subgroups are the direct result of the most recent duplication events. This association is supported by the absence of the multiple copies of the PE and PPE families in the genome of the fast-growing mycobacterium *M. smegmatis*. Dot blot analyses showed that the PPE gene present in ESAT-6 gene cluster region 5 is able to distinguish between mycobacteria belonging to the slow-growing or fast-growing species, indicating a function for these genes and/or the ESAT-6 gene clusters in the phenotypical differences distinguishing these two groups of mycobacteria.

In conclusion, this study has highlighted several important aspects of mycobacterial genomics and has greatly contributed to the current body of knowledge concerning the mechanisms of antigen secretion. This work not only presented the identification and characterization of a novel mycobacterial protease gene family, it also led to the identification of an immunologically important gene cluster within which this family is situated. It provided proof for the presence of the genes of important secreted antigens within dedicated transporter-encoding operons and also identified the promoter driving the expression of at least one of the important T-cell antigens of the ESAT-6 gene family. Most importantly, it provided invaluable insight into the mechanisms of sec-independent protein secretion in the mycobacteria. Finally, this study provided evidence that the duplications of the ESAT-6 gene clusters/PE/PPE proteins are involved in the division between fast and slow-growing mycobacterial species. This work lays the foundation for further research into the characterization of the specific functions of the members of the mycosin, ESAT-6, PE and PPE multigene families. The data presented in this study, supported by the discussed literature, indicates that these gene families may be important in disease pathogenesis and may present interesting candidates for drug design and to evaluate as components for anti- tuberculosis subunit vaccines.

LIST OF REFERENCES

A

Abou-Zeid, C., Garbe, T., Lathigra, R., Wiker, H. G., Harboe, M., Rook, G. A., and Young, D. B. (1991). Genetic and immunological analysis of *Mycobacterium tuberculosis* fibronectin-binding proteins. *Infect.Immun.* **59**, 2712-2718.

Agrawal, S., Thomas, N. S., Dhanikula, A. B., Kaul, C. L., and Panchagnula, R. (2001). Antituberculosis drugs and new drug development. *Curr.Opin.Pulm.Med.* **7**, 142-147.

Ahmad, S., Amoudy, H. A., Thole, J. E., Young, D. B., and Mustafa, A. S. (1999). Identification of a novel protein antigen encoded by a *Mycobacterium tuberculosis*-specific RD1 region gene. *Scand.J.Immunol.* **49**, 515-522.

Alderson, M. R., Bement, T., Day, C. H., Zhu, L., Molesh, D., Skeiky, Y. A., Coler, R., Lewinsohn, D. M., Reed, S. G., and Dillon, D. C. (2000). Expression cloning of an immunodominant family of *Mycobacterium tuberculosis* antigens using human CD4(+) T cells. *J.Exp.Med.* **191**, 551-560.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J.Mol.Biol.* **215**, 403-410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.

Alvarez, E. and Tavel, E. (1885). Reche'rches sur le bacille'de Lustgarden. *Archives de Physiologie Normal et Pathologique* **6**, 303-321.

Andersen, P., Askgaard, D., Ljungqvist, L., Bentzon, M. W., and Heron, I. (1991a). T-cell proliferative response to antigens secreted by *Mycobacterium tuberculosis. Infect.Immun.* **59**, 1558-1563.

Andersen, P., Askgaard, D., Ljungqvist, L., Bennedsen, J., and Heron, I. (1991b). Proteins released from *Mycobacterium tuberculosis* during growth. *Infect.Immun.* **59**, 1905-1910.

Andersen, P. and Heron, I. (1993). Specificity of a protective memory immune response against *Mycobacterium tuberculosis. Infect.Immun.* **61** , 844-851.

Andersen, P. (1994a). The T cell response to secreted antigens of *Mycobacterium tuberculosis. Immunobiology* **191**, 537-547.

Andersen, P. (1994b). Effective vaccination of mice against *Mycobacterium tuberculosis* infection with a soluble mixture of secreted mycobacterial proteins. *Infect.Immun.* **62**, 2536-2544.

Andersen, P., Andersen, A. B., Sorensen, A. L., and Nagai, S. (1995). Recall of long-lived immunity to *Mycobacterium tuberculosis* infection in mice. *J.Immunol.* **154**, 3359-3372.

Andersen, P. (1997). Host responses and antigens involved in protective immunity to *Mycobacterium tuberculosis* . *Scand.J.Immunol.* **45**, 115-131.

Andersen, P., Munk, M. E., Pollock, J. M., and Doherty, T. M. (2000). Specific immune-based diagnosis of tuberculosis. *Lancet* **356**, 1099-1104.

Anilkumar, G., Chauhan, M. M., and Ajitkumar, P. (1998). Cloning and expression of the gene coding for FtsH protease from *Mycobacterium tuberculosis* H37Rv. *Gene* **214**, 7-11.

Arend, S. M., Geluk, A., van Meijgaarden, K. E., van Dissel, J. T., Theisen, M., Andersen, P., and Ottenhoff, T. H. (2000a). Antigenic equivalence of human T-cell responses to *Mycobacterium tuberculosis*-specific RD1-encoded protein antigens ESAT-6 and culture filtrate protein 10 and to mixtures of synthetic peptides. *Infect.Immun.* **68**, 3314-3321.

Arend, S. M., Andersen, P., van Meijgaarden, K. E., Skjot, R. L., Subronto, Y. W., van Dissel, J. T., and Ottenhoff, T. H. (2000b). Detection of active tuberculosis infection by T cell responses to early- secreted antigenic target 6-kDa protein and culture filtrate protein 10. *J.Infect.Dis.* **181**, 1850-1854.

Arend, S. M., Ottenhoff, T. H., Andersen, P., and van Dissel, J. T. (2001a). Uncommon presentations of tuberculosis: the potential value of a novel diagnostic assay based on the *Mycobacterium tuberculosis*-specific antigens ESAT-6 and CFP-10. *Int.J.Tuberc.Lung Dis.* **5**, 680-686.

Arend, S. M., Engelhard, A. C., Groot, G., de Boer, K., Andersen, P., Ottenhoff, T. H., and van Dissel, J. T. (2001b). Tuberculin skin testing compared with T-cell responses to *Mycobacterium tuberculosis*-specific and nonspecific antigens for detection of latent infection in persons with recent tuberculosis contact. *Clin.Diagn.Lab Immunol.* **8**, 1089-1096.

B

Bairoch, A. (1991). PROSITE: a dictionary of sites and patterns in proteins. *Nucleic Acids Res.* **19 Suppl**, 2241-2245.

Baker, D., Silen, J. L., and Agard, D. A. (1992). Protease pro region required for folding is a potent inhibitor of the mature enzyme. *Proteins* **12**, 339-344.

Bange, F. C., Collins, F. M., and Jacobs, W. R., Jr. (1999). Survival of mice infected with *Mycobacterium smegmatis* containing large DNA fragments from *Mycobacterium tuberculosis*. *Tuber.Lung Dis.* **79**, 171-180.

Barnes, P. F., Bloch, A. B., Davidson, P. T., and Snider, D. E., Jr. (1991). Tuberculosis in patients with human immunodeficiency virus infection. *N.Engl.J.Med.* **324**, 1644-1650.

Barr, P. J. (1991). Mammalian subtilisins: the long-sought dibasic processing endoproteases. *Cell* **66**, 1-3.

Barrett, A. J. and Rawlings, N. D. (1991). Types and families of endopeptidases. *Biochem.Soc.Trans.* **19**, 707-715.

Barrett, A. J. (1994). Classification of peptidases. *Methods Enzymol.* **244**, 1-15.

Barrett, A. J. and Rawlings, N. D. (1995). Families and clans of serine peptidases. *Arch.Biochem.Biophys.* **318**, 247-250.

Barry, C. E., III and Mdluli, K. (1996). Drug sensitivity and environmental adaptation of mycobacterial cell wall components. *Trends Microbiol* **4**, 275-281.

Barry, C. E., III (2001a). Interpreting cell wall 'virulence factors' of *Mycobacterium tuberculosis*. *Trends Microbiol* **9**, 237-241.

Barry, C. E., III (2001b). *Mycobacterium smegmatis*: and absurd model for tuberculosis? *Trends Microbiol* **9**, 473-474.

Barsom, E. K. and Hatfull, G. F. (1997). A putative ABC-transport operon of *Mycobacterium smegmatis*. *Gene* **185**, 127-132.

Bashyam, M. D., Kaushal, D., Dasgupta, S. K., and Tyagi, A. K. (1996). A study of mycobacterial transcriptional apparatus: identification of novel features in promoter elements. *J.Bacteriol.* **178**, 4847-4853.

Behr, M. A., Wilson, M. A., Gill, W. P., Salamon, H., Schoolnik, G. K., Rane, S., and Small, P. M. (1999). Comparative genomics of BCG vaccines by whole-genome DNA microarray. *Science* **284**, 1520-1523.

Berthet, F. X., Rasmussen, P. B., Rosenkrands, I., Andersen, P., and Gicquel, B. (1998). A *Mycobacterium tuberculosis* operon encoding ESAT-6 and a novel low- molecular-mass culture filtrate protein (CFP-10). *Microbiology* **144 (Pt 11)**, 3195-3203.

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 Pt 12**, 3205-3216.

Bishai, W. R., Dannenberg, A. M., Jr., Parrish, N., Ruiz, R., Chen, P., Zook, B. C., Johnson, W., Boles, J. W., and Pitt, M. L. (1999). Virulence of *Mycobacterium tuberculosis* CDC1551 and H37Rv in rabbits evaluated by Lurie's pulmonary tubercle count method. *Infect.Immun.* **67**, 4931-4934.

Boesen, H., Jensen, B. N., Wilcke, T., and Andersen, P. (1995). Human T-cell responses to secreted antigen fractions of *Mycobacterium tuberculosis*. *Infect.Immun.* **63**, 1491-1497.

Bolhuis, A., Matzen, A., Hyyrylainen, H. L., Kontinen, V. P., Meima, R., Chapuis, J., Venema, G., Bron, S., Freudl, R., and van Dijl, J. M. (1999). Signal peptide peptidase- and ClpP-like proteins of *Bacillus subtilis* required for efficient translocation and processing of secretory proteins. *J.Biol.Chem.* **274**, 24585-24592.

Braibant, M., Gilot, P., and Content, J. (2000). The ATP binding cassette (ABC) transport systems of *Mycobacterium tuberculosis*. *FEMS Microbiol Rev.* **24**, 449-467.

Brandt, L., Oettinger, T., Holm, A., Andersen, A. B., and Andersen, P. (1996). Key epitopes on the ESAT-6 antigen recognized in mice during the recall of protective immunity to *Mycobacterium tuberculosis*. *J.Immunol.* **157**, 3527-3533.

Brandt, L., Elhay, M., Rosenkrands, I., Lindblad, E. B., and Andersen, P. (2000). ESAT-6 subunit vaccination against *Mycobacterium tuberculosis*. *Infect.Immun.* **68**, 791-795.

Braun, P. and Tommassen, J. (1998). Function of bacterial propeptides. *Trends Microbiol* **6**, 6-8.

Braunstein, M. and Belisle, J. T. (2000). Genetics of protein secretion. In 'Molecular genetics of mycobacteria'. (Eds. G. F. Hatfull and W. R. Jacobs, Jr.) (ASM Press)

Braxton, S. and Wells, J. A. (1992). Incorporation of a stabilizing Ca(2+)-binding loop into subtilisin BPN'. *Biochemistry* **31**, 7796-7801.

Brennan, M. J., Delogu, G., Chen, Y., Bardarov, S., Kriakov, J., Alavi, M., and Jacobs, W. R., Jr. (2001). Evidence that mycobacterial PE_PGRS proteins are cell surface constituents that influence interactions with other cells. *Infect.Immun.* **69** , 7326-7333.

Brennan, P. J. and Nikaido, H. (1995). The envelope of mycobacteria. *Annu.Rev.Biochem.* **64**, 29-63.

Brosch, R., Gordon, S. V., Billault, A., Garnier, T., Eiglmeier, K., Soravito, C., Barrell, B. G., and Cole, S. T. (1998). Use of a *Mycobacterium tuberculosis* H37Rv bacterial artificial chromosome library for genome mapping, sequencing, and comparative genomics. *Infect.Immun.* **66**, 2221-2229.

Brosch, R., Philipp, W. J., Stavropoulos, E., Colston, M. J., Cole, S. T., and Gordon, S. V. (1999). Genomic analysis reveals variation between *Mycobacterium tuberculosis* H37Rv and the attenuated *M. tuberculosis* H37Ra strain. *Infect.Immun.* **67**, 5768-5774.

Brosch, R., Gordon, S. V., Pym, A., Eiglmeier, K., Garnier, T., and Cole, S. T. (2000a). Comparative genomics of the mycobacteria. *Int.J.Med.Microbiol.* **290**, 143-152.

Brosch, R., Gordon, S. V., Eiglmeier, K., Garnier, T., and Cole, S. T. (2000b). Comparative genomics of the leprosy and tubercle bacilli. *Res.Microbiol.* **151**, 135-142.

Brosch, R., Gordon, S. V., Buchrieser, C., Pym, A. S., Garnier, T., and Cole, S. T. (2000c). Comparative genomics uncovers large tandem chromosomal duplications in *Mycobacterium bovis* BCG Pasteur. *Yeast* **17**, 111-123.

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**, 452-458.

Brown, G. D., Dave, J. A., Gey van Pittius, N. C., Stevens, L., Ehlers, M. R., and Beyers, A. D. (2000). The mycosins of *Mycobacterium tuberculosis* H37Rv: a family of subtilisin-like serine proteases. *Gene* **254**, 147-155.

C

Calmette, A. and Guerin, C. (1920). *Ann.Inst.Pasteur* **34**, 553.

Calmette, A. (1927). 'La vaccination preventive contre la tuberculose.' (Mason et C^{ie}: Paris.)

Calmette, A. (1928). La vaccination preventive de la tuberculose par BCG (Bacille Calmette-Guerin). *Ann.Inst.Pasteur* **12**, 1-58.

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. *Mol.Microbiol.* **34**, 257-267.

Cameron, R. M., Stevenson, K., Inglis, N. F., Klausen, J., and Sharp, J. M. (1994). Identification and characterization of a putative serine protease expressed *in vivo* by *Mycobacterium avium* subsp. *paratuberculosis*. *Microbiology* **140 (Pt 8)**, 1977-1982.

Carter, P. and Wells, J. A. (1990). Functional interaction among catalytic residues in subtilisin BPN'. *Proteins* 7, 335-342.

Chaparas, S. D., Maloney, C. J., and Hedrick, S. R. (1970). Specificity of tuberculins and antigens from various species of mycobacteria. *Am.Rev.Respir.Dis.* **101**, 74-83.

Chiodini, R. J., Van Kruiningen, H. J., and Merkal, R. S. (1984). Ruminant paratuberculosis (Johne's disease): the current status and future prospects. *Cornell Vet.* **74**, 218-262.

Christie, P. J. (2001). Type IV secretion: intercellular transfer of macromolecules by systems ancestrally related to conjugation machines. *Mol.Microbiol* **40**, 294-305.

Chubb, A. J., Woodman, Z. L., Silva Tatley, F. M., Hoffmann, H. J., Scholle, R. R., and Ehlers, M. R. (1998). Identification of *Mycobacterium tuberculosis* signal sequences that direct the export of a leaderless beta-lactamase gene product in *Escherichia coli*. *Microbiology* **144 (Pt 6)**, 1619-1629.

Clemens, D. L. and Horwitz, M. A. (1995). Characterization of the *Mycobacterium tuberculosis* phagosome and evidence that phagosomal maturation is inhibited. *J.Exp.Med.* **181**, 257-270.

Colangeli, R., Spencer, J. S., Bifani, P., Williams, A., Lyashchenko, K., Keen, M. A., Hill, P. J., Belisle, J., and Gennaro, M. L. (2000). MTSA-10, the product of the Rv3874 gene of *Mycobacterium tuberculosis*, elicits tuberculosis-specific, delayed-type hypersensitivity in guinea pigs. *Infect.Immun.* **68**, 990-993.

Colditz, G. A., Brewer, T. F., Berkey, C. S., Wilson, M. E., Burdick, E., Fineberg, H. V., and Mosteller, F. (1994). Efficacy of BCG vaccine in the prevention of tuberculosis. Meta-analysis of the published literature. *JAMA* **271**, 698-702.

Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V., Eiglmeier, K., Gas, S., Barry, C. E., III, 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., and Barrell, B. G. (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**, 537-544.

Cole, S. T. (1998). Comparative mycobacterial genomics. *Curr.Opin.Microbiol.* **1**, 567-571.

Cole, S. T. and Barrell, B. G. (1998). Analysis of the genome of *Mycobacterium tuberculosis* H37Rv. *Novartis.Found.Symp.* **217**, 160-172.

Cole, S. T. (1999). Learning from the genome sequence of *Mycobacterium tuberculosis* H37Rv. *FEBS Lett* **452**, 7-10.

Cole, S. T., Eiglmeier, K., Parkhill, J., James, K. D., Thomson, N. R., Wheeler, P. R., Honore, 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 Barrell, B. G. (2001). Massive gene decay in the leprosy bacillus. *Nature* **409**, 1007-1011.

Connell, N. D. (1994). *Mycobacterium*: isolation, maintenance, transformation, and mutant selection. *Methods Cell Biol.* **45**, 107-125.

Creemers, J. W., Siezen, R. J., Roebroek, A. J., Ayoubi, T. A., Huylebroeck, D., and van de Ven, W. J. (1993). Modulation of furin-mediated proprotein processing activity by site- directed mutagenesis. *J.Biol.Chem.* **268**, 21826-21834.

Crubezy, E., Ludes, B., Poveda, J. D., Clayton, J., Crouau-Roy, B., and Montagnon, D. (1998). Identification of *Mycobacterium* DNA in an Egyptian Pott's disease of 5,400 years old. *C.R.Acad.Sci.III* **321**, 941-951.

Cutler, C. W., Arnold, R. R., and Schenkein, H. A. (1993). Inhibition of C3 and IgG proteolysis enhances phagocytosis of *Porphyromonas gingivalis*. *J.Immunol.* **151**, 7016-7029.

Cywes, C., Hoppe, H. C., Daffe, M., and Ehlers, M. R. (1997). Nonopsonic binding of *Mycobacterium tuberculosis* to complement receptor type 3 is mediated by capsular polysaccharides and is strain dependent. *Infect.Immun.* **65**, 4258-4266.

D

d'Aubenton, C. Y., Brody, E., and Thermes, C. (1990). Prediction of rho-independent *Escherichia coli* transcription terminators. A statistical analysis of their RNA stem-loop structures. *J.Mol.Biol.* **216**, 835-858.

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

Daffe, M. and Etienne, G. (1999). The capsule of *Mycobacterium tuberculosis* and its implications for pathogenicity. *Tuber.Lung Dis.* **79**, 153-169.

Dannenberg, A. M., Jr. and Collins, F. M. (2001). Progressive pulmonary tuberculosis is not due to increasing numbers of viable bacilli in rabbits, mice and guinea pigs, but is due to a continuous host response to mycobacterial products. *Tuberculosis.(Edinb.)* **81**, 229-242.

de Kievit, T. R. and Iglewski, B. H. (2000). Bacterial quorum sensing in pathogenic relationships. *Infect.Immun.* **68**, 4839-4849.

Dillon, D. C., Alderson, M. R., Day, C. H., Bement, T., Campos-Neto, A., Skeiky, Y. A., Vedvick, T., Badaro, R., Reed, S. G., and Houghton, R. (2000). Molecular and immunological characterization of *Mycobacterium tuberculosis* CFP-10, an immunodiagnostic antigen missing in *Mycobacterium bovis* BCG. *J.Clin.Microbiol* **38**, 3285-3290.

Doenhoff, M. J. (1998). Granulomatous inflammation and the transmission of infection: schistosomiasis--and TB too? *Immunol.Today* **19**, 462-467.

Doldi, K., Leroux, M., Augustin, R., Kirchner, H., and Kalden, J. R. (1985). Proliferation and interferon production in whole blood samples and isolated lymphocyte preparations. *J.Interferon Res.* **5**, 55-64.

Doran, T., Tizard, M., Millar, D., Ford, J., Sumar, N., Loughlin, M., and Hermon-Taylor, J. (1997). IS900 targets translation initiation signals in *Mycobacterium avium* subsp. *paratuberculosis* to facilitate expression of its *hed* gene. *Microbiology* **143 (Pt 2)**, 547-552.

Doran, T. J., Hodgson, A. L., Davies, J. K., and Radford, A. J. (1992). Characterisation of a novel repetitive DNA sequence from *Mycobacterium bovis. FEMS Microbiol Lett* **75**, 179-185.

Doran, T. J., Davies, J. K., Radford, A. J., and Hodgson, A. L. (1994). Putative functional domain within ORF2 on the *Mycobacterium* insertion sequences IS900 and IS902. *Immunol.Cell Biol.* **72**, 427-434.

Downing, K. J., McAdam, R. A., and Mizrahi, V. (1999). *Staphylococcus aureus* nuclease is a useful secretion reporter for mycobacteria. *Gene* **239**, 293-299.

Doyle, R. J., Chaloupka, J., and Vinter, V. (1988). Turnover of cell walls in microorganisms. *Microbiol Rev.* **52**, 554-567.

Dubos, R. and Dubos, J. (1952). 'The White Plague: Tuberculosis, Man, and Society.' (Little, Brown, and Company: Boston.)

Dunny, G. M. and Leonard, B. A. (1997). Cell-cell communication in gram-positive bacteria. *Annu.Rev.Microbiol* **51**, 527-564.

E

Eder, J., Rheinnecker, M., and Fersht, A. R. (1993). Folding of subtilisin BPN': role of the pro-sequence. *J.Mol.Biol.* **233**, 293-304.

Eder, J. and Fersht, A. R. (1995). Pro-sequence-assisted protein folding. *Mol.Microbiol.* **16**, 609-614.

Ehlers, M. R. (1993). The wolf at the door. Some thoughts on the biochemistry of the tubercle bacillus. *S.Afr.Med.J.* **83**, 900-903.

Elhay, M. J., Oettinger, T., and Andersen, P. (1998). Delayed-type hypersensitivity responses to ESAT-6 and MPT64 from *Mycobacterium tuberculosis* in the guinea pig. *Infect.Immun.* **66**, 3454-3456.

Espitia, C., Laclette, J. P., Mondragon-Palomino, M., Amador, A., Campuzano, J., Martens, A., Singh, M., Cicero, R., Zhang, Y., and Moreno, C. (1999). The PE-PGRS glycine-rich proteins of *Mycobacterium tuberculosis*: a new family of fibronectin-binding proteins? *Microbiology* **145 (Pt 12)**, 3487-3495.

Eun, H. M., Yapo, A., and Petit, J. F. (1978). DD-Carboxypeptidase activity of membrane fragments of *Mycobacterium smegmatis*. Enzymatic properties and sensitivity to beta-lactam antibiotics. *Eur.J.Biochem.* **86**, 97-103.

F

Fan, N., Cutting, S., and Losick, R. (1992). Characterization of the *Bacillus subtilis* sporulation gene *spoVK*. *J.Bacteriol.* **174**, 1053-1054.

Felsenstein, J. (1989). PHYLIP -- Phylogeny Inference Package (Version 3.2). *Cladistics* **5**, 164-166.

Fenton, M. J. and Vermeulen, M. W. (1996). Immunopathology of tuberculosis: roles of macrophages and monocytes. *Infect.Immun.* **64**, 683-690.

Fine, P. E. (1995). Bacille Calmette-Guerin vaccines: a rough guide. *Clin.Infect.Dis.* **20**, 11-14.

Flynn, J. L. and Chan, J. (2001). Immunology of tuberculosis. *Annu.Rev.Immunol.* **19**, 93-129.

Frangioni, J. V. and Neel, B. G. (1993). Solubilization and purification of enzymatically active glutathione S- transferase (pGEX) fusion proteins. *Anal.Biochem.* **210**, 179-187.

Frederick, G. D., Rombouts, P., and Buxton, F. P. (1993). Cloning and characterisation of *pepC*, a gene encoding a serine protease from *Aspergillus niger. Gene* **125**, 57-64.

G

Garbe, T. R., Barathi, J., Barnini, S., Zhang, Y., Abou-Zeid, C., Tang, D., Mukherjee, R., and Young, D. B. (1994). Transformation of mycobacterial species using hygromycin resistance as selectable marker. *Microbiology* **140 (Pt 1)**, 133-138.

Gelber, R. H. (1994). Chemotherapy of lepromatous leprosy: recent developments and prospects for the future. *Eur.J.Clin.Microbiol.Infect.Dis.* **13**, 942-952.

Gey van Pittius, N. C., Gamieldien, J., Hide, W., Brown, G. D., Siezen, R. J., and Beyers, A. D. (2001). The ESAT-6 gene cluster of *Mycobacterium tuberculosis* and other high G+C Gram-positive bacteria. *Genome Biol.* **2**, 0044.

Goguen, J. D., Hoe, N. P., and Subrahmanyam, Y. V. (1995). Proteases and bacterial virulence: a view from the trenches. *Infect.Agents Dis.* **4**, 47-54.

Gomez, M., Johnson, S., and Gennaro, M. L. (2000). Identification of secreted proteins of *Mycobacterium tuberculosis* by a bioinformatic approach. *Infect.Immun.* **68**, 2323-2327.

Gordon, S. V., Brosch, R., Billault, A., Garnier, T., Eiglmeier, K., and Cole, S. T. (1999a). Identification of variable regions in the genomes of tubercle bacilli using bacterial artificial chromosome arrays. *Mol.Microbiol.* **32**, 643-655.

Gordon, S. V., Eiglmeier, K., Brosch, R., Garnier, T., Honore, N., Barrell, B., and Cole, S. T. (1999b). Genomics of *Mycobacterium tuberculosis* and *Mycobacterium leprae.* In 'Mycobacteria: molecular biology and virulence'. (Eds. C. Ratledge and J. Dale.) pp. 93-109. (Blackwell Science Ltd: Oxford.)

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

Goren, M. B., D'Arcy, H. P., Young, M. R., and Armstrong, J. A. (1976). Prevention of phagosome-lysosome fusion in cultured macrophages by sulfatides of *Mycobacterium tuberculosis*. *Proc.Natl.Acad.Sci.U.S.A* **73**, 2510-2514.

Gormley, E., Sandall, L., Hong, C., Lawton, D., and Murray, A. (1997). Identification and differentiation of mycobacteria using the *PAN* promoter sequence from *Mycobacterium paratuberculosis* as a DNA probe. *FEMS Microbiol.Lett.* **147**, 63-68.

Gron, H. and Breddam, K. (1992). Interdependency of the binding subsites in subtilisin. *Biochemistry* **31**, 8967-8971.

Guerin, C. (1928). Prophylaxe gegen Tuberculose Infectionen bei Rindern mittels BCG. *Wien.Klin.Wschr.* **21**, 731-735.

Guerin, C. (1948). Le BCG et la prevention de la tuberculose. *Rev.Atomes* **27**, 183-188.

Guerin, C. (1980). The history of BCG. In 'BCG vaccine: Tuberculosis-Cancer'. (Ed. S. R. Rosenthal.) pp. 35-43. (PSG Publishing Company: Littleton.)

H

Haas, D. W. and Des Prez, R. M. (1994). Tuberculosis and acquired immunodeficiency syndrome: a historical perspective on recent developments. *Am.J.Med.* **96**, 439-450.

Hansen, G. H. A. (1874). Undersogelser angaende spedalskhedens aasager. *Norsk Magazin for Laegervidenskaben* **4 (Suppl.)**, 1-88.

Hansen, G. H. A. (1880). *Bacillus leprae*. *Virchows Archiv* **79**, 32-42.

Harasym, M., Zhang, L. H., Chater, K., and Piret, J. (1990). The *Streptomyces coelicolor* A3(2) *bldB* region contains at least two genes involved in morphological development. *J.Gen.Microbiol* **136 (Pt 8)**, 1543-1550.

Harboe, M., Oettinger, T., Wiker, H. G., Rosenkrands, I., and Andersen, P. (1996). Evidence for occurrence of the ESAT-6 protein in *Mycobacterium tuberculosis* and virulent *Mycobacterium bovis* and for its absence in *Mycobacterium bovis* BCG. *Infect.Immun.* **64**, 16-22.

Harth, G. and Horwitz, M. A. (1997). Expression and efficient export of enzymatically active *Mycobacterium tuberculosis* glutamine synthetase in *Mycobacterium smegmatis* and evidence that the information for export is contained within the protein. *J.Biol.Chem.* **272,** 22728-22735.

Havlir, D. V. (1994). *Mycobacterium avium* complex: advances in therapy. *Eur.J.Clin.Microbiol.Infect.Dis.* **13**, 915-924.

Hermans, P. W., van Soolingen, D., and van Embden, J. D. (1992). Characterization of a major polymorphic tandem repeat in *Mycobacterium tuberculosis* and its potential use in the epidemiology of *Mycobacterium kansasii* and *Mycobacterium gordonae*. *J.Bacteriol.* **174**, 4157-4165.

Herrmann, J. L., O'Gaora, P., Gallagher, A., Thole, J. E., and Young, D. B. (1996). Bacterial glycoproteins: a link between glycosylation and proteolytic cleavage of a 19 kDa antigen from *Mycobacterium tuberculosis*. *EMBO J.* **15**, 3547-3554.

Higgins, D. G. and Sharp, P. M. (1988). CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* **73**, 237-244.

Hondalus, M. K., Bardarov, S., Russell, R., Chan, J., Jacobs, W. R., Jr., and Bloom, B. R. (2000). Attenuation of and protection induced by a leucine auxotroph of *Mycobacterium tuberculosis*. *Infect.Immun.* **68**, 2888-2898.

Hopewell, P. C. (1994). The baby and the bath water. The case for retaining categorical services for tuberculosis control in a reformed health care system. *Am.J.Respir.Crit Care Med.* **150**, 895.

Horwitz, M. A., Lee, B. W., Dillon, B. J., and Harth, G. (1995). Protective immunity against tuberculosis induced by vaccination with major extracellular proteins of *Mycobacterium tuberculosis*. *Proc.Natl.Acad.Sci.U.S.A* **92**, 1530-1534.

Hu, Z., Haghjoo, K., and Jordan, F. (1996). Further evidence for the structure of the subtilisin propeptide and for its interactions with mature subtilisin. *J.Biol.Chem.* **271**, 3375-3384.

Hubbard, R. D., Flory, C. M., and Collins, F. M. (1992). Immunization of mice with mycobacterial culture filtrate proteins. *Clin.Exp.Immunol.* **87**, 94-98.

Huebner, R. E., Schein, M. F., and Bass, J. B., Jr. (1993). The tuberculin skin test. *Clin.Infect.Dis.* **17**, 968-975.

I

Iadarola, P., Lungarella, G., Martorana, P. A., Viglio, S., Guglielminetti, M., Korzus, E., Gorrini, M., Cavarra, E., Rossi, A., Travis, J., and Luisetti, M. (1998). Lung injury and degradation of extracellular matrix components by *Aspergillus fumigatus* serine proteinase. *Exp.Lung Res.* **24**, 233-251.

Ichihara, S., Suzuki, T., Suzuki, M., and Mizushima, S. (1986). Molecular cloning and sequencing of the *sppA* gene and characterization of the encoded protease IV, a signal peptide peptidase, of *Escherichia coli*. *J.Biol.Chem.* **261**, 9405-9411.

Ikemura, H., Takagi, H., and Inouye, M. (1987). Requirement of pro-sequence for the production of active subtilisin E in *Escherichia coli*. *J.Biol.Chem.* **262**, 7859-7864.

Imboden, P. and Schoolnik, G. K. (1998). Construction and characterization of a partial *Mycobacterium tuberculosis* cDNA library of genes expressed at reduced oxygen tension. *Gene* **213**, 107-117.

J

Jacob-Dubuisson, F., Locht, C., and Antoine, R. (2001). Two-partner secretion in Gram-negative bacteria: a thrifty, specific pathway for large virulence proteins. *Mol.Microbiol* **40**, 306-313.

Jacobs, W. R., Jr., Kalpana, G. V., Cirillo, J. D., Pascopella, L., Snapper, S. B., Udani, R. A., Jones, W., Barletta, R. G., and Bloom, B. R. (1991). Genetic systems for mycobacteria. *Methods Enzymol.* **204**, 537-555.

Johnson, K., Charles, I., Dougan, G., Pickard, D., O'Gaora, P., Costa, G., Ali, T., Miller, I., and Hormaeche, C. (1991). The role of a stress-response protein in *Salmonella typhimurium* virulence. *Mol.Microbiol* **5**, 401-407.

K

Kamath, A. T., Feng, C. G., Macdonald, M., Briscoe, H., and Britton, W. J. (1999). Differential protective efficacy of DNA vaccines expressing secreted proteins of *Mycobacterium tuberculosis* . *Infect.Immun.* **67**, 1702-1707.

Kanatani, A., Yoshimoto, T., Nagai, H., Ito, K., and Tsuru, D. (1992). Location of the protease II gene (*ptrB*) on the physical map of the *Escherichia coli* chromosome. *J.Bacteriol.* **174**, 7881.

Kannan, K. B., Katoch, V. M., Sharma, V. D., and Bharadwaj, V. P. (1987). Extracellular enzymes of mycobacteria. *FEMS Microbiol Lett* **48**, 31-33.

Katial, R. K., Hershey, J., Purohit-Seth, T., Belisle, J. T., Brennan, P. J., Spencer, J. S., and Engler, R. J. (2001). Cell-mediated immune response to tuberculosis antigens: comparison of skin testing and measurement of in vitro gamma interferon production in whole-blood culture. *Clin.Diagn.Lab Immunol.* **8**, 339-345.

Kato-Maeda, M., Rhee, J. T., Gingeras, T. R., Salamon, H., Drenkow, J., Smittipat, N., and Small, P. M. (2001). Comparing genomes within the species *Mycobacterium tuberculosis*. *Genome Res.* **11**, 547-554.

Kenney, T. J. and Churchward, G. (1996). Genetic analysis of the *Mycobacterium smegmatis* rpsL promoter. *J.Bacteriol.* **178**, 3564-3571.

Khan, A. R. and James, M. N. (1998). Molecular mechanisms for the conversion of zymogens to active proteolytic enzymes. *Protein Sci.* **7**, 815-836.

Kilian, M., Mestecky, J., and Russell, M. W. (1988). Defense mechanisms involving Fc-dependent functions of immunoglobulin A and their subversion by bacterial immunoglobulin A proteases. *Microbiol.Rev.* **52**, 296-303.

Kim, K. I., Park, S. C., Kang, S. H., Cheong, G. W., and Chung, C. H. (1999). Selective degradation of unfolded proteins by the self- compartmentalizing HtrA protease, a periplasmic heat shock protein in *Escherichia coli*. *J.Mol.Biol.* **294**, 1363-1374.

Kirchner, H., Kleinicke, C., and Digel, W. (1982). A whole-blood technique for testing production of human interferon by leukocytes. *J.Immunol.Methods* **48**, 213-219.

Knipfer, N. and Shrader, T. E. (1997). Inactivation of the 20S proteasome in *Mycobacterium smegmatis*. *Mol.Microbiol.* **25**, 375-383.

Koch, R. (1882). Die Aetiologie der Tuberkulose. *Ber.Klin.Wochenschrift* **15**.

Kochi, A. (1994). Tuberculosis: distribution, risk factors, mortality. *Immunobiology* **191**, 325-336.

Kok, J. and De Vos, W. M. (1994). The proteolytic sysytem of lactic acid bacteria. In 'Genetics and Biotechnology of Lactic Acid Bacteria'. (Eds. M. J. Gasson and W. M. De Vos.) pp. 169-210. (Blackie Academic and Professional: Glasgow.)

Kortt, A. A., Caldwell, J. B., Lilley, G. G., Edwards, R., Vaughan, J., and Stewart, D. J. (1994). Characterization of a basic serine proteinase (pl approximately 9.5) secreted by virulent strains of *Dichelobacter nodosus* and identification of a distinct, but closely related, proteinase secreted by benign strains. *Biochem.J.* **299 (Pt 2)**, 521-525.

L

Lantz, M. S. (1997). Are bacterial proteases important virulence factors? *J.Periodontal Res.* **32**, 126-132.

Lawrence, J. G. and Roth, J. R. (1996). Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics* **143**, 1843-1860.

Lein, A. D., von Reyn, C. F., Ravn, P., Horsburgh, C. R., Jr., Alexander, L. N., and Andersen, P. (1999). Cellular immune responses to ESAT-6 discriminate between patients with pulmonary disease due to *Mycobacterium avium* complex and those with pulmonary disease due to *Mycobacterium tuberculosis. Clin.Diagn.Lab Immunol.* **6**, 606-609.

Li, Z., Howard, A., Kelley, C., Delogu, G., Collins, F., and Morris, S. (1999). Immunogenicity of DNA vaccines expressing tuberculosis proteins fused to tissue plasminogen activator signal sequences. *Infect.Immun.* **67**, 4780-4786.

Linton, K. J. and Higgins, C. F. (1998). The *Escherichia coli* ATP-binding cassette (ABC) proteins. *Mol.Microbiol* **28**, 5-13.

Lipsitch, M. and Levin, B. R. (1998). Population dynamics of tuberculosis treatment: mathematical models of the roles of non-compliance and bacterial heterogeneity in the evolution of drug resistance. *Int.J.Tuberc.Lung Dis.* **2**, 187-199.

Lu, W., Apostol, I., Qasim, M. A., Warne, N., Wynn, R., Zhang, W. L., Anderson, S., Chiang, Y. W., Ogin, E., Rothberg, I., Ryan, K., and Laskowski, M., Jr. (1997). Binding of amino acid side-chains to S1 cavities of serine proteinases. *J.Mol.Biol.* **266**, 441-461.

M

Maeda, H. and Molla, A. (1989). Pathogenic potentials of bacterial proteases. *Clin.Chim.Acta* **185**, 357-367.

Maeda, H. and Yamamoto, T. (1996). Pathogenic mechanisms induced by microbial proteases in microbial infections. *Biol.Chem.Hoppe Seyler* **377**, 217-226.

Maeda, H. (1996). Role of microbial proteases in pathogenesis. *Microbiol.Immunol.* **40** , 685-699.

Mahairas, G. G., Sabo, P. J., Hickey, M. J., Singh, D. C., and Stover, C. K. (1996). Molecular analysis of genetic differences between *Mycobacterium bovis* BCG and virulent *M. bovis. J.Bacteriol.* **178**, 1274-1282.

Makino, S., Makino, T., Abe, K., Hashimoto, J., Tatsuta, T., Kitagawa, M., Mori, H., Ogura, T., Fujii, T., Fushinobu, S., Wakagi, T., Matsuzawa, H., and Makinoa, T. (1999). Second transmembrane segment of FtsH plays a role in its proteolytic activity and homo-oligomerization. *FEBS Lett.* **460**, 554-558.

Manca, C., Tsenova, L., Barry, C. E., III, Bergtold, A., Freeman, S., Haslett, P. A., Musser, J. M., Freedman, V. H., and Kaplan, G. (1999). *Mycobacterium tuberculosis* CDC1551 induces a more vigorous host response in vivo and in vitro, but is not more virulent than other clinical isolates. *J.Immunol.* **162**, 6740-6746.

Mattow, J., Jungblut, P. R., Schaible, U. E., Mollenkopf, H. J., Lamer, S., Zimny-Arndt, U., Hagens, K., Muller, E. C., and Kaufmann, S. H. (2001). Identification of proteins from *Mycobacterium tuberculosis* missing in attenuated *Mycobacterium bovis* BCG strains. *Electrophoresis* **22**, 2936-2946.

Mazzaccaro, R. J., Gedde, M., Jensen, E. R., van Santen, H. M., Ploegh, H. L., Rock, K. L., and Bloom, B. R. (1996). Major histocompatibility class I presentation of soluble antigen facilitated by *Mycobacterium tuberculosis* infection. *Proc.Natl.Acad.Sci.U.S.A* **93**, 11786-11791.

McAuliffe, O., Ross, R. P., and Hill, C. (2000). Lantibiotics: structure, biosynthesis and mode of action. *FEMS Microbiol Rev.* **714**, 1-24.

McKinney, J. D., Jacobs, W. R., Jr., and Bloom, B. R. (1998). Persisting problems in tuberculosis. In 'Emerging Infections'. pp. 51-146. (Academic Press)

Mdluli, K. E., Treit, J. D., Kerr, V. J., and Nano, F. E. (1995). New vectors for the *in vitro* generation of alkaline phosphatase fusions to proteins encoded by G+C-rich DNA. *Gene* **155**, 133-134.

Misra, N., Habib, S., Ranjan, A., Hasnain, S. E., and Nath, I. (1996). Expression and functional characterisation of the *clpC* gene of *Mycobacterium leprae*: ClpC protein elicits human antibody response. *Gene* **172**, 99-104.

Miyagawa, S., Nishino, N., Kamata, R., Okamura, R., and Maeda, H. (1991). Effects of protease inhibitors on growth of *Serratia marcescens* and *Pseudomonas aeruginosa* . *Microb.Pathog.* **11**, 137-141.

Molla, A., Yamamoto, T., Akaike, T., Miyoshi, S., and Maeda, H. (1989). Activation of hageman factor and prekallikrein and generation of kinin by various microbial proteinases. *J.Biol.Chem.* **264**, 10589-10594.

Montville, T. J. and Chen, Y. (1998). Mechanistic action of pediocin and nisin: recent progress and unresolved questions. *Appl.Microbiol.Biotechnol.* **50**, 511-519.

Mori, H. and Ito, K. (2001). The Sec protein-translocation pathway. *Trends Microbiol* **9**, 494-500.

Morris, S., Kelley, C., Howard, A., Li, Z., and Collins, F. (2000). The immunogenicity of single and combination DNA vaccines against tuberculosis. *Vaccine* **18**, 2155-2163.

Mulder, M. A., Zappe, H., and Steyn, L. M. (1997). Mycobacterial promoters. *Tuber.Lung Dis.* **78**, 211-223.

Murray, A., Winter, N., Lagranderie, M., Hill, D. F., Rauzier, J., Timm, J., Leclerc, C., Moriarty, K. M., Gheorghiu, M., and Gicquel, B. (1992). Expression of *Escherichia coli* beta-galactosidase in *Mycobacterium bovis* BCG using an expression system isolated from *Mycobacterium paratuberculosis* which induced humoral and cellular immune responses. *Mol.Microbiol.* **6**, 3331-3342.

Murray, C. J. and Salomon, J. A. (1998). Modeling the impact of global tuberculosis control strategies. *Proc.Natl.Acad.Sci.U.S.A* **95**, 13881-13886.

Mustafa, A. S., Oftung, F., Amoudy, H. A., Madi, N. M., Abal, A. T., Shaban, F., Rosenkrands, I, and Andersen, P. (2000). Multiple epitopes from the *Mycobacterium tuberculosis* ESAT-6 antigen are recognized by antigen-specific human T cell lines. *Clin.Infect.Dis.* **30 Suppl 3**, S201-S205.

Mustafa, A. S. (2001). Biotechnology in the development of new vaccines and diagnostic reagents against tuberculosis. *Curr.Pharm.Biotechnol.* **2**, 157-173.

N

Nair, E. R., Banerjee, S., Kumar, S., Reddy, M. V., and Harinath, B. C. (2001). Purification and characterization of a 31 kDa mycobacterial excretory- secretory antigenic protein with a diagnostic potential in pulmonary tuberculosis. *Indian J.Chest Dis.Allied Sci.* **43**, 81-90.

Neurath, H. (1989). Proteolytic processing and physiological regulation. *Trends Biochem.Sci.* **14**, 268-271.

Neuwald, A. F., Aravind, L., Spouge, J. L., and Koonin, E. V. (1999). AAA+: A class of chaperone-like ATPases associated with the assembly, operation, and disassembly of protein complexes. *Genome Res.* **9**, 27-43.

Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng* **10**, 1-6.

Nielsen, H. and Krogh, A. (1998). Prediction of signal peptides and signal anchors by a hidden Markov model. *Ismb.* **6**, 122-130.

Noble, J. A., Innis, M. A., Koonin, E. V., Rudd, K. E., Banuett, F., and Herskowitz, I. (1993). The *Escherichia coli hflA* locus encodes a putative GTP-binding protein and two membrane proteins, one of which contains a protease-like domain. *Proc.Natl.Acad.Sci.U.S.A* **90**, 10866-10870.

Nodwell, J. R., McGovern, K., and Losick, R. (1996). An oligopeptide permease responsible for the import of an extracellular signal governing aerial mycelium formation in *Streptomyces coelicolor. Mol.Microbiol.* **22**, 881-893.

Nodwell, J. R. and Losick, R. (1998). Purification of an extracellular signaling molecule involved in production of aerial mycelium by *Streptomyces coelicolor. J.Bacteriol.* **180**, 1334-1337.

North, R. J., Ryan, L., LaCource, R., Mogues, T., and Goodrich, M. E. (1999). Growth rate of mycobacteria in mice as an unreliable indicator of mycobacterial virulence. *Infect.Immun.* **67**, 5483-5485.

O

Oettinger, T., Jorgensen, M., Ladefoged, A., Haslov, K., and Andersen, P. (1999). Development of the *Mycobacterium bovis* BCG vaccine: review of the historical and biochemical evidence for a genealogical tree. *Tuber.Lung Dis.* **79**, 243-250.

Onwubalili, J. K., Scott, G. M., and Robinson, J. A. (1985). Deficient immune interferon production in tuberculosis. *Clin.Exp.Immunol.* **59**, 405-413.

Orme, I. M. (1988a). Characteristics and specificity of acquired immunologic memory to *Mycobacterium tuberculosis* infection. *J.Immunol.* **140**, 3589-3593.

Orme, I. M. (1988b). Induction of nonspecific acquired resistance and delayed-type hypersensitivity, but not specific acquired resistance in mice inoculated with killed mycobacterial vaccines. *Infect.Immun.* **56**, 3310-3312.

Orme, I. M., Andersen, P., and Boom, W. H. (1993). T cell response to *Mycobacterium tuberculosis. J.Infect.Dis.* **167**, 1481-1497.

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.U.S.A* **96**, 2896-2901.

P

Page, R. D. (1996). TreeView: an application to display phylogenetic trees on personal computers. *Comput.Appl.Biosci.* **12**, 357-358.

Pal, P. G. and Horwitz, M. A. (1992). Immunization with extracellular proteins of *Mycobacterium tuberculosis* induces cell-mediated immune responses and substantial protective immunity in a guinea pig model of pulmonary tuberculosis. *Infect.Immun.* **60**, 4781-4792.

Pardee, A. B., Jacob, F., and Monod, J. (1959). The genetic control and cytoplasmic expression of "inducibility" in the synthesis of β-galactosidase by *E. coli. J.Mol.Biol.* **1**, 165-178.

Parish, T. and Stoker, N. G. (1999). Mycobacteria: bugs and bugbears (two steps forward and one step back). *Mol.Biotechnol.* **13**, 191-200.

Philipp, W. J., Nair, S., Guglielmi, G., Lagranderie, M., Gicquel, B., and Cole, S. T. (1996). Physical mapping of *Mycobacterium bovis* BCG Pasteur reveals differences from the genome map of *Mycobacterium tuberculosis* H37Rv and from *M. bovis*. *Microbiology* **142 (Pt 11)**, 3135-3145.

Picken, R. N., Tsang, A. Y., and Yang, H. L. (1988). Speciation of organisms within the *Mycobacterium avium-Mycobacterium intracellulare-Mycobacterium scrofulaceum* (MAIS) complex based on restriction fragment length polymorphisms. *Mol.Cell Probes* **2**, 289-304.

Pitulle, C., Dorsch, M., Kazda, J., Wolters, J., and Stackebrandt, E. (1992). Phylogeny of rapidly growing members of the genus *Mycobacterium*. *Int.J.Syst.Bacteriol.* **42**, 337-343.

Plano, G. V., Day, J. B., and Ferracci, F. (2001). Type III export: new uses for an old pathway. *Mol.Microbiol* **40**, 284-293.

Ponnighaus, J. M., Fine, P. E., Sterne, J. A., Wilson, R. J., Msosa, E., Gruer, P. J., Jenkins, P. A., Lucas, S. B., Liomba, N. G., and Bliss, L. (1992). Efficacy of BCG vaccine against leprosy and tuberculosis in northern Malawi. *Lancet* **339**, 636-639.

Pope, M. K., Green, B., and Westpheling, J. (1998). The *bldB* gene encodes a small protein required for morphogenesis, antibiotic production, and catabolite control in *Streptomyces coelicolor*. *J.Bacteriol.* **180**, 1556-1562.

Poquet, I., Saint, V., Seznec, E., Simoes, N., Bolotin, A., and Gruss, A. (2000). HtrA is the unique surface housekeeping protease in *Lactococcus lactis* and is required for natural protein processing. *Mol.Microbiol.* **35** , 1042-1051.

Poulet, S. and Cole, S. T. (1995). Characterization of the highly abundant polymorphic GC-rich-repetitive sequence (PGRS) present in *Mycobacterium tuberculosis*. *Arch.Microbiol.* **163**, 87-95.

Power, S. D., Adams, R. M., and Wells, J. A. (1986). Secretion and autoproteolytic maturation of subtilisin. *Proc.Natl.Acad.Sci.U.S.A* **83**, 3096-3100.

Powers, J. C., Odake, S., Oleksyszyn, J., Hori, H., Ueda, T., Boduszek, B., and Kam, C. (1993). Proteases--structures, mechanism and inhibitors. *Agents Actions Suppl* **42**, 3-18.

Pugsley, A. P. and Schwartz, M. (1985). Export and secretion of proteins by bacteria. *FEMS Microbiol Rev.* **32**, 3-38.

Pugsley, A. P. (1993). The complete general secretory pathway in gram-negative bacteria. *Microbiol Rev.* **57**, 50-108.

R

Ramakrishnan, L., Federspiel, N. A., and Falkow, S. (2000). Granuloma-specific expression of *Mycobacterium* virulence proteins from the glycine-rich PE-PGRS family. *Science* **288**, 1436-1439.

Ramos, L. S. (1994). Characterization of mycobacteria species by HPLC and pattern recognition. *J.Chromatogr.Sci.* **32**, 219-227.

Rauzier, J., Gormley, E., Gutierrez, M. C., Kassa-Kelembho, E., Sandall, L. J., Dupont, C., Gicquel, B., and Murray, A. (1999). A novel polymorphic genetic locus in members of the *Mycobacterium tuberculosis* complex. *Microbiology* **145 (Pt 7)**, 1695-1701.

Ravn, P., Demissie, A., Eguale, T., Wondwosson, H., Lein, D., Amoudy, H. A., Mustafa, A. S., Jensen, A. K., Holm, A., Rosenkrands, I., Oftung, F., Olobo, J., von Reyn, F., and Andersen, P. (1999). Human T cell responses to the ESAT-6 antigen from *Mycobacterium tuberculosis*. *J.Infect.Dis.* **179**, 637-645.

Rawlings, N. D. and Barrett, A. J. (1993). Evolutionary families of peptidases. *Biochem.J.* **290 (Pt 1)**, 205-218.

Rawlings, N. D. and Barrett, A. J. (1994). Families of serine peptidases. *Methods Enzymol.* **244**, 19-61.

Raynaud, C., Etienne, G., Peyron, P., Laneelle, M. A., and Daffe, M. (1998). Extracellular enzyme activities potentially involved in the pathogenicity of *Mycobacterium tuberculosis*. *Microbiology* **144 (Pt 2)**, 577-587.

Reich, M., Wright, G. L., Jr., and Affronti, L. F. (1981). Proteolysis and recycling of *Mycobacterium tuberculosis* culture filtrate tuberculoproteins. *Microbios* **32**, 173-179.

Reyrat, J.-M. and Kahn, D. (2001). *Mycobacterium smegmatis*: an absurd model for tuberculosis? *Trends Microbiol.* **9**, 472-473.

Riley, R. I., Mills, C. C., Nyka, W., Weinstock, N., Storey, P. B., Sultan, L. U., Riley, M. C., and Wells, W. F. (1959). Aerial dissemination of pulmonary tuberculosis: A two-year study of contagion in a tuberculosis ward. *Am.J.Hygiene* **70**, 185-196.

Rindi, L., Lari, N., and Garzelli, C. (1999). Search for genes potentially involved in *Mycobacterium tuberculosis* virulence by mRNA differential display. *Biochem.Biophys.Res.Commun.* **258**, 94-101.

Roberts, A. D., Sonnenberg, M. G., Ordway, D. J., Furney, S. K., Brennan, P. J., Belisle, J. T., and Orme, I. M. (1995). Characteristics of protective immunity engendered by vaccination of mice with purified culture filtrate protein antigens of *Mycobacterium tuberculosis*. *Immunology* **85**, 502-508.

Roche, P. W., Triccas, J. A., and Winter, N. (1995). BCG vaccination against tuberculosis: past disappointments and future hopes. *Trends Microbiol.* **3**, 397-401.

Rodriguez, G. M., Gold, B., Gomez, M., Dussurget, O., and Smith, I. (1999). Identification and characterization of two divergently transcribed iron regulated genes in *Mycobacterium tuberculosis*. *Tuber.Lung Dis.* **79**, 287-298.

Romain, F., Augier, J., Pescher, P., and Marchal, G. (1993). Isolation of a proline-rich mycobacterial protein eliciting delayed- type hypersensitivity reactions only in guinea pigs immunized with living mycobacteria. *Proc.Natl.Acad.Sci.U.S.A* **90**, 5322-5326.

Rosenkrands, I., Weldingh, K., Jacobsen, S., Hansen, C., Florio, V., Gianetri, I., and Andersen, P. (2000a). Mapping and identification of *Mycobacterium tuberculosis* proteins by two-dimensional gel electrophoresis, microsequencing and immunodetection. *Electrophoresis* **21**, 935-948.

Rosenkrands, I., King, A., Weldingh, K., Moniatte, M., Moertz, E., and Andersen, P. (2000b). Towards the proteome of *Mycobacterium tuberculosis*. *Electrophoresis* **21**, 3740-3756.

Ross, B. C., Raios, K., Jackson, K., and Dwyer, B. (1992). Molecular cloning of a highly repeated DNA element from *Mycobacterium tuberculosis* and its use as an epidemiological tool. *J.Clin.Microbiol.* **30**, 942-946.

Rothel, J. S., Jones, S. L., Corner, L. A., Cox, J. C., and Wood, P. R. (1990). A sandwich enzyme immunoassay for bovine interferon-gamma and its use for the detection of tuberculosis in cattle. *Aust.Vet.J.* **67**, 134-137.

Rothel, J. S., Jones, S. L., Corner, L. A., Cox, J. C., and Wood, P. R. (1992). The gamma-interferon assay for diagnosis of bovine tuberculosis in cattle: conditions affecting the production of gamma-interferon in whole blood culture. *Aust.Vet.J.* **69**, 1-4.

Roudiak, S. G. and Shrader, T. E. (1998). Functional role of the N-terminal region of the Lon protease from *Mycobacterium smegmatis*. *Biochemistry* **37**, 11255-11263.

Roudiak, S. G., Seth, A., Knipfer, N., and Shrader, T. E. (1998). The lon protease from *Mycobacterium smegmatis*: molecular cloning, sequence analysis, functional expression, and enzymatic characterization. *Biochemistry* **37**, 377-386.

Rowland, S. S., Ruckert, J. L., and Burall, B. N., Jr. (1997). Identification of an elastolytic protease in stationary phase culture filtrates of *M. tuberculosis*. *FEMS Microbiol.Lett.* **151**, 59-64.

S

Sahl, H. G. and Bierbaum, G. (1998). Lantibiotics: biosynthesis and biological activities of uniquely modified peptides from gram-positive bacteria. *Annu.Rev.Microbiol.* **52**, 41-79.

Salo, W. L., Aufderheide, A. C., Buikstra, J., and Holcomb, T. A. (1994). Identification of *Mycobacterium tuberculosis* DNA in a pre-Columbian Peruvian mummy. *Proc.Natl.Acad.Sci.U.S.A* **91**, 2091-2094.

Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989). 'Molecular Cloning: A Laboratory Manual.' 2 Edn. (Cold Spring Harbour, New York.)

Sandkvist, M. (2001). Biology of type II secretion. *Mol.Microbiol* **40**, 271-283.

Sathish, M., Esser, R. E., Thole, J. E., and Clark-Curtiss, J. E. (1990). Identification and characterization of antigenic determinants of *Mycobacterium leprae* that react with antibodies in sera of leprosy patients. *Infect.Immun.* **58**, 1327-1336.

Schagger, H. and von Jagow, G. (1987). Tricine-sodium dodecyl sulfate-polyacrylamide gel electrophoresis for the separation of proteins in the range from 1 to 100 kDa. *Anal.Biochem.* **166**, 368-379.

Schechter, I. and Berger, A. (1967). On the size of the active site in proteases. I. Papain. *Biochem.Biophys.Res.Commun.* **27**, 157-162.

Schluger, N. W. and Rom, W. N. (1998). The host immune response to tuberculosis. *Am.J.Respir.Crit Care Med.* **157**, 679-691.

Schulzer, M., Fitzgerald, J. M., Enarson, D. A., and Grzybowski, S. (1992). An estimate of the future size of the tuberculosis problem in sub- Saharan Africa resulting from HIV infection. *Tuber.Lung Dis.* **73**, 52-58.

See, Y. P. and Jackowski, G. (1989). Protein molecular weight determination by sodium dodecyl sulfate polyacrylamide gel electrophoresis. In 'Protein Structure - A Practical Approach.'. (Ed. T. E. Creighton.) pp. 1-21. (IRL Press: Oxford.)

Seidah, N. G. and Chretien, M. (1994). Pro-protein convertases of subtilisin/kexin family. *Methods Enzymol.* **244**, 175-188.

Shephard, E. G., Beer, S. M., Anderson, R., Strachan, A. F., Nel, A. E., and de Beer, F. C. (1989). Generation of biologically active C-reactive protein peptides by a neutral protease on the membrane of phorbol myristate acetate- stimulated neutrophils. *J.Immunol.* **143**, 2974-2981.

Shetty, K. T., Antia, N. H., and Krishnaswamy, P. R. (1981). Occurrence of gamma-glutamyl transpeptidase activity in several mycobacteria including *Mycobacterium leprae*. *Int.J.Lepr.Other Mycobact.Dis.* **49**, 49-56.

Shinnick, T. M. and Good, R. C. (1994). Mycobacterial taxonomy. *Eur.J.Clin.Microbiol.Infect.Dis.* **13**, 884-901.

Siezen, R. J., De Vos, W. M., Leunissen, J. A., and Dijkstra, B. W. (1991). Homology modelling and protein engineering strategy of subtilases, the family of subtilisin-like serine proteinases. *Protein Eng* **4**, 719-737.

Siezen, R. J., Creemers, J. W., and van de Ven, W. J. (1994). Homology modelling of the catalytic domain of human furin. A model for the eukaryotic subtilisin-like proprotein convertases. *Eur.J.Biochem.* **222**, 255-266.

Siezen, R. J., Rollema, H. S., Kuipers, O. P., and De Vos, W. M. (1995). Homology modelling of the *Lactococcus lactis* leader peptidase NisP and its interaction with the precursor of the lantibiotic nisin. *Protein Eng* **8**, 117-125.

Siezen, R. J., Kuipers, O. P., and De Vos, W. M. (1996). Comparison of lantibiotic gene clusters and encoded proteins. *Antonie Van Leeuwenhoek* **69**, 171-184.

Siezen, R. J. and Leunissen, J. A. (1997). Subtilases: the superfamily of subtilisin-like serine proteases. *Protein Sci.* **6**, 501-523.

Skeiky, Y. A., Lodes, M. J., Guderian, J. A., Mohamath, R., Bement, T., Alderson, M. R., and Reed, S. G. (1999). Cloning, expression, and immunological evaluation of two putative secreted serine protease antigens of *Mycobacterium tuberculosis*. *Infect.Immun.* **67**, 3998-4007.

Skjøt, R. L., Oettinger, T., Rosenkrands, I., Ravn, P., Brock, I., Jacobsen, S., and Andersen, P. (2000). Comparative evaluation of low-molecular-mass proteins from *Mycobacterium tuberculosis* identifies members of the ESAT-6 family as immunodominant T-cell antigens. *Infect.Immun.* **68**, 214-220.

Skjøt, R. L., Agger, E. M., and Andersen, P. (2001). Antigen discovery and tuberculosis vaccine development in the post- genomic era. *Scand.J.Infect.Dis.* **33**, 643-647.

Snapper, S. B., Melton, R. E., Mustafa, S., Kieser, T., and Jacobs, W. R., Jr. (1990). Isolation and characterization of efficient plasmid transformation mutants of *Mycobacterium smegmatis*. *Mol.Microbiol.* **4**, 1911-1919.

Sokal, J. E. (1975). Editorial: Measurement of delayed skin-test responses. *N.Engl.J.Med.* **293**, 501-502.

Sonnhammer, E. L., von Heijne, G., and Krogh, A. (1998). A hidden Markov model for predicting transmembrane helices in protein sequences. *Ismb.* **6**, 175-182.

Sørensen, A. L., Nagai, S., Houen, G., Andersen, P., and Andersen, A. B. (1995). Purification and characterization of a low-molecular-mass T-cell antigen secreted by *Mycobacterium tuberculosis*. *Infect.Immun.* **63**, 1710-1717.

Sperandio, V., Mellies, J. L., Nguyen, W., Shin, S., and Kaper, J. B. (1999). Quorum sensing controls expression of the type III secretion gene transcription and protein secretion in enterohemorrhagic and enteropathogenic *Escherichia coli*. *Proc.Natl.Acad.Sci.U.S.A* **96**, 15196-15201.

Springer, B., Stockman, L., Teschner, K., Roberts, G. D., and Bottger, E. C. (1996). Two-laboratory collaborative study on identification of mycobacteria: molecular versus phenotypic methods. *J.Clin.Microbiol.* **34**, 296-303.

Stead, W. W. (1997). The origin and erratic global spread of tuberculosis. How the past explains the present and is the key to the future. *Clin.Chest Med.* **18**, 65-77.

Steiner, D. F., Smeekens, S. P., Ohagi, S., and Chan, S. J. (1992). The new enzymology of precursor processing endoproteases. *J.Biol.Chem.* **267**, 23435-23438.

Strohl, W. R. (1992). Compilation and analysis of DNA sequences associated with apparent streptomycete promoters. *Nucleic Acids Res.* **20**, 961-974.

Sturgiil-Koszycki, S., Schlesinger, P. H., Chakraborty, P., Haddix, P. L., Collins, H. L., Fok, A. K., Allen, R. D., Gluck, S. L., Heuser, J., and Russell, D. G. (1994). Lack of acidification in *Mycobacterium* phagosomes produced by exclusion of the vesicular proton-ATPase. *Science* **263**, 678-681.

Suzuki, T., Itoh, A., Ichihara, S., and Mizushima, S. (1987). Characterization of the *sppA* gene coding for protease IV, a signal peptide peptidase of *Escherichia coli*. *J.Bacteriol.* **169**, 2523-2528.

Swofford, D. L. (1998). PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

T

Tatusov, R. L., Galperin, M. Y., Natale, D. A., and Koonin, E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28**, 33-36.

Teitelbaum, R., Cammer, M., Maitland, M. L., Freitag, N. E., Condeelis, J., and Bloom, B. R. (1999). Mycobacterial infection of macrophages results in membrane-permeable phagosomes. *Proc.Natl.Acad.Sci.U.S.A* **96**, 15190-15195.

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*. *Tuber.Lung Dis.* **79**, 329-342.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673-4680.

Timm, J., Lim, E. M., and Gicquel, B. (1994). *Escherichia coli*-mycobacteria shuttle vectors for operon and gene fusions to *lacZ*: the pJEM series. *J.Bacteriol.* **176**, 6749-6753.

Timm, J., Gomez, M., and Smith, I. (1999). Gene expression and regulation. In 'Mycobacteria: molecular biology and virulence'. (Eds. C. Ratledge and J. Dale.) pp. 59-92. (Blackwell Science Ltd: Oxford.)

Travis, J., Potempa, J., and Maeda, H. (1995). Are bacterial proteinases pathogenic factors? *Trends Microbiol.* **3**, 405-407.

Twining, S. S. (1984). Fluorescein isothiocyanate-labeled casein assay for proteolytic enzymes. *Anal.Biochem.* **143**, 30-34.

U, V

Ulrichs, T., Munk, M. E., Mollenkopf, H., Behr-Perst, S., Colangeli, R., Gennaro, M. L., and Kaufmann, S. H. (1998). Differential T cell responses to Mycobacterium tuberculosis ESAT6 in tuberculosis patients and healthy donors. *Eur.J.Immunol.* **28**, 3949-3958.

Ulrichs, T., Anding, P., Porcelli, S., Kaufmann, S. H., and Munk, M. E. (2000). Increased numbers of ESAT-6- and purified protein derivative-specific gamma interferon-producing cells in subclinical and active tuberculosis infection. *Infect.Immun.* **68**, 6073-6076.

Valway, S. E., Sanchez, M. P., Shinnick, T. F., Orme, I., Agerton, T., Hoy, D., Jones, J. S., Westmoreland, H., and Onorato, I. M. (1998). An outbreak involving extensive transmission of a virulent strain of *Mycobacterium tuberculosis*. *N.Engl.J.Med.* **338**, 633-639.

van de Ven, W. J., Voorberg, J., Fontijn, R., Pannekoek, H., van den Ouweland, A. M., van Duijnhoven, H. L., Roebroek, A. J., and Siezen, R. J. (1990). Furin is a subtilisin-like proprotein processing enzyme in higher eukaryotes. *Mol.Biol.Rep.* **14**, 265-275.

van Pinxteren, L. A., Ravn, P., Agger, E. M., Pollock, J., and Andersen, P. (2000). Diagnosis of tuberculosis based on the two specific antigens ESAT-6 and CFP10. *Clin.Diagn.Lab Immunol.* **7**, 155-160.

van der Meer, Jr., Polman, J., Beerthuyzen, M. M., Siezen, R. J., Kuipers, O. P., and De Vos, W. M. (1993). Characterization of the *Lactococcus lactis* nisin A operon genes *nisP*, encoding a subtilisin-like serine protease involved in precursor processing, and *nisR*, encoding a regulatory protein involved in nisin biosynthesis. *J.Bacteriol.* **175**, 2578-2588.

Vekemans, J., Lienhardt, C., Sillah, J. S., Wheeler, J. G., Lahai, G. P., Doherty, M. T., Corrah, T., Andersen, P., McAdam, K. P., and Marchant, A. (2001). Tuberculosis contacts but not patients have higher gamma interferon responses to ESAT-6 than do community controls in The Gambia. *Infect.Immun.* **69**, 6554-6557.

Vissa, V. D. and Brennan, P. J. (2001). The genome of *Mycobacterium leprae*: a minimal mycobacterial gene set. *Genome Biol.* **2**, REVIEWS1023.

Vordermeier, H. M., Cockle, P. J., Whelan, A. O., Rhodes, S., and Hewinson, R. G. (2000). Toward the Development of Diagnostic Assays to Discriminate between *Mycobacterium bovis* Infection and Bacille Calmette-Guerin Vaccination in Cattle. *Clin.Infect.Dis.* **30 Suppl 3**, S291-S298.

Vordermeier, H. M., Whelan, A., Cockle, P. J., Farrant, L., Palmer, N., and Hewinson, R. G. (2001). Use of synthetic peptides derived from the antigens ESAT-6 and CFP-10 for differential diagnosis of bovine tuberculosis in cattle. *Clin.Diagn.Lab Immunol.* **8**, 571-578.

W

Wandersman, C. (1989). Secretion, processing and activation of bacterial extracellular proteases. *Mol.Microbiol.* **3**, 1825-1831.

Wang, L. and Lutkenhaus, J. (1998). FtsK is an essential cell division protein that is localized to the septum and induced as part of the SOS response. *Mol.Microbiol.* **29**, 731-740.

Wang, R. F., Cao, W. W., and Cerniglia, C. E. (1995). Phylogenetic analysis of polycyclic aromatic hydrocarbon degrading mycobacteria by 16S rRNA sequencing. *FEMS Microbiol.Lett.* **130**, 75-80.

Wards, B. J., de Lisle, G. W., and Collins, D. M. (2000). An *esat6* knockout mutant of *Mycobacterium bovis* produced by homologous recombination will contribute to the development of a live tuberculosis vaccine. *Tuber.Lung Dis.* **80**, 185-189.

Wayne, L. G. and Kubica, G. P. (1986). The mycobacteria. In 'Bergey's Manual of Systematic Bacteriology'. (Eds. P. H. A. Sneath, N. S. Mair, M. E. Sharpe, and J. G. Holt.) pp. 1436-57. (Williams and Wilkins: Baltimore.)

Wendei, K. A., Aiwood, K. S., Gachuhi, R., Chaisson, R. E., Bishai, W. R., and Sterling, T. R. (2001). Paradoxical worsening of tuberculosis in HIV-infected persons. *Chest* **120**, 193-197.

WHO (1993). WHO declares tuberculosis a global emergency. *Soz.Praventivmed.* **38**, 251-252.

Wiker, H. G., Wilson, M. A., and Schoolnlk, G. K. (2000). Extracytoplasmic proteins of *Mycobacterium tuberculosis* - mature secreted proteins often start with aspartic acid and proline. *Microbiology* **146 (Pt** 7), 1525-1533.

Willcox, P. A. (2000). Drug-resistant tuberculosis. *Curr.Opin.Pulm.Med.* **6**, 198-202.

Willey, J., Schwedock, J., and Losick, R. (1993). Multiple extracellular signals govern the production of a morphogenetic protein involved in aerial mycelium formation by *Streptomyces coelicolor. Genes Dev.* **7**, 895-903.

Wingrove, J. A., DiScipio, R. G., Chen, Z., Potempa, J., Travis, J., and Hugli, T. E. (1992). Activation of complement components C3 and C5 by a cysteine proteinase (gingipain-1) from *Porphyromonas* (*Bacteroides*) *gingivalis*. *J.Biol.Chem.* **267**, 18902-18907.

Wixon, J. (2000). Genomes 2000 International Conference on Microbial and Model Genomes. *Yeast* **17**, 124-133.

Wolf, D. H. (1992). Proteases as biological regulators. Introductory remarks. *Experientia* **48**, 117-118.

Wong, S. L. and Doi, R. H. (1986). Determination of the signal peptidase cleavage site in the preprosubtilisin of *Bacillus subtilis*. *J.Biol.Chem.* **261**, 10176-10181.

Wu, L. J. and Errington, J. (1997). Septal localization of the SpoIIIE chromosome partitioning protein in *Bacillus subtilis*. *EMBO J.* **16**, 2161-2169.

Y, Z

Yew, W. W. and Chau, C. H. (1995). Drug-resistant tuberculosis in the 1990s. *Eur.Respir.J.* **8**, 1184-1192.

Zhou, A., Paquet, L., and Mains, R. E. (1995). Structural elements that direct specific processing of different mammalian subtilisin-like prohormone convertases. *J.Biol.Chem.* **270**, 21509-21516.

Zubay, G. (1993). Immunobiology. In 'Biochemistry'. (Eds. K. Kane, M. Johnson, and S. Padden.) pp. 963-78. (Wm. C. Brown Publishers: Dubuque.)

Zumarraga, M., Bigi, F., Alito, A., Romano, M. I., and Cataldi, A. (1999). A 12.7 kb fragment of the *Mycobacterium tuberculosis* genome is not present in *Mycobacterium bovis*. *Microbiology* **145** **(Pt 4)**, 893-897.