Studies on GntR family of transcriptional regulators from *Mycobacterium tuberculosis*

Thesis submitted to

Manipal University, Manipal

for the degree of

Doctor of Philosophy

by

Vaibhav Vindal

[Registration number: 040100010]

Centre for DNA Fingerprinting and Diagnostics
Hyderabad-500076
INDIA

DECLARATION

The research work embodied in this thesis entitled, "**Studies on GntR family of transcriptional regulators from *Mycobacterium tuberculosis***", has been carried out by me at the Centre for DNA Fingerprinting and Diagnostics, Hyderabad, under the guidance of Dr. Akash Ranjan. I hereby declare that this work is original and has not been submitted in part or full for any other degree or diploma of any other university or institution.

[Vaibhav Vindal]

CERTIFICATE

This is to certify that this thesis entitled, **“Studies on GntR family of transcriptional regulators from *Mycobacterium tuberculosis*”**, submitted by **Mr. Vaibhav Vindal** for the Degree of Doctor of Philosophy to Manipal University is based on the work carried out by him at the Centre for DNA Fingerprinting and Diagnostics, Hyderabad. This work is original and has not been submitted in part or full for any degree or diploma of any other university or institution.

[Dr. Akash Ranjan]	**[Dr. Shekhar C. Mande]**
Thesis Supervisor	Dean, Academic Affairs

Acknowledgements

The work presented in this thesis was accomplished with the help of many colleagues and friends. I would like to take this opportunity to thank all the people who have helped me directly or indirectly in their various capacities during the tenure of my Ph.D.

First and foremost, I would like to express my heartfelt gratitude to my research supervisor Dr. Akash Ranjan for his constant support, guidance and encouragement for finishing this thesis work. I am indeed fortunate to acquire the knowledge of computational as well as "wet lab" experiments under his outstanding guidance, which otherwise, would never have been made possible. The freedom of thought and amiable environment extended by him was immense, which provided me confidence in analyzing any research problem that I encountered. He is a wonderful teacher and I thank him for teaching me the various aspects of biology.

I am grateful to the former Director, Centre for DNA Fingerprinting and Diagnostics, Dr. Syed E. Hasnain for providing facilities and encouragement to carry out my thesis work in this esteemed institute. I would like to express my thanks to present Director, CDFD, Dr. J Gowrishankar for his continued input to carry out research work. I am thankful to the Vice Chancellor, Registrar and the Head, Dept. of Biotechnology at Manipal University for permitting me to register as a PhD graduate student there.

I gratefully acknowledge UGC and DST for providing fellowship during my PhD and financial support to present a poster at EMBL, Heidelberg Germany, respectively. I also extend my thanks to Dr. Salam for his kind support during my foreign trip.

I am deeply indebted and would like to express my sincere thanks to Dr. Gayatri for her constant encouragement and help. I am immensely grateful to Dr. Murali Bashyam for his kind help and the joyful discussions that we had. I also thank Dr. Sanjeev Khosla for his constructive criticism and excellent advice from time to time. I am grateful to other seniors and staff at CDFD, Dr. Ranjan Sen, Dr. Nagarajaram, Dr. Shekhar Mande, Dr. Sangeetha and Dr. Abhijeet for the constructive conversations that we had and which has immensely inspired me in my research work. I also take this opportunity to thank Nitin Anna for his excellent support during the tenure.

A few sentences are not sufficient to rightly describe the good times that I had with my lab-mates. I thank Sarita for the scientific and humorous debates that we had. I sincerely thank my senior Sailu who inspired me to choose the topic of research during my initial days. I cherish all the friendly and warm conversations with Uma, she has been supportive and helping in nature. I thank past members of my lab Suma, Ashwanth, Madhu, Sidharth Jayashri, Rohini, Indira, for their kind support at different times. I profoundly appreciate the camaraderie I shared with Jamshaid, Vijay, Rohan, Jayavardhan-JVR. Though I interacted quite less with Sanjay, Shashi, Sudeshana, Ravi and Manju-the newly joined students; their names too are now part of my address book.

Many cheerful thanks to my batchmates Aisha, Akif, Anoop, Archana, Arvind, Prabhat, Sudhish and late Showkat for their friendly company, especially at the time of course work, which was a memorable moment. I acknowledge all past and present CDFDians- Gokul, Sridhar, Kshama, Sandeep, Madhav, Santosh, Bibhushita, Pramod, Nancy, Sheeba for their friendly environment and for sharing fun that made my life comfortable during my stay here. Many thanks are due to Uma, Jeffrey, Rohan and Gokul for their fruitful suggestion to improve the thesis.

I am happy to have developed everlasting friendship with Ashwanth and Suma. They were not only colleagues of mine who contributed towards the completion of many of my experiments, but always been there in my need. I have been blessed with many friends who have extended their unconditional support during all the times. I am very grateful to Sudhish-Rachana, Jayendra-Nidhi and Gokul for having created a homely environment, when I am away from home. We all had a wonderful time together. Gokul-Anna and myself had good time, specially while preparing food. I also take the opportunity to thank my CCMB friends: Jobin, Soumya, Geetha, Sahasransu for the time I enjoyed in their company. I really run out of words to say anything about Jobin. Though he is miles away, his love and prayers are always with me.

I would like to acknowledge the big help provided by Sharief, Hassan, Hamid, Gowrish and Imran. I also thank the staff from different sections of CDFD: like Bioinformatics, NGTF, Instrumentation, Administration, Accounts, EMPC, Security and Canteen.

I thank my family for always staying so close to me, despite the distance. Though on many occasions, I was not present to help them, they were always with me at my time of need. At last, I thank almighty for his benevolent grace. I dedicate my thesis to none else than Lord Jesus. I cannot define his mercy towards me who always turns my sorrows into joy.

[Vaibhav Vindal]

Table of contents

Chapter 1

General Introduction

Chapter 2

GntR family transcriptional regulators from *Mycobacterium tuberculosis*

Chapter 3

Identification of GntR operator sites: *In vitro* validation for Rv0586 and Rv0792c

Chapter 4

Operator site recognition: Insight from DNA binding domain

Chapter 5

GntR family of transcriptional regulators from non-infectious *Mycobacterium smegmatis*

Appendix

List of Figures

List of Tables

ABBREVIATIONS

Measurements:

°C	Degree Celsius
µM	micro molar
Å	angstrom
g	gram(s)
h	hour(s)
kDa	kilo dalton
L	litre
M	molar
mg	milli gram
mg/ml	milligram per milliliter
min	minute
ml	milli liter
mM	milli molar
mm	milli meter
mV	milli volt
MW	molecular weight
ng	nano gram
nm	nanometer
O.D.	optical density
rmsd	root mean square deviation
rpm	rotation per minute
s	second(s)
µl	microliter
µM	micromolar

General:

AA	amino acid
APS	ammonium persulfate
ATP	adenosine triphosphate
BCG	bacillus-calmette-guerin
BLAST	basic local alignment search tool
bp	base pair(s)
BSA	bovine serum albumin
C terminal	carboxy terminal
DBD	DNA binding domain
DDW	double distilled water
DNA	deoxyribonucleic acid
dNTP	deoxynucleotide triphosphate
DTT	dithiothreitol
E-b/O	effector binding and/or oligomerization
EDTA	ethylene diamine tetraacetic acid
EMSA	electrophoretic mobility shift assay
EtBr	ethidium bromide
Fam	family
GD	GntR-like domain
IPTG	isopropyl β-D-thiogalacto-pyranoside
Kb	kilo basepairs

kDa	kilo Dalton(s)
LB	Luria Bertanni medium
NCBI	National Center for Biotechnology Information
Ni-NTA	nickel nitrilo-tri-acetate
ORF	open reading frame
PAGE	polyacrylamide gel electrophoresis
PCR	polymerase chain reaction
pdb	protein database
PFAM	protein families
PMSF	phenylmethylsulfonyl flouride
RNA	ribonucleic acid
rpm	revolutions per minute
RT	room temperature
SDS	sodium dodecyl sulfate
SF	sub family
SP	swiss prot
TAE	tris-acetate-EDTA buffer
TB	Terrific broth medium
TBE	tris-borate-EDTA-buffer
TEMED	N,N,N',N' tetramethylethylenediamine
TIGR	The Institute for Genomic Research
Tris	tris(hydroxymethyl)aminomethane
U	units
UV	ultraviolet
WHO	World Health Organization
α	alfa
β	beta

Organism:

mav	*M. avium 104*
map	*M. avium subsp. paratuberculosis*
mbv	*M. bovis*
mtu	*M. tuberculosis*
bcg	*M. bovis BCG Pasteur 1173P2*
mgl	*M. PYR-GCK*
jls	*M. sp. JLS*
kms	*M. sp KMS*
mcs	*M. sp. MCS*
msm	*M. smegmatis*
cdc	*M. tuberculosis CDC1551*
mtf	*M. tuberculosis F11*
mra	*M. tuberculosis H37Ra*
mul	*M. ulcerans Agy99*
van	*M. vanbaalenii PYR*
far	*N. farciana*

Nucleotides: Standard single letter codes
Amino acids: Standard three or single letter codes

PREFACE

Mycobacterium tuberculosis is an incredibly successful pathogen causing tuberculosis in humans. In spite of more than 100 years of research, tuberculosis remains a significant world health burden and one of the major causes of human suffering, with an approximate mortality rate of 2 million people per year. Although many antibiotics are available to combat this disease, new cases of tuberculosis are being reported every year.

The genome-sequencing project of this bacillus, *M. tuberculosis H37Rv*, was completed in 1998. Since then it has been possible for biologists to explore the coding and non-coding sequences to understand the role of uncharacterized genes. Though deciphering information from genome sequences is an important challenge in functional genomics, increasing number of available complete genome sequences eases the task through comparative genomics. In the past decade, besides *M. tuberculosis*, genome sequences of many other mycobacterial species have been completed. The complete genome sequence of this bacillus is annotated with a large number of putative transcriptional regulators indicating that much of the gene regulation at the level of transcription is yet to be understood. As the biological activity of most genes in adaptive responses is regulated by a set of transcriptional regulators, understanding transcriptional regulation is an important step towards deciphering the biology of the organism.

The genome of *M. tuberculosis* is annotated with a large number of poorly characterized transcriptional regulators, including GntR family regulators. Members of this family play a wide range of roles in cellular physiology. The GntR family of bacterial regulators is named after the *Bacillus subtilis* transcriptional regulator, GntR. This family of regulatory proteins consists of a conserved N-terminal DNA binding domain and C-terminal effector-binding/oligomerisation domain. The work embodied in the thesis aims to understand the GntR family of transcriptional regulators from *M. tuberculosis* and in particular their upstream DNA targets. Further, in order to validate the results obtained, *in vitro* experiments were designed. These computational approaches with experimental validations are described systematically in five chapters. Chapter 1 of the thesis introduces the topic of research that was carried out and it also defines the main objectives of the study. It starts with the importance of Robert Koch’s findings in the history of tuberculosis. It illustrates the current statistical data from WHO which encourage biologists to study tuberculosis in order to develop new interventional strategies and therapeutics. The importance of transcriptional regulators in the understanding of cellular physiology is highlighted. It summarizes the importance of studying *M. tuberculosis* genome and its transcriptional regulators. The structural features associated with GntR family of transcriptional regulators have been described. It highlights the need of protein family/subfamily classification in post-genomic era.

Classification of proteins into families provides valuable clues for better understanding of protein structure and function. It improves the identification of proteins that are difficult to characterize based on pair-wise sequence alignments. Such annotation aids in identifying features associated to the family/subfamily. This precise identification of protein family guides further wet-bench experiments. This thesis classifies the transcriptional regulators belonging to the GntR family from *M. tuberculosis* and explores their upstream DNA targets. Previously on the basis of sequence and structural similarities, six subfamilies have been defined within the GntR family of transcriptional regulators by Rigali and co-workers in 2002. This family of protein includes the subfamilies: FadR, HutC, MocR, YtrA, AraR, and PlmA. Chapters 2 successfully reveals the representation of various subfamilies of GntRs in *M. tuberculosis* proteome. It covers a comprehensive *in silico* sequence analyses of the *M. tuberculosis* GntRs. All putative members possessing features associated to this family of transcriptional regulators, GntRs, were classified into subfamilies. The subfamily members thus classified were scanned for their secondary structural elements that provide an effective means to retrieve information associated with the respective subfamilies. It also demonstrates an evolutionary gene duplication event of a FadR subfamily member in *M. tuberculosis*. Since most of the GntR transcriptional regulators function as a dimer molecule, the finding of a duplication event followed by fusion is an interesting step towards the working of a monomer, as opposed to a dimer.

One of the major challenges to characterize any transcriptional regulators is to identify its DNA targets. Classification of GntR family transcriptional regulators, described in chapter 2, provides a novel approach to identify the DNA targets exhibiting nucleotide preferences as per subfamily, besides the comparison of orthologous upstream regions. Hence, after the identification of subfamilies of GntRs, features associated to operator sites of these subfamilies of regulators were investigated in Chapter 3. This chapter deals with the identification of operator sites for these transcriptional regulators from *M. tuberculosis* and the closest putative orthologs across the sequenced mycobacterial species. Ortholog detection is not only important to identify the operator site but also assist in functional annotation. To strengthen the computational predictions, two of the transcriptional regulators were chosen as representative of the two largest subfamilies of GntRs. These two genes were cloned and expressed in bacterial expression system. Predicted DNA motifs were experimentally investigated using purified recombinant protein.

DNA-protein interactions depend upon the precise interaction between amino acids of the DNA binding protein and the nucleotides in the DNA target site. Therefore, to understand DNA-protein interactions, structural insight is required. Chapter 4 explores the FadR regulator, a well-characterized member of FadR subfamily of regulator from *E. coli*. Structure of this regulator along with DNA has been determined using x-ray crystallography. Considering the best availability and the suitability of the structure of FadR-DNA complex as template, this chapter investigates the structural features associated with DNA recognition for one of the classified regulators. It attempts to summarize the potential amino acids playing critical role in the specific DNA recognition. It also investigates DNA binding abilities of one of the regulator to a range of DNA targets identified in the upstream region of other orthologous transcriptional regulators possessing closely related DNA binding domains.

Mycobacterium smegmatis is a saprophytic species that has been used for years as a model to study *Mycobacterium tuberculosis* in various aspects. The availability of genome sequence of this species offers the possibility of extensive comparative studies.

Hence, besides major emphasis on the GntR regulators from *M. tuberculosis*, Chapter 5 explores the GntR family of transcriptional regulators from *M. smegmatis*. It identifies the GntR family of regulators in *M. smegmatis* followed by classification into respective subfamilies. Furthermore, orthologs of these GntRs were also identified in other species of mycobacteria. Identified GntR orthologs of *M. smegmatis* could serve as a model to decipher the molecular regulation by its orthologs in other pathogenic mycobacteria. Further, potential operator sites for these transcriptional regulators were also analyzed.

Present research work has important implications on understanding the regulatory elements of GntRs. It is the first comprehensive report to describe GntR family of regulators in mycobacteria, in particular *M. tuberculosis* and *M. smegmatis*. The results presented in this study extend the genome annotation of mycobacteria and would add to the present knowledge of the GntR family of transcriptional regulators in microbes. Presented results suggest several directions and opportunity for future research work.

Chapter 1

General Introduction

Tuberculosis is an ancient disease which is known to humans for centuries. Tubercular decay found in the skulls and spines of Egyptian mummies is one of the crucial evidences of its ancient existence [Zimmerman 1979; Nerlich *et al*., 1997; Zink *et al*., 2003; Ziskind and Halioua, 2007]. This devastating disease has taken human life for more than 4,000 years. In spite of the modern advancement in science and medicine, tuberculosis has not yet been eradicated. This disease is not only limited to humans but is also known to infect animals such as cattle and birds [Mackintosh *et al*., 2004; Schmidt *et al*., 2008]. It is estimated that one-third of all human beings are infected with latent state of tuberculosis [WHO report 2008]. The latent tuberculosis infection is a condition where an individual harbors the pathogen without developing any symptoms [Locht *et al*., 2007; Gomez and McKinney, 2004; Frieden *et al*., 2003]. In this state, the bacteria are inactive; kept in check by the body's immune defense system. However, subsequently due to a variety of reasons, the bacillus may become active and result in full-blown disease over time. People having HIV/AIDS with a weakened immune system are particularly vulnerable to activation of this latent form leading to tuberculosis disease. It has thus become one of the major causes of death in HIV-infected individuals [Chintu and Mwaba, 2005; Lalloo and Pillay, 2008]. In fact not only HIV/AIDS, anything which weakens the body's immune defense system can cause latent tuberculosis infection to become active and result in full-blown tuberculosis disease [Keane and Bresnihan, 2008].

1.1 *Mycobacterium tuberculosis*

Identification of *Mycobacterium tuberculosis* in 1882 by Robert Koch, a German physician, scientist and Noble laureate, was one of the great achievements in tuberculosis research. In his well-known speech he said, :"If the importance of a disease for mankind is measured by the number of fatalities it causes, then tuberculosis must be considered much more important than those most feared infectious diseases, plague, cholera and the like. One in seven of all human beings die from tuberculosis. If one only considers the productive middle-age groups, tuberculosis carries away one-third and often more" (www.nobleprize.org).

The release of the drug rifampicin in 1970s and the completion of sequencing of the *M. tuberculosis* genome in 1998 were another two major achievements that revolutionized tuberculosis research [Chakhaiyar and Hasnain, 2004] and the complete genome sequence of this pathogen especially, has provided a strong foundation to improve the understanding of the complex biology of the tuberculosis infection [Young, 2001]. It has also contributed towards deciphering the evolution and pathogenesis of *M. tuberculosis*, and facilitated development of new diagnostic test with increased specificity for tuberculosis [Ernst *et al.*, 2007].

1.2 Statistical overview for tuberculosis prevalence

The status of tuberculosis is under close surveillance by the World Health Organization (WHO) which has for many years assessed tuberculosis incidence, prevalence, and mortality worldwide. In the year 2005, it was estimated that this disease was responsible for around 1.6 million deaths. This increased to 1.7 million by 2006. Globally in 2007, 9.2 million new cases of tuberculosis were reported. It included 4.1 million of new smear positive cases. On an average one-third of the world population is currently infected with tuberculosis. Among these, large numbers of deaths were because of the lethal effect of

HIV/AIDS combined with tuberculosis. This disease has casted its shadow worldwide. Considering different parts of the world, the African region (23%), the Western Pacific region (25%) and the South-East Asia region (36%) together accounted for large number of all notified new and relapse cases and for similar proportions of new smear positive cases in 2006 (WHO reports: http://www.who.int/tb/publications/en/). The wide spread geographical distribution and the pathogenecity trends encourage biologists to study tuberculosis in order to develop new interventional strategies.

1.3 Multiple drug resistance tuberculosis

In the past few years, drug resistant strains of *M. tuberculosis* have become prevalent. In fact, resistance is so wide spread that it is now being identified as multi-drug resistant (MDR-TB) and extreme-drug resistant (XDR-TB) strains [Pelly *et al*., 2004; Jain and Mondal, 2008; Martins *et al*., 2008]. The frightening increase of drug resistance tuberculosis has severely threatened public health control. The frequency of *M. tuberculosis* isolates that is resistant to one or more anti-tuberculosis drugs is increasing day by day. Drug resistance is not confined to one region but is being reported from many tuberculosis-affected areas [Singla *et al*. 2003; Sajduda *et al*. 2004; Blondal 2007; Afanas'ev *et al*. 2007; Sulochana *et al*., 2007; Umubyeyi *et al*., 2008]. Indeed, some of the *M. tuberculosis* strains have developed resistance to all major anti-tuberculosis drugs available. Fortunately many infections can still be cured with extensive and long term chemotherapy. However a need for a new anti-tuberculosis treatment is evident [Okunade *et al*., 2004; Ly and McMurray, 2008].

1.4 Post Genomic Era: Comparative genomics

In the present post-genomic era, genome-sequencing projects are progressing at a fast pace. Hence the availability of the genome sequences of pathogens has generated huge amount of data that is available to scientists via public domain databases. The availability of genome sequences poses challenges and opportunities to a computational biologist to understand the genome function and its complexity [Tsoka and Ouzounis, 2000; Young, 2001]. Comparative genomics, in recent years has gained popularity and momentum for its use in delineating the challenges posed in understanding the complexity of genome function. It is a branch of study that helps a researcher to compare and build the relationship between the genomes of various species and to extract the useful information hidden in the genome sequences. Also, it plays a major role in understanding the functional aspect of un-annotated genes present in the genome [Saghatelian and Cravatt, 2005]. That said, comparative genomics not only deals with the similarity but also the differences in the proteins, DNA, non-coding regulatory regions and the like [Gelfand, 1999]. For example, conserved protein sequence signatures help in determining the family and the class of the newly discovered protein, whereas conserved non-coding DNA motif along the upstream region of the regulatory genes help in determining the regulatory sequences.

As far as the utility of comparative genomic approaches in understanding the genome of *M. tuberculosis* is concerned, at present, sequencing of complete genomes of many mycobacterial species like *M. tuberculosis*, *M. smegmatis*, etc., have been accomplished, and further, large numbers of sequencing projects to unravel the complete genome sequence of several other mycobacterial species are currently in progress (http://www.genomesonline.org/). Comparative genomic analyses of these species will help to identify the genetic basis of phenotypic variation, which in turn helps to highlight the important targets to develop new intervention strategies [Brosch *et al*., 2000]. *M. tuberculosis* genome was predicted to encode about 4000 proteins [Cole *et al*., 1998; Camus *et al*., 2002]. The genome sequence annotation revealed the existence of large number of un-annotated genes (http://genolist.pasteur.fr/TubercuList/). Therefore, comparative genomic analyses for this genome sequence would help to decipher the role of uncharacterized genes.

1.5 Pathogen success: transcriptional regulation

Genes in prokaryotes are arranged in an operon [Jacob *et al*., 1960; Jacob *et al*., 1964]. It is a series of genes that are transcribed together as a single polycistronic mRNA. It consists of initiation signal (promoter), regulatory sequences (operator), genes to be transcribed, and a termination signal (Figure 1.1). There are two classes of genes in the operon, structural genes and regulatory genes. Structural genes code proteins those are required for enzymatic and structural functions in the cell. The regulator gene encodes the regulatory proteins that regulate the gene expression of the cell.

Success of a pathogen lies partly in its ability to sense the varying environmental condition and to adapt accordingly [Bruggemann *et al*., 2006]. This adaptation depends upon the coordinated behavior of gene expression achieved by collective role of a set of transcriptional regulators which the genome encodes. These transcriptional regulators bind to specific sites on the DNA, generally located near the promoter region and thereby control the transcription of genetic information from DNA to RNA [Latchman, 1997;

Karin, 1990]. They can either repress (repressor) or activate (activator) the RNA polymerase activity [Ptashne and Gann, 1997].

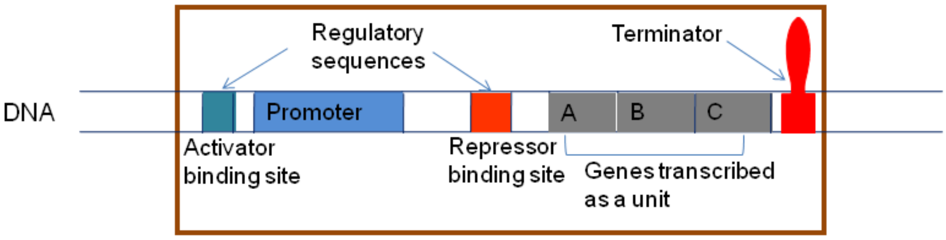


Figure 1.1: Operon: a unit of transcription in prokaryotes.

The genome annotation of *Mycobacterium tuberculosis* revealed the presence of over 150 transcriptional regulators [Cole *et al*., 1998; Camus *et al*., 2002]. Success of this bacillus partly lies in coordinated gene regulation via these regulators. Functional role of majority of these regulators, their regulatory elements, and target genes remain largely unknown (http://genolist.pasteur.fr/TubercuList/). The availability of a large number of sequenced mycobacterial genomes allows us to conduct systematic studies on gene regulatory systems [Stormo and Tan, 2002].

1.6 Protein family identification

Evolutionary events such as gene duplication and combination, have resulted in the formation of many new proteins [Taylor and Raes, 2004; Orengo and Thornton 2005; Bashton and Chothia, 2007]. This implies that on the basis of protein sequence and structure, relationship among the proteins can be categorized as families whose members descended from a common ancestor [Saier, 1996; Vogel and Chothia, 2006].

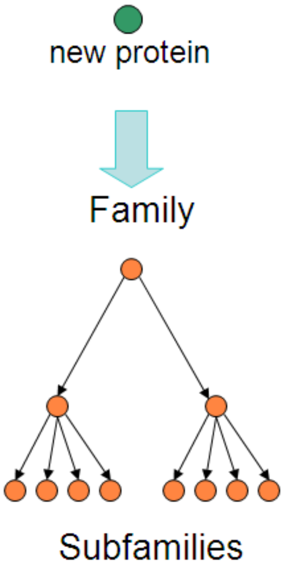


Figure 1.2: Placing a new protein in a protein hierarchy.

Identification of protein family is the key to the functional annotation and the exploration of diversity of protein function. It provides valuable clues for the determination of structure and function [Saqi and Wild 2005; Marsden *et al*., 2006]. Classification of proteins into its family has many advantages: (1) it improves the identification of proteins that are difficult to annotate based on pair-wise alignments; (2) it enriches the sequence database by enhancing protein family based annotation; (3) it provides a valuable way to recover relevant biological information from vast amount of data; (4) it reveals the underlying gene families, the analyses of which is important for comparative genomics [Wu *et al*., 2003; Tatusov *et al*., 1997]. It is clear from the facts mentioned that with accelerated accumulation of genome sequence data; there is a need to carry out protein family/subfamily identification (Figure 1.2).

1.7 Prokaryotic regulator families

Proper regulation of transcription is crucial for the cell to adapt to its environment. For single cell bacteria, such regulation must be highly responsive because its environment can change instantly and drastically. This change can be in terms of temperature, nutrients, water availability, and the presence of toxic substances. To adapt to the environment, cell encodes transcriptional regulators to regulate the gene expression [Zhou and Yang, 2006]. It is observed that large number of transcriptional regulators from prokaryotes utilize helix-turn-helix (HTH) motif to bind to their target DNA sites [Perez-Rueda and Collado-Vides, 2000; Perez-Rueda and Collado-Vides, 2001; Karmirantzou and Hamodrakas, 2001; Aravind *et al*., 2005]. HTH domain-containing families of transcriptional regulators are involved in various disparate biological processes [Rosinski and Atchley, 1999].

Primarily, prokaryotic transcriptional regulators are classified into protein families [Elofsson and Sonnhammer, 1999]. These families are generally defined by sequence homology. So far, many protein families of transcriptional regulators have been identified in prokaryotes [Perez-Rueda and Collado-Vides, 2000], like GntR [Haydon and Guest, 1991; Rigali *et al*., 2002], LysR [Schell, 1993], AraC [Martin and Rosner, 2001], TetR [Aramaki *et al*., 1995; Ramos *et al*., 2005], LuxR [Fuqua *et al*., 1994], LacI [Nguyen and Saier, 1995], ArsR [Busenlehner *et al*., 2003], IclR [Sunnarborg *et al*., 1990], MerR [Brown *et al*., 2003], AsnC [Friedberg *et al*., 2001], MarR [Alekshun and Levy, 1999], NtrC [Morett and Segovia, 1993], OmpR [Martinez-Hackert and Stock, 1997], DeoR [van Rooijen and de Vos, 1990], and CRP [Korner *et al*., 2003] (Table 1.1).

Table 1.1: Prokaryotic regulator families

Family	DBD motif	Position	Action	Some regulated functions
GntR	HTH	N-terminal	Repressor	General metabolism
LysR	HTH	N-terminal	Activator/repressor	Carbon and nitrogen metabolism
AraC/XylS	HTH	C-terminal	Activator	Carbon metabolism, stress response and pathogenesis
TetR	HTH	C-terminal	Repressor	Biosynthesis of antibiotics, efflux pumps, osmotic stress
LuxR	HTH	C-terminal	Activator	Quorum sensing, biosynthesis and metabolism
LacI	HTH	N-terminal	Repressor	Carbon source utilization
ArsR	HTH	Central	Repressor	Metal resistance
IcIR	HTH	N-terminal	Repressor/activator	Carbon metabolism, efflux pumps
MerR	HTH	N-terminal	Repressor	Resistance and detoxification
AsnC	HTH	N-terminal	Activator/repressor	Amino acid biosynthesis
MarR	HTH	Central	Activator/repressor	Multiple antibiotic resistance
NtrC (EBP)	HTH	C-terminal	Activator	Nitrogen assimilation, aromatic amino acid synthesis, flagella, catabolic pathways, phage response
OmpR	Winged helix	C-terminal	Activator	Heavy metal and virulence (response regulator of a two-component system)
DeoR	HTH	N-terminal	Repressor	Sugar metabolism
Cold shock	RNA binding domain (CSD)	Variable	Activator	Low-temperature resistance
Crp	HTH	C-terminal	Activator/repressor	Global responses, catabolite repression and anaerobiosis

(Adapted from Ramos JL *et al*. 2005)

1.8 GntR family of transcriptional regulators

GntR family, named after gluconate regulator, is a very interesting protein family [Buck and Guest, 1989; Haydon and Guest, 1991]. Transcriptional regulators of this family show a similar N-terminal region containing winged helix-turn-helix DNA-binding domain but have a highly diverse C-terminal region with an effector binding and/or oligomerization domain [Rigali *et al*., 2002]. This family is best defined by a profile covering the N-terminal DNA binding domain. Due to their diverse C-terminal region different members of GntR family regulators can be divided into six sub-families: FadR, HutC, MocR, YtrA, AraR and PlmA [Lee *et al*., 2003; Rigali *et al*., 2004]. All the members of these subfamilies are reported from Gram-positive and Gram-negative bacteria except PlmA subfamily. Members of PlmA subfamily were observed from cynobacterial species [Lee *et al*., 2003].

Diversity of the C-terminal domain leads these regulators to recognize diverse range of effector molecules. Large numbers of these regulators are involved in the regulation of gene expression in response to oxidized substrates related to either amino acid metabolism or various metabolic pathways such as glycolate [Pellicer *et al*., 1999], pyruvate [Quail and Guest, 1995], lactate [Nunez *et al*., 2001], malonate [Lee *et al*., 2000] or gluconate [Fujita *et al*., 1986; Reizer *et al*., 1991]. These regulators also respond to small diverse molecules, like histidine (HutC) [Allison and Phillips, 1990], long chain fatty acids [Quail *et al*., 1994], trehalose 6-phosphate [Matthijs *et al*., 2000; Schock and Dahl, 1996] or alkylphosphonate [Chen *et al*., 1990]. Some of the regulators of the PlmA subfamily are also reported to be involved in regulation of plasmid maintenance function in *Anabaena sp. strain PCC 7120* [Lee *et al*., 2003].

1.8.1 Domain organization of GntR regulators

In general, GntR family transcriptional regulators possess a DNA binding domain (DBD) and the effector binding and/or oligomerization (E-b/O) domain. DNA binding domain consists of three α-helices with small β-strand making specific winged helix turn helix domain. This domain is known to be conserved across the regulators of the family (Figure 1.3) [Rigali *et al*., 2002].

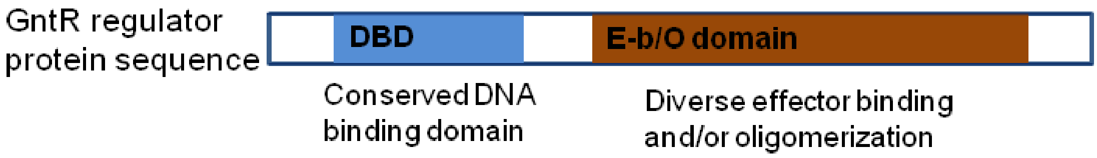

GntR regulator protein sequence
DBD
E-b/O domain
Conserved DNA binding domain
Diverse effector binding and/or oligomerization

Figure 1.3: Schematic representation of the domain organization of GntR family regulators

In contrast to the DNA binding domain, effector binding and/or oligomerization (E-b/O) domain is very diverse in sequence as well as in structure. Each subfamily regulator possesses the specific secondary structural pattern characteristic to that subfamily. E-b/O domain of FadR subfamily consists of all α-helices with an average length of about 160 amino acids. Based on the number of α-helices FadR regulators have been classified into two groups; the FadR group and VanR group. An important difference between them is that VanR has six α-helices in the effector binding and/or oligomerization domain whereas the FadR subgroup has seven α-helices [Rigali *et al*., 2002]. A second type of subfamily regulator, HutC, contains both α-helices and β-strands secondary structure elements in its E-b/O domain. This domain adopts a fold similar to chorismate lyases (*Escherichia coli* UbiC) hence it is named UbiC transcription regulator-associated (UTRA) domain [Aravind and Anantharaman, 2003]. Regulators of third type, MocR, are generally large in size. The average length of the E-b/O domain of these regulators is about 350 amino acids long. These regulators contain both α-helices and β-strands and exhibit homology to class I aminotransferase proteins [Sung *et al*., 1991], which requires pyridoxal 5'-phosphate (PLP) as a co-factor. These regulators exhibit a PLP attachment site with a conserved lysine residue [Magarvey *et al*., 2001]. Regulators of fourth type, YtrA, are much shorter in protein size as compared to other members of GntR family. The approximate size of the C-terminal region is about 50 amino acid residues with only two α-helices. How these regulators manage in such a small size to oligomerise is interesting. Recent report about one of the members of this subfamily, CGL2947 from *Corynebacterium glutamicum*, provides insight to the dimerization and the ability to bind effectors with small C-terminal domain [Gao *et al*., 2007]. Known representations of the other subfamilies of GntR regulators are very few [Rigali *et al*., 2004]. PlmA subfamily regulators show highest sequence similarity with YtrA and MocR subfamily. It is also believed that this subfamily arose from the ancestral sequences shared by one of these subfamilies [Lee *et al*., 2003]. AraR subfamily regulators exhibit a chimeric organization comprising a small N-terminal DNA-binding domain that contains a winged helix-turn-helix motif similar to that seen with the GntR family and a larger C-terminal domain homologous to that of the LacI/GalR family. As GntR family regulators are determined with the signature sequence of N-terminal DNA binding domain, these are part of the GntR family regulators [Minezaki *et al*., 2005; Mota *et al*., 1999].

1.8.2 GntR regulators in mycobacteria

So far some of the GntR family regulators in mycobacteria have been characterized. One of the FadR subfamily regulators, PipR, is known to be involved in the regulation of piperidine and pyrrolidine metabolism in *M. smegmatis* [Poupin *et al*., 1999]. Another GntR regulator from *M. smegmatis*, PhnF, is reported to act as a repressor of the *phnDCE* operon. Involvement of this operon has been shown for the adaptation of *M. smegmatis* to phosphate-limited conditions [Gebhard and Cook, 2008]. One of the GntR family regulators from *M. tuberculosis*, Rv0165c, is identified as a repressor. It is responsible for the intracellular repression of the *mce1* operon [Casali *et al*., 2006]. Expression of the *M. tuberculosis mce1* operon is crucial to obtain the host proinflammatory response that is significant for the establishment of a persistent infection [Shimono *et al*., 2003].

1.8.3 GntR regulator-DNA interaction: a structural insight

DNA-protein interaction is a precise interaction between amino acids from the DNA binding protein and the nucleotides in the DNA target site. Hence in order to better understand the DNA-protein interactions, structural knowledge of the DNA protein complex is vital. In the past few years, there has been a significant advancement towards determining the structure of biological sequences [Todd *et al*. 2005]. Recently some of the GntR family protein structures have been determined (PDB codes: 2HS5, 2P19, 2DI3, 3DBW, 3DDV, 3EET, 2FA1, 3C7J, 3CNV, 2DU9, 1H9G, 1HW1, 1H9T, etc.). Still, GntR regulator-DNA structural knowledge is very limited. Among the GntR family of regulators, *E. coli* FadR is one of the well-characterized proteins. Cocrystallization of this protein with its operator site provided the major insight available in DNA-protein interaction of GntR family of transcriptional regulators [van Aalten *et al.*, 2001; Xu *et al.*, 2001].

1.9 Subfamily classification of the GntR family of regulators

With the advent of large number of sequencing projects, sequence data is being generated at accelerated pace. This large volume of sequence data is difficult to handle manually, thereby it require computational analyses to sort them into different groups. Generally, signature sequences may be used to classify new proteins into various families. Identification of GntR family regulators is also carried out with the help of conserved signature sequence of the DNA binding domain. But the heterogeneity of C-terminal leads to a further classification of GntR family of regulators into sub-families. It requires careful examination of the protein sequences. Using phylogenetic analyses, for the first time Rigali and co-workers attempted to classify the GntR family of proteins into four specific subfamilies, FadR, HutC, MocR, YtrA [Rigali *et al.*, 2002]. Later, based on conserved DNA binding domain, two more subfamilies were added into the account of GntR family of regulators, AraR and PlmA [Lee *et al.*, 2003; Rigali *et al.*, 2004]. Although further new regulators have been identified based on the variations observed in the C-terminal domains and are being defined as a new subfamily of GntR regulators [Hillerich and Westpheling, 2006], in this study of GntR family of regulators, I have restricted my analyses to the well-established six subfamilies of GntR regulators (Figure 1.4).

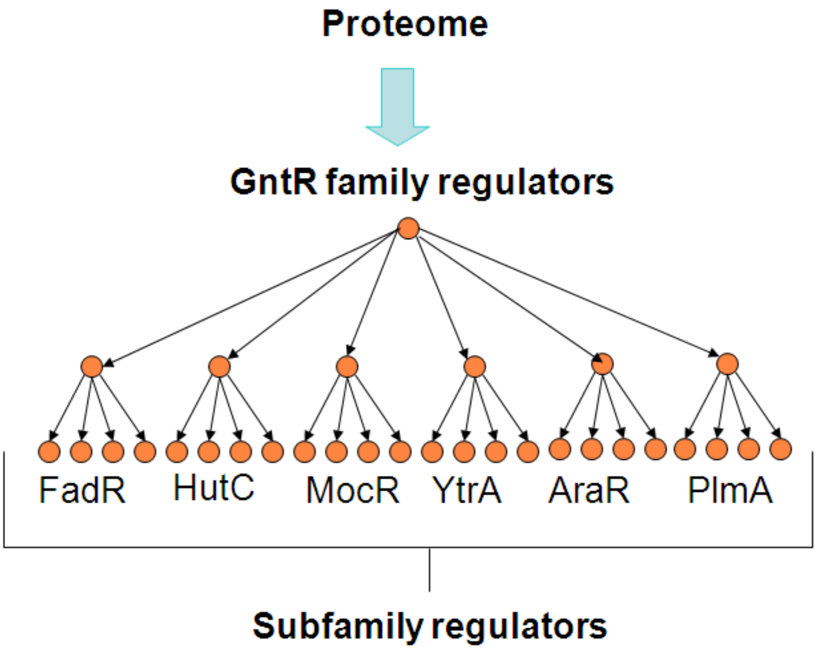


Figure 1.4: Classification of GntR family proteins into subfamilies.

1.10 GntR family constrain operator sites

Knowledge of regulatory elements is vital to understand the regulon of any transcriptional regulator. In the past decade, different approaches have been used to address this problem [Thieffry *et al*., 1998; Bulyk *et al*., 2004; Siddharthan *et al*., 2005;]. Generally, identification of these DNA elements relies on an extensive set of known target genes [Yellaboina *et al*., 2004; Bailey *et al*., 2006; Ranjan *et al*., 2006]. Thus, for identifying novel transcriptional regulators most of these approaches are futile. Using family-wise approach to address DNA targets for such novel transcriptional regulator could be quite significant [Kaplan *et al.,* 2005]. It is observed that most structurally related protein families of transcriptional regulators show similarity in their DNA binding domain that influences similar DNA target recognition [Sandelin and Wasserman, 2004]. Therefore incorporating the familial profile, if any, is useful to locate the DNA targets in the genome.

Although, GntR family of regulators possess similar DNA binding domain that facilitates regulators to recognize similar DNA targets, it is difficult to find consensus for all the family members. This practical difficulty can be explained due to diversity in the C-terminal effector binding and/or oligomerization domain which imposes diverse sterical constraints. However, this problem can be overcome by further division of GntR family into subfamilies, wherein each classified subfamily member possess similar C-terminal domain effector binding and/or oligomerization domain and DNA binding domain. It was observed by Rigali and co-workers, that regulators belonging to a subfamily exhibit nucleotide preferences in their DNA targets [FadR (TNGT-$N_{(0-3)}$-ACNA), HutC (GTNTANAC) and YtrA (GTNNTA$N_{(0-3)}$TANNAC)]. However due to lack of experimental data, consensus for all the subfamilies have not been determined [Rigali *et al*., 2002; Rigali *et al*., 2004].

1.11 Conservation of operator sites among the upstream sequences of orthologous genes

Being auto-regulatory in nature, many GntR family regulators are reported to interact with their upstream operator sites. These regulatory elements are short and variable DNA motifs which are generally considered to be conserved across the upstream regions of orthologs. Analysis of these DNA regions is a very efficient approach to locate these

DNA elements [Miziara *et al*., 2004; McCue *et al*., 2001], which relies on the evolutionary conservation of regulatory elements in related species [Bailey *et al*., 2006]. The rapidly growing number of available complete genome sequences facilitates to perform this task. McCue and co-workers showed that selection of upstream sequences from three species is optimal for identification of regulatory elements based on evolutionary conservation. However, it is important to decide the species of interest that are likely to be useful for sequence alignment [McCue *et al*., 2002], because more divergent species will have less sequence homology between orthologous genes. Therefore, the key is to select species that are related enough to detect homology, but divergent enough to maximize non-alignment "noise".

1.12 Objective and overview of the present work

Recently a large number of mycobacterial genome sequences including *M. tuberculosis* have been sequenced and are available via public domain databases and repositories. Availability of these large sequence data enables biologists to employ the comparative functional genomics to explore the coding and non-coding sequences to understand the role of uncharacterized genes. The genome of *M. tuberculosis* is annotated with large number poorly characterized transcriptional regulators, including GntR family regulators. The GntR family of bacterial regulators is named after the *Bacillus subtilis* transcriptional regulator, GntR. This family of regulatory proteins consists of the conserved N-terminal DNA binding domain and the diverse C-terminal effector-binding/oligomerisation domain.

The aim of my study is to enrich the knowledge of the repertoire of GntR regulators in mycobacteria, particularly in *M. tuberculosis*. The primary objectives of my study are first, to classify the GntR family of transcriptional regulators from *M. tuberculosis* and *M. smegmatis* into specific subfamilies based on sequence and secondary structural features. Such annotation aids in identifying the features associated to the family/subfamily. Second, to identify the upstream operator sites for the GntR regulators using clues from the classification based on sequence and secondary structural features. And finally, to experimentally validate computational predictions using *in vitro* DNA binding experiments for a select few observations.

Chapter 2

GntR family transcriptional regulators from Mycobacterium tuberculosis

Part of the work is published as:

Vindal V, Ranjan S and Ranjan A (2006) *In silico* analysis and classification of *Mycobacterium tuberculosis* putative GntR regulators. *Tuberculosis.* 87, 242–247.

2.1 INTRODUCTION

The genome sequence of *Mycobacterium tuberculosis* was completed in 1998 [Cole *et al.*, 1998] and still years after, a large number of *M. tuberculosis* proteins are annotated as hypothetical and are poorly characterized. These include proteins belonging to the putative GntR family (http://genolist.pasteur.fr/TubercuList/). Usually, clues for the functional characterization of any newly discovered protein can be obtained from its sequence similarity with experimentally well characterized proteins of other species. However, there are a large number of proteins, which do not share considerable sequence similarity to the proteins of other species and therefore are documented as hypothetical proteins. Classification of such proteins into their specific families/subfamilies can be considered as the first task towards their identification and characterization [Wu *et al.*, 2003]. Many such proteins have been computationally identified and classified into protein families on the basis of their sequence similarity. Among these, there are proteins belonging to the GntR family that are named after the gluconate regulator. Members of the GntR family exhibit a similar N-terminal region containing a winged helix-turn-helix DNA-binding (Db) domain but have a highly diverse C-terminal region containing the effector-binding and/or oligomerization (E-b/O) domain. Due to their diverse C-terminal region, different members of GntR family regulators can bind to different effectors and thereby can be classified further into subfamilies, such as FadR, HutC, MocR, YtrA, AraR and PlmA **[**Lee *et al.*, 2003; Rigali *et al.*, 2004**]**. A high sequence similarity of the N-terminal region enables relatively easy sequence-based identification of the protein member to a part of the GntR family. An important challenge is to further classify these proteins into more informative subfamilies based on the sequence of the highly divergent

C-terminal region. This heterogeneity led Rigali and co-workers for the first time to attempt the classification of GntR family of regulators in bacterial genome [Rigali *et al*., 2002]. It provides useful clues for the identification of the DNA targets of the transcriptional regulators, as most of the regulators of these families are known to exhibit nucleotide preferences in the operator sites. This chapter undertakes the study of known and classified GntR family of transcriptional regulators from literature as a representative of various subfamilies. It reports the analyses of C-terminal domain as well as N-terminal DNA binding domain and classifies all GntRs into subfamilies. This subfamily classification has major implications on identification of their upstream DNA targets.

2.2 EXPERIMENTAL PROCEDURE

2.2.1 Selection of GntR family members in *M. tuberculosis*

Figure 2.1 schematically represents the approach employed to classify the GntR family of transcriptional regulators. Primarily *M. tuberculosis* proteome was scanned for proteins possessing the GntR domain using GntR protein family profile. This GntR family profile was obtained from pfam web server. All the hits identified with E-value less than 10^{-5} were considered [Eddy, S.R., 1998]. Rest of the GntR family members used in this study were collected from the SWISS-PROT/TrEMBL/GenBankTM sequence databases by SwissProt number (Table 2.1).

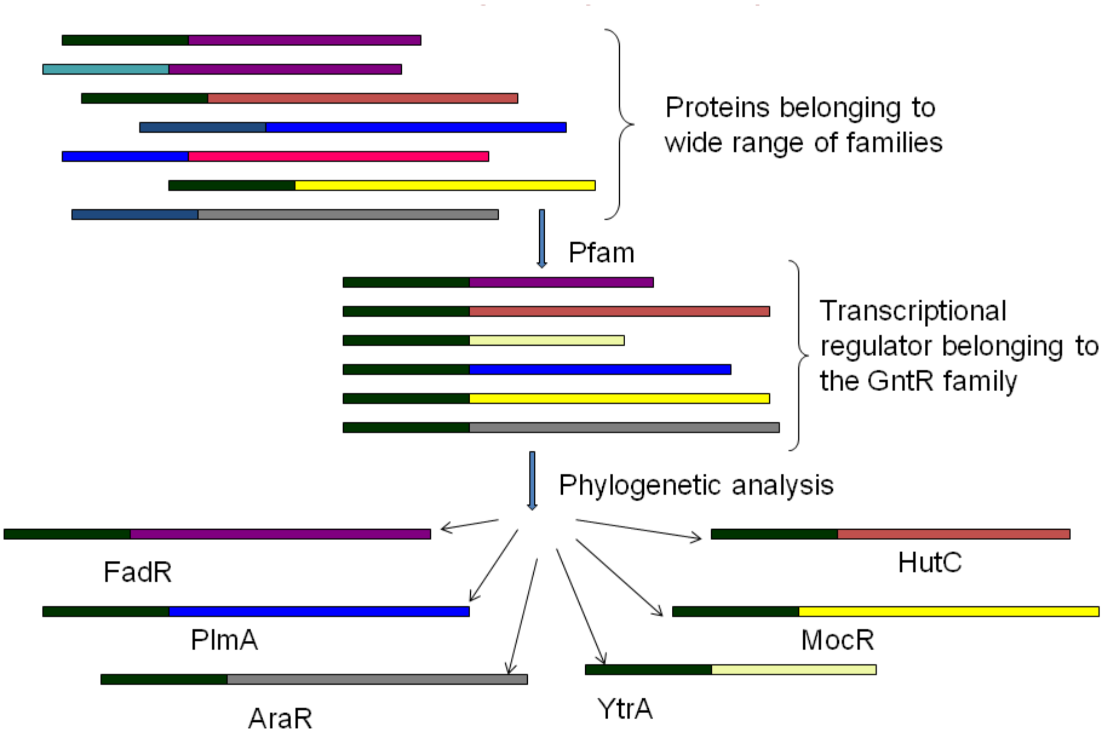

Proteins belonging to wide range of families
Pfam
Transcriptional regulator belonging to the GntR family
Phylogenetic analysis
FadR
HutC
PlmA
MocR
AraR
YtrA

Figure 2.1: Schematic representation of the identification and classification strategy for GntR family of transcriptional regulators.

2.2.2 Dot plot analysis

To examine the substantial region of similarity graphically at a glance, global sequence comparison was carried out with Dot Plot analysis program DOTMATCHER. A web interface of this EMBOSS program is freely available at http://bioweb.pasteur.fr/seqanal/interfaces/dotmatcher.html.

2.2.3 Secondary structure prediction

In order to analyze the protein sequences based on not only their sequence homology but also their secondary structure, the secondary structures of all bacterial GntR regulators, including *M tuberculosis,* were studied using 3DPSSM [Kelley *et at.*, 2000], Jpred [Cuff *et al*., 1998] and SsPro **[**Cheng *et al*., 2005**].** Consensus results thus obtained were considered for greater validity.

2.2.4 Multiple sequence alignments and phylogenetic tree construction

Multiple sequence alignment was generated with MULTALIN [Corpet, 1988] followed by manual improvement according to the predicted secondary structure consensus. Distances between aligned proteins were computed using PROTDIST program with Dayhoff PAM matrix [Young *et al*., 1979]. The FITCH program estimated phylogenies from distances in the matrix data using the Fitch-Margoliash algorithm **[**Fitch and Margoliash, 1967**]**. Further bootstrap values involving 1000 replicates were calculated using program SEQBOOT. This program is intended to generate multiple data sets that are resampled versions of the input data set. All programs used in the study are a part of the PHYLIP package developed by Feldenstein for inferring phylogenies [Felsenstein, 1989]. The tree was drawn using the TREEVIEW program [Page, 1996].

2.2.5 Structure based multiple sequence alignment of the classified subfamilies

Using MULTALIN **[**Corpet, 1988**]**, multiple sequence alignment of the C-terminal domains of all classified *M. tuberculosis* GntRs was carried out with their respective subfamily representatives. This alignment was adjusted as per the secondary structure predictions and the consensus sequence was derived from multiple sequence alignments. High and low consensus levels were fixed arbitrarily at 80% and 40% of identity and are represented respectively by capital and lowercase letters. Consensus symbol ! used for anyone of IV; $ is anyone of LM; % is anyone of FY; # is anyone of NDQEBZ. All abbreviations are listed in Table 2.1.

Table 2.1: A list of GntR regulators used as a representative belonging to the various subfamilies (SF)

SF	*Organism (abbreviation)*	Protein	A. Acid	Swiss Prot ID
FadR	*Acinetobacter sp. (strain ADP1) (Asp)*	VanR	251	O24839
	Mycobacterium smegmatis (Msm)	PipR	245	Q9XDB1
	Chelatobacter heintzii (Che)	NtaR	210	P54988
	Rhizobium leguminosarum (Rle)	MatR	222	Q9JP74
	Escherichia coli (Eco)	LldR	258	P0ACL7
	Escherichia coli (Eco)	PdhR	254	P0ACL9
	Escherichia coli (Eco)	GlcC	254	P0ACL5
	Escherichia coli O157:H7 (Eco)	FadR	238	P0A8V8
	Escherichia coli (Eco)	DgoR	229	P31460
HutC	*Pseudomonas putida (Ppu)*	HutC	248	P22773
	Streptomyces ambofaciens (Sam)	KorSA	259	Q07191
	Escherichia coli (Eco)	PhnF	241	P16684
	Salmonella typhi (Sty)	PhnR	239	P96061
	Bacillus subtilis (Bsu)	TreR	238	P39796
	Streptomyces lividans (Sli)	XlnR	252	Q9ACN8
	Bacillus subtilis (Bsu)	YvoA	243	O34817
	Escherichia coli (Eco)	FarR	240	P13669
MocR	*Rhizobium meliloti (Rme)*	MocR	493	P49309
	Streptomyces venezuelae (Sve)	PdxR	532	Q9FDB4
	Salmonella typhimurium (Sty)	PtsJ	430	P40193
	Bacillus subtilis (Bsu)	YcxD	444	Q08792
	Bacillus subtilis (Bsu)	YcnF	479	P94426
	Bacillus subtilis (Bsu)	YdfD	482	P96681
	Bacillus subtilis (Bsu)	YhdI	469	O07578
	Escherichia coli (Eco)	YjiR	470	P39389
	Rhodobacter sphaeroides (Rsp)	YrdX	456	Q01856
YtrA	*Bacillus halodurans (Bha)*	BH0651	123	Q9KF35
	Bacillus halodurans (Bha)	BH2647	123	Q9K9J9
	Staphylococcus aureus (Sau)	SAV1934	126	Q99SV4
	Bacillus subtilis (Bsu)	YhcF	121	P54590
	Bacillus subtilis (Bsu)	YtrA	130	O34712
AraR	*Bacillus subtilis (Bsu)*	P96711	362	P96711
	Bacillus halodurans (Bha)	Q9KBQ0	375	Q9KBQ0
	Bacillus stearothermophilus (Bst)	Q9S470	364	Q9S470
PlmA	*Synechocystis sp. strain PCC 6803 (Ssp)*	sll1961	388	P73804
	Anabaena sp. strain PCC 7120(Asp)	Q8YXY0	328	Q8YXY0
	Synechococcus elongates (Sel)	Q8DH43	367	Q8DH43
	Trichodesmium erythraeum IMS101 (Ter)	Q3HFX5	327	Q3HFX5

2.3 RESULTS AND DISCUSSION

2.3.1 GntR family of transcriptional regulators

In general, size of the GntR proteins varies in range from 98 to 470 amino acid residues, carrying the conserved GntR family signature or GntR domain **[**Perez-Rueda and Collado-Vides, 2000**]**. This conserved GntR family signature enables the identification of these regulators. My study also began with analyses of *M. tuberculosis* proteins carrying GntR family signatures. Of all the *M. tuberculosis* proteins, seven proteins Rv0043c, Rv0165c, Rv0494, Rv0586, Rv0792c, Rv1152, and Rv3060c were identified with GntR family signature. However this result was consistent with earlier listed putative GntR family transcriptional regulators from *M. tuberculosis* at TubercuList site (http://genolist.pasteur.fr/TubercuList/). Additionally, one of the regulator, Rv3060c, was observed with two GntR-like domains (Table 2.2). Apart from possessing two GntR-like domains, this protein was nearly double the size of the general range of GntR family of regulators [Perez-Rueda and Collado-Vides, 2000].

2.3.2 Rv3060c shows internal duplication

Larger size with two GntR-like domains suggested an interesting possibility of internal gene duplication. To test this idea, a self-dot plot of Rv3060c was generated. The dot plot showed two off diagonal lines in addition to a diagonal line suggesting that the N-terminal and the C-terminal halves of Rv3060c are similar to each other (Figure 2.2). This indicated that two genes encoding the GntR-like regulators have fused to produce a larger gene. But whether this fusion occurred as a result of gene duplication followed by fusion or is a fusion of two distant *gntR*-like genes is yet to be answered. Since most of the GntR transcriptional regulators function as a dimer molecule, it would be interesting to find whether Rv3060c works as a monomer as opposed to a dimer **[**Raman *et al.* 1997**]**.

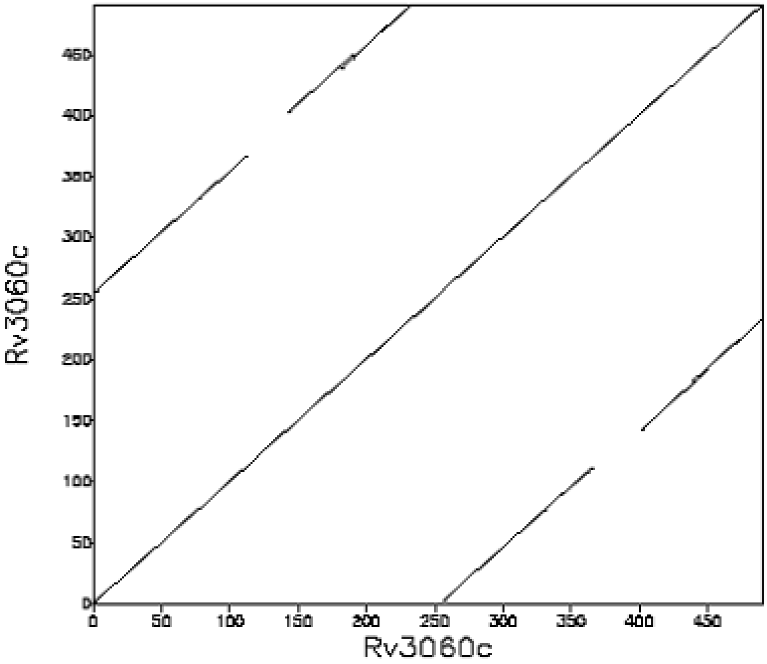


Figure 2.2: A Dotmatcher plot of GntR Rv3060c shows internal duplication. A self Dotmatcher plot of Rv3060c, using window size of 35 and threshold value of 30, show two sets of non-overlapping off diagonal lines, a pattern which is associated with internal duplication.

Table 2.2: ***Mycobacterium tuberculosis*** **GntRs identified using GntR family profile**.

Protein	Swiss Prot ID	Amino Acid	Score	E-value	GD
Rv0043c	P67737	244	52.4	6.60E-13	1
Rv0165c	Q79G00	264	59.8	3.90E-15	1
Rv0494	P67739	242	74.9	1.10E-19	1
Rv0586	P67741	240	74.9	1.10E-19	1
Rv0792c	O86331	269	63.9	2.30E-16	1
Rv1152	O06550	121	56.2	4.70E-14	1
Rv3060c	P95098	490	42	9.00E-10	2

(Note: GD - GntR-like domain)

2.3.3 Classification of putative *M. tuberculosis* GntRs into subfamilies

In order to classify putative GntR regulators into subfamilies, a distance based unrooted tree of putative GntRs and other known and classified GntRs was constructed (Tables 2.1, Figure 2.3). To construct the tree all putative GntRs were first aligned with other classified GntR proteins and subsequently the alignment was manually improved based on predicted secondary structure [Rigali *et al*., 2002]. Since Rv3060c showed internal duplication it was represented as two sequences Rv3060c I (residue nos. 1 - 260) and Rv3060c II (residue nos. 261 - 490), representing N-terminal and C-terminal part of the Rv3060c protein. The improved alignment of known bacterial GntR regulators with *M. tuberculosis* putative GntRs was used to construct an unrooted tree, which revealed various GntR proteins organized into different clusters (Figure 2.3). The clusters represent subfamilies that emerge from the constructed tree. Each cluster in the tree represents a set of branches belonging to a subfamily. In this tree all putative members of the *M. tuberculosis* GntR family were classified into three subfamilies FadR, HutC and YtrA and none of them clusters with other GntR subfamilies such as MocR, AraR and PlmA. Constructed tree was further validated with bootstrapping, involving 1000 replicates. All the values were also incorporated in the constructed tree. Figure 2.3 shows that all subfamily branches are clustered with high bootstrap values.

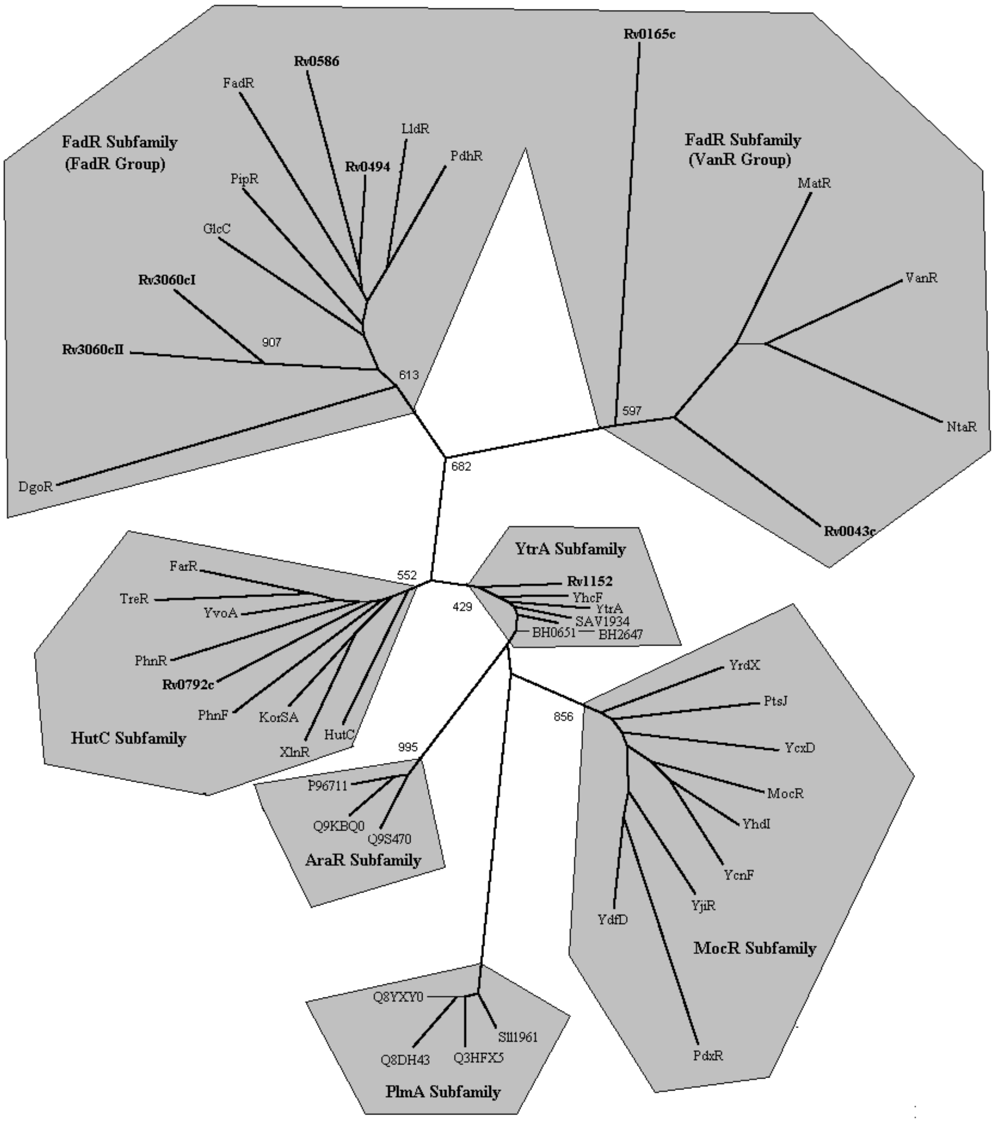


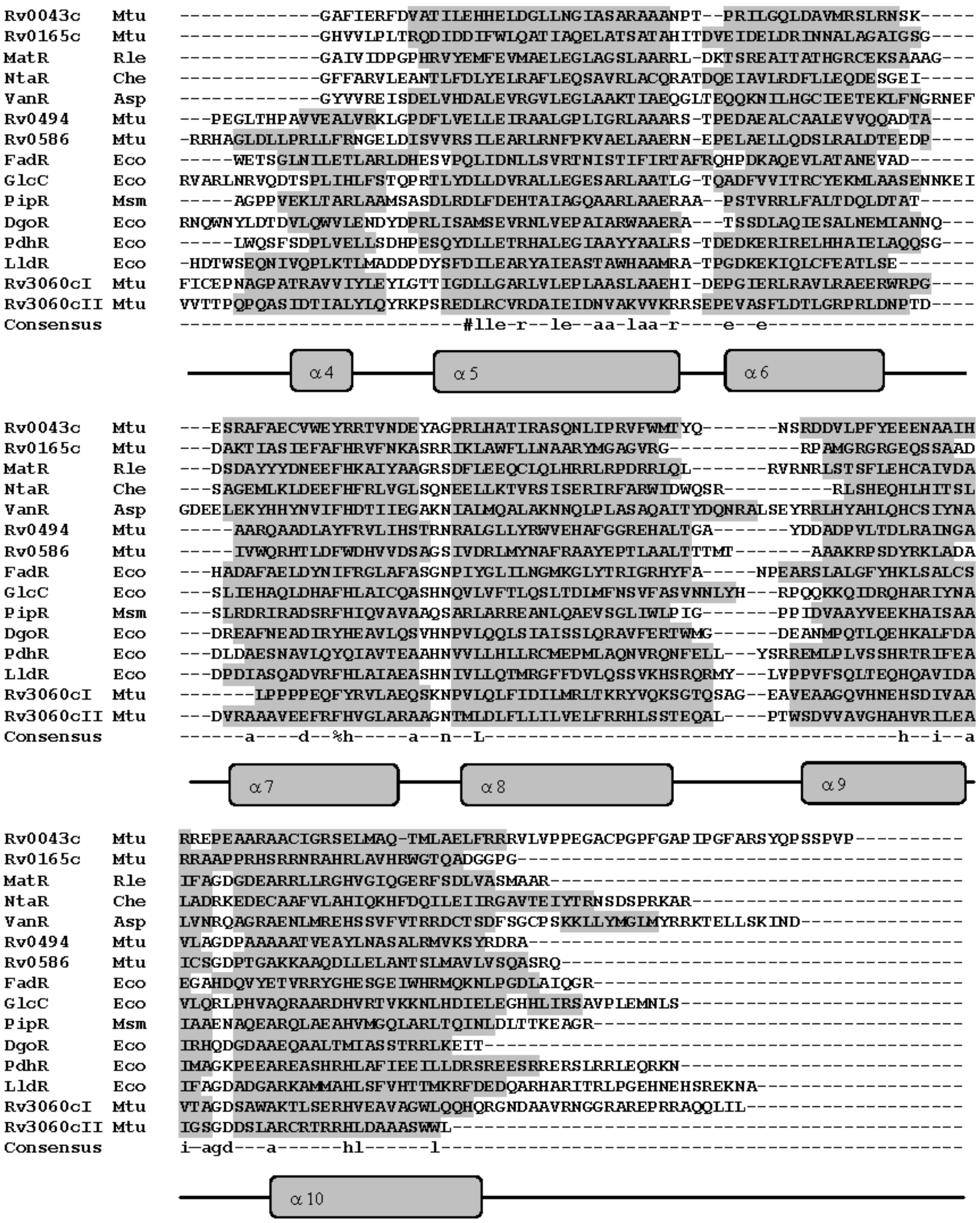
Figure 2.3: Classification of *Mycobacterium tuberculosis* putative GntRs: The unrooted tree of proteins belonging to the GntR family in different bacterial genomes including *M. tuberculosis* is shown here. All GntR regulators are clustered into six subfamilies. FadR subfamily is branched again into two subgroups (FadR group and VanR group). (Note: The GntR proteins and their swiss-prot accession no. are mentioned in table 2.1 and 2.2)

2.3.4 FadR subfamily regulators of *M. tuberculosis*

Five out of the seven putative GntR genes - Rv0043c, Rv0165c, Rv0494, Rv0586 and Rv3060c (I and II) - were classified as members of FadR subfamily in *M. tuberculosis* genome (Figure 2.3). The FadR subfamily is the most represented GntR and constitutes one of the largest subfamilies of bacterial GntR, which is also evident in case of *M. tuberculosis*. The large FadR subfamily consists of proteins with all α-helical C-terminal domains with an average length of about 160 amino acids. Based on the number of helices, FadR has been previously reported to encompass two groups, namely FadR and VanR **[**Rigali *et al.*, 2002**]**. An important difference between them is that VanR group has six α-helices in the C-terminal E-b/O domain whereas FadR group has seven α-helices **[**Rigali *et al.*, 2002**]**. Among proteins belonging to *M. tuberculosis* FadR subfamily, Rv0043c and Rv0165c showed signatures similar to the VanR group and the remaining three regulators Rv0494, Rv0586 and Rv3060c (I and II) to the FadR group (Figure 2.4). The tree (Figure 2.3) also revealed that the two parts of Rv3060c, Rv3060c I and Rv3060c II, are spaced closer to each other than any other GntR protein and share a common and an immediate evolutionary ancestor which clearly suggests that the Rv3060c arose by gene duplication followed by gene fusion and not by mere gene fusion of the two GntR with different origins. Both the gene parts also exhibit more than 90% bootstrap value.

2.3.5 HutC subfamily regulator of *M. tuberculosis*

One of the seven putative GntR genes, Rv0792c, was classified in the HutC subfamily of transcription regulators. In this subfamily the C-terminal domain contains both α-helical and β-strand structures and Rv0792c showed distinguishable predicted secondary structural features specific to this subfamily (Figure 2.5) **[**Rigali *et al*., 2002**]**. The C-terminal domain of HutC subfamily regulators adopts the same fold as chorismate lyases (*Escherichia coli UbiC*), hence it is named UbiC transcription regulator associated (UTRA) domain **[**Rigali *et al*., 2002**]**.



```
Rv0043c    Mtu   -------------GAFIERFDVATILEHHELDGLLNGIASARAAANPT--PRILGQLDAVMRSLRNSK-----
Rv0165c    Mtu   -------------GHVVLPLTRQDIDDIFWLQATIAQELATSATAHITDVEIDELDRINNALAGAIGSG----
MatR       Rle   -------------GAIVIDPGPHRVYEMFEVMAELEGLAGSLAARRL-DKTSREAITATHGRCEKSAAAG---
NtaR       Che   -------------GFFARVLEANTLFDLYELRAFLEQSAVRLACQRATDQEIAVLRDFLLEQDESGEI-----
VanR       Asp   -------------GYVVREISDELVHDALEVRGVLEGLAAKTIAEQGLTEQQKNILHGCIEETEKLFNGRNEF
Rv0494     Mtu   ---PEGLTHPAVVEALVRKLGPDFLVELLEIRAALGPLIGRLAAARS-TPEDAEALCAALEVVQQADTA----
Rv0586     Mtu   -RRHAGLDLLPRLLFRNGELDISVVRSILEARLRNFPKVAELAAERN-EPELAELLQDSLRALDTEEDP----
FadR       Eco   -----WETSGLNILETLARLDHESVPQLIDNLLSVRTNISTIFIRTAFRQHPDKAQEVLATANEVAD------
GlcC       Eco   RVARLNRVQDTSPLIHLFSTQPRTLYDLLDVRALLEGESARLAATLG-TQADFVVITRCYEKMLAASENNKEI
PipR       Msm   -----AGPPVEKLTARLAAMSASDLRDLFDEHTAIAGQAARLAAERAA--PSTVRRLFALTDQLDTAT-----
DgoR       Eco   RNQWNYLDTDVLQWVLENDYDPRLISAMSEVRNLVEPAIARWAAERA---TSSDLAQIESALNEMIANNQ---
PdhR       Eco   -----LWQSFSDPLVELLSDHPESQYDLLETRHALEGIAAYYAALRS-TDEDKERIRELHHAIELAQQSG---
LldR       Eco   -HDTWSEQNIVQPLKTLMADDPDYSFDILEARYAIEASTAWHAAMRA-TPGDKEKIQLCFEATLSE-------
Rv3060cI   Mtu   FICEPNAGPATRAVVIYLEYLGTTIGDLLGARLVLEPLAASLAAEHI-DEPGIERLRAVLRAEERWRPG----
Rv3060cII  Mtu   VVTTPQPQASIDTIALYLQYRKPSREDLRCVRDAIEIDNVAKVVKRRSEPEVASFLDTLGRPRLDNPTD----
Consensus        --------------------------#lle-r--le--aa-laa-r----e--e-------------------
```

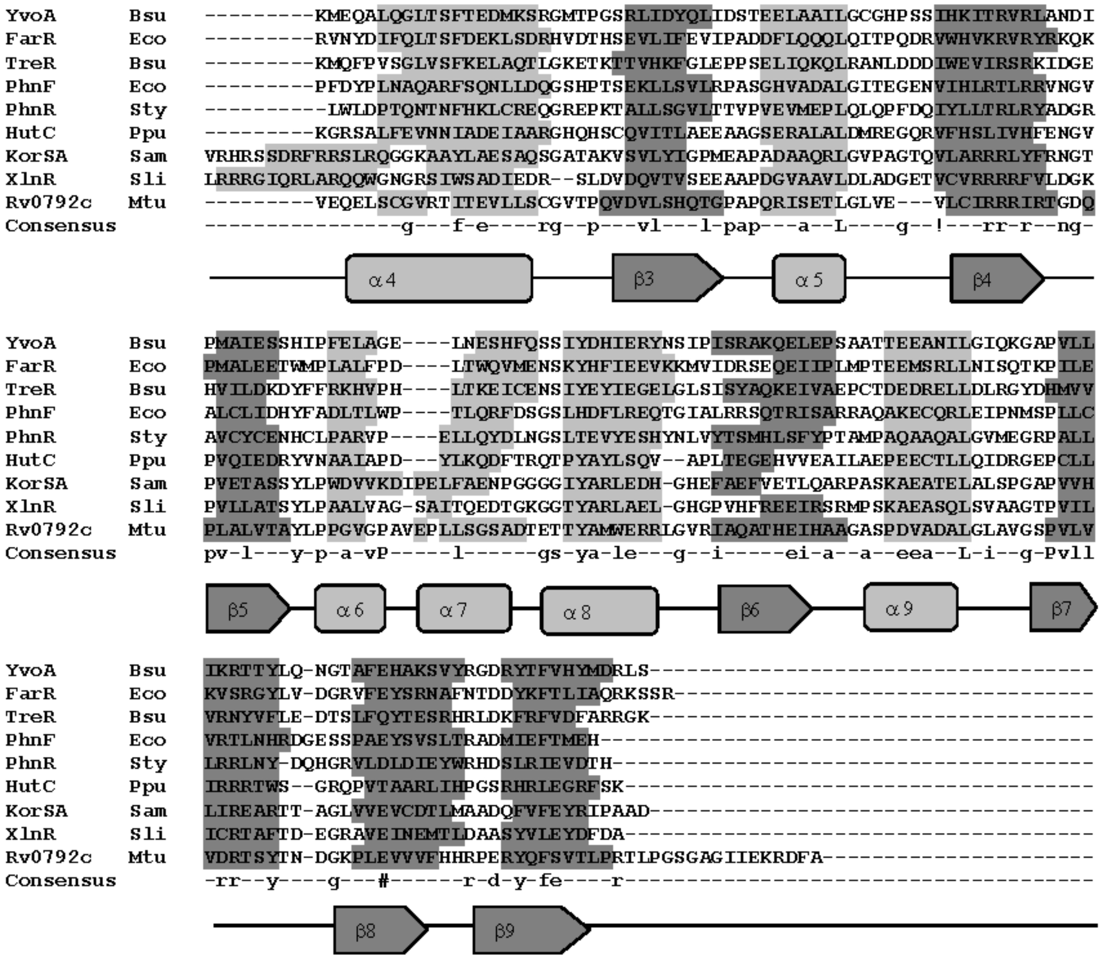
```
Rv0043c    Mtu   ---ESRAFAECVWEYRRTVNDEYAGPRLHATIRASQNLIPRVFWMTYQ-------NSRDDVLPFYEEENAAIH
Rv0165c    Mtu   ---DAKTIASIEFAFHRVFNKASRRIKLAWFLLNAARYMGAGVRG------------RPAMGRGRGEQSSAAD
MatR       Rle   ---DSDAYYYDNEEFHKAIYAAGRSDFLEEQCLQLHRRLRPDRRLQL-------RVRNRLSTSFLEHCAIVDA
NtaR       Che   ---SAGEMLKLDEEFHFRLVGLSQNEELLKTVRSISERIRFARWIDWQSR----------RLSHEQHLHITSL
VanR       Asp   GDEELEKYHHYNVIFHDTIIEGAKNIALMQALAKNNQLPLASAQAITYDQNRALSEYRRLHYAHLQHCSIYNA
Rv0494     Mtu   -----AARQAADLAYFRVLIHSTRNRALGLLYRWVEHAFGGREHALTGA-------YDDADPVLTDLRAINGA
Rv0586     Mtu   -----IVWQRHTLDFWDHVVDSAGSIVDRLMYNAFRAAYEPTLAALTTTMT-------AAAKRPSDYRKLADA
FadR       Eco   ---HADAFAELDYNIFRGLAFASGNPIYGLILNGMKGLYTRIGRHYFA-----NPEARSLALGFYHKLSALCS
GlcC       Eco   ---SLIEHAQLDHAFHLAICQASHNQVLVFTLQSLTDIMFNSVFASVNNLYH---RPQQKKQIDRQHARIYNA
PipR       Msm   ---SLRDRIRADSRFHIQVAVAAQSARLARREANLQAEVSGLIWLPIG-------PPIDVAAYVEEKHAISAA
DgoR       Eco   ---DREAFNEADIRYHEAVLQSVHNPVLQQLSIAISSLQRAVFERTWMG------DEANMPQTLQEHKALFDA
PdhR       Eco   ---DLDAESNAVLQYQIAVTEAAHNVVLLHLLRCMEPMLAQNVRQNFELL---YSRREMLPLVSSHRTRIFEA
LldR       Eco   ---DPDIASQADVRFHLAIAEASHNIVLLQTMRGFFDVLQSSVKHSRQRMY---LVPPVFSQLTEQHQAVIDA
Rv3060cI   Mtu   -------LPPPPEQFYRVLAEQSKNPVLQLFIDILMRLTKRYVQKSGTQSAG--EAVEAAGQVHNEHSDIVAA
Rv3060cII  Mtu   ---DVRAAAVEEFRFHVGLARAAGNTMLDLFLLLILVELFRRHLSSTEQAL----PTWSDVVAVGHAHVRILEA
Consensus        ------a----d--%h-----a--n--L--------------------------------------h--i--a
```

```
Rv0043c    Mtu   RREPEAARAACIGRSELMAQ-TMLAELFRRRVLVPPEGACPGPFGAPIPGFARSYQPSSPVP----------
Rv0165c    Mtu   RRAAPPRHSRRNRAHRLAVHRWGTQADGGPG-----------------------------------------
MatR       Rle   IFAGDGDEARRLLRGHVGIQGERFSDLVASMAAR--------------------------------------
NtaR       Che   LADRKEDECAAFVLAHIQKHFDQILEIIRGAVTEIYTRNSDSPRKAR-------------------------
VanR       Asp   LVNRQAGRAENLMREHSSVFVTRRDCTSDFSGCPSKKLLYMGIMYRRKTELLSKIND---------------
Rv0494     Mtu   VLAGDPAAAAATVEAYLNASALRMVKSYRDRA----------------------------------------
Rv0586     Mtu   ICSGDPTGAKKAAQDLLELANTSLMAVLVSQASRQ-------------------------------------
FadR       Eco   EGAHDQVYETVRRYGHESGEIWHRMQKNLPGDLAIQGR----------------------------------
GlcC       Eco   VLQRLPHVAQRAARDHVRTVKKNLHDIELEGHHLIRSAVPLEMNLS--------------------------
PipR       Msm   IAAENAQEARQLAEAHVMGQLARLTQINLDLTTKEAGR----------------------------------
DgoR       Eco   IRHQDGDAAEQAALTMIASSTRRLKEIT--------------------------------------------
PdhR       Eco   IMAGKPEEAREASHRHLAFIEEILLDRSREESRRERSLRRLEQRKN--------------------------
LldR       Eco   IFAGDADGARKAMMAHLSFVHTTMKRFDEDQARHARITRLPGEHNEHSREKNA-------------------
Rv3060cI   Mtu   VTAGDSAWAKTLSERHVEAVAGWLQQHQRGNDAAVRNGGRAREPRRAQQLIL--------------------
Rv3060cII  Mtu   IGSGDDSLARCRTRRHLDAAASWWL-----------------------------------------------
Consensus        i—agd---a------hl------l------------------------------------------------
```

Figure 2.4: Structure based sequence alignment of C-terminal domains of the classified transcriptional regulators from FadR subfamily. (Note: details of consensus symbol are given in method section 2.2.4; abbreviations for organism are given in table 2.1)

2.3.6 YtrA subfamily regulator of *M. tuberculosis*

The seventh putative GntR gene, Rv1152, was classified in the YtrA subfamily of transcriptional regulators. The YtrA subfamily of GntR is one of the least represented in bacterial genomes. YtrA subfamily regulator possesses a reduced C-terminal domain with only two α-helices. The average length of the putative E-b/O domain is about 50 amino acids. Only Rv1152 among the seven transcriptional regulators showed similar secondary structural features specific to the YtrA subfamily (Figure 2.6) **[**Rigali *et al*., 2002**]**. In *Bacillus subtilis* the YtrA regulator form a part of the large self-regulated operons that consist of genes encoding ATP binding cassette (ABC) transport systems in addition to YtrA **[**Yoshida *et al.,* 2000**]**. However, in *M. tuberculosis* the classified YtrA subfamily regulator, Rv1152, is not physically linked to genes encoding ABC transport systems or any other gene as a part of its operon.

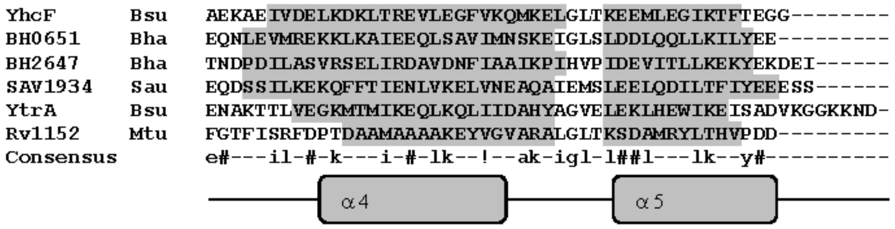


```
YvoA      Bsu   ---------KMEQALQGLTSFTEDMKSRGMTPGSRLIDYQLIDSTEELAAILGCGHPSSIHKITRVRLANDI
FarR      Eco   ---------RVNYDIFQLTSFDEKLSDRHVDTHSEVLIFEVIPADDFLQQQLQITPQDRVWHVKRVRYRKQK
TreR      Bsu   ---------KMQFPVSGLVSFKELAQTLGKETKTTVHKFGLEPPSELIQKQLRANLDDDIWEVIRSRKIDGE
PhnF      Eco   ---------PFDYPLNAQARFSQNLLDQGSHPTSEKLLSVLRPASGHVADALGITEGENVIHLRTLRRVNGV
PhnR      Sty   ----------LWLDPTQNTNFHKLCREQGREPKTALLSGVLTTVPVEVMEPLQLQPFDQIYLLTRLRYADGR
HutC      Ppu   ---------KGRSALFEVNNIADEIAARGHQHSCQVITLAEEAAGSERALALDMREGQRVFHSLIVHFENGV
KorSA     Sam   VRHRSSDRFRRSLRQGGKAAYLAESAQSGATAKVSVLYIGPMEAPADAAQRLGVPAGTQVLARRRLYFRNGT
XlnR      Sli   LRRRGIQRLARQQWGNGRSIWSADIEDR--SLDVDQVTVSEEAAPDGVAAVLDLADGETVCVRRRRFVLDGK
Rv0792c   Mtu   ---------VEQELSCGVRTITEVLLSCGVTPQVDVLSHQTGPAPQRISETLGLVE---VLCIRRRIRTGDQ
Consensus       ----------------g---f-e----rg--p---vl---l-pap---a--L----g--!---rr-r--ng-

YvoA      Bsu   PMAIESSHIPFELAGE----LNESHFQSSIYDHIERYNSIPISRAKQELEPSAATTEEANILGIQKGAPVLL
FarR      Eco   PMALEETWMPLALFPD----LTWQVMENSKYHFIEEVKKMVIDRSEQEIIPLMPTEEMSRLLNISQTKPILE
TreR      Bsu   HVILDKDYFFRKHVPH----LTKEICENSIYEYIEGELGLSISYAQKEIVAEPCTDEDRELLDLRGYDHMVV
PhnF      Eco   ALCLIDHYFADLTLWP----TLQRFDSGSLHDFLREQTGIALRRSQTRISARRAQAKECQRLEIPNMSPLLC
PhnR      Sty   AVCYCENHCLPARVP----ELLQYDLNGSLTEVYESHYNLVYTSMHLSFYPTAMPAQAAQALGVMEGRPALL
HutC      Ppu   PVQIEDRYVNAAIAPD---YLKQDFTRQTPYAYLSQV--APLTEGEHVVEAILAEPEECTLLQIDRGEPCLL
KorSA     Sam   PVETASSYLPWDVVKDIPELFAENPGGGGIYARLEDH-GHEFAEFVETLQARPASKAEATELALSPGAPVVH
XlnR      Sli   PVLLATSYLPAALVAG-SAITQEDTGKGGTYARLAEL-GHGPVHFREEIRSRMPSKAEASQLSVAAGTPVIL
Rv0792c   Mtu   PLALVTAYLPPGVGPAVEPLLSGSADTETTYAMWERRLGVRIAQATHEIHAAGASPDVADALGLAVGSPVLV
Consensus       pv-l---y-p-a-vP-----l------gs-ya-le---g--i-----ei-a--a--eea--L-i--g-Pvll

YvoA      Bsu   IKRTTYLQ-NGTAFEHAKSVYRGDRYTFVHYMDRLS------------------------------------
FarR      Eco   KVSRGYLV-DGRVFEYSRNAFNTDDYKFTLIAQRKSSR----------------------------------
TreR      Bsu   VRNYVFLE-DTSLFQYTESRHRLDKFRFVDFARRGK------------------------------------
PhnF      Eco   VRTLNHRDGESSPAEYSVSLTRADMIEFTMEH----------------------------------------
PhnR      Sty   LRRLNY-DQHGRVLDLDIEYWRHDSLRIEVDTH---------------------------------------
HutC      Ppu   IRRRTWS--GRQPVTAARLIHPGSRHRLEGRFSK--------------------------------------
KorSA     Sam   LIREARTT-AGLVVEVCDTLMAADQFVFEYRIPAAD------------------------------------
XlnR      Sli   ICRTAFTD-EGRAVEINEMTLDAASYVLEYDFDA--------------------------------------
Rv0792c   Mtu   VDRTSYTN-DGKPLEVVVFHHRPERYQFSVTLPRTLPGSGAGIIEKRDFA----------------------
Consensus       -rr--y---g---#------r-d-y-fe----r---------------------------------------
```

Figure 2.5: Structure based sequence alignment of C-terminal domains of the classified transcriptional regulators from HutC subfamily. (Note: details of consensus symbol are given in method section 2.2.4; abbreviations for organism are given in table 2.1)



```
YhcF        Bsu   AEKAEIVDELKDKLTREVLEGFVKQMKELGLTKEEMLEGIKTFTEGG--------
BH0651      Bha   EQNLEVMREKKLKAIEEQLSAVIMNSKEIGLSLDDLQQLLKILYEE---------
BH2647      Bha   TNDPDILASVRSELIRDAVDNFIAAIKPIHVPIDEVITLLKEKYEKDEI------
SAV1934     Sau   EQDSSILKEKQFFTIENLVKELVNEAQAIEMSLEELQDILTFIYEEESS------
YtrA        Bsu   ENAKTTLVEGKMTMIKEQLKQLIIDAHYAGVELEKLHEWIKEISADVKGGKKND-
Rv1152      Mtu   FGTFISRFDPTDAAMAAAAKEYVGVARALGLTKSDAMRYLTHVPDD---------
Consensus         e#---il-#-k---i-#-lk--!--ak-igl-l##l---lk--y#---------
```

Figure 2.6: Structure based sequence alignment of C-terminal domains of the classified transcriptional regulators from YtrA subfamily. (Note: details of consensus symbol are given in method section 2.2.4; abbreviations for organism are given in table 2.1)

2.4 CONCLUSION

This study aimed to understand the uncharacterized GntR family of transcriptional regulators from *M. tuberculosis*. All members of the GntR family of transcriptional regulators from *M. tuberculosis* were identified in the whole proteome. Using known and classified GntRs, these regulators were classified into functionally meaningful subfamilies. This classification was also supported with high bootstrapping values to all the classified subfamilies. The regulators illustrated consist of secondary structural features known to be associated to their respective subfamily. The sequence analyses of one of the FadR subfamily regulator, Rv3060c showed an internal gene duplication that possibly arose due to gene duplication followed by fusion. It is a first comprehensive sequence analyses of the GntR family of regulators that constitute the first step towards the understanding of operator site of GntR family regulators in *M. tuberculosis*. This extended protein subfamily classification, is an important step towards understanding the nucleotide preferences observed within operator sites for FadR, HutC and YtrA subfamily of transcriptional regulators that would be studied in the next chapter.

Chapter 3

Identification of GntR operator sites: In vitro validation for Rv0586 and Rv0792c

Part of the work is published as:

Vindal V, Kumar E A and Ranjan A (2008) Identification of operator sites within the upstream region of the putative mce2R from mycobacteria. FEBS Lett., 582:1117-1122

Vindal V, Kumar E A and Ranjan A (2008) Identification of a *cis*-acting site in the upstream sequence of a gene encoding a novel *Mycobacterium tuberculosis* transcriptional regulator Rv0792c (communicated).

3.1 INTRODUCTION

As the number of sequenced mycobacterial genomes has grown, employing comparative genomics has become more promising as a means to address operator site identification [Bailey *et al.,* 2006; Gelfand, 1999]. Nevertheless finding the operator site is one of the crucial steps towards characterization of any transcriptional regulator [Pritsker *et al.*, 2004]. In recent years much advancement has been made to identify these regulatory DNA sequences. In general, besides seeking conserved DNA sequences within the upstream DNA region of a regulator, acquiring clues from the features associated with family/subfamily is an effective approach towards identification of operator sites [Kaplan *et al.,* 2005]. In view of the available *M. tuberculosis* genome sequence, the GntR family of transcriptional regulators was classified into subfamilies (Chapter 2). Seven transcriptional regulators of GntR family were identified with features associated with FadR, HutC and YtrA subfamilies. The present chapter describes a combinatorial approach that was employed to understand the operator sites. Identification of operator sites of these transcriptional regulators helps in identifying the genes or operons regulated by these regulators. Conventionally these binding sites are determined using a labor-intensive DNAase1 footprinting technique.

Generally members of this family of transcriptional regulators are known to be auto-regulatory. Hence, besides regulating the expression of a set of genes in the genome, they regulate their own expression. Therefore these proteins interact with their upstream DNA sequences. This DNA-protein interaction depends upon precise recognition of the nucleotides of the target DNA by the amino acids of the DNA binding protein. In case of GntR family of regulators, it was observed that these regulators exhibit nucleotide preferences known to be associated with their specific subfamilies. Using these associated features the present chapter identifies potential operator sites for *M. tuberculosis* GntRs and their putative orthologs in the sequenced mycobacterial species, and the closely related non-mycobacterial species of *Nocardia farciana*. Furthermore, among the seven classified members of the GntR regulators, two novel regulators were subjected to experimental validations. These two novel regulators represent two major subfamilies of GntR. One of them, Rv0586, is reported to be associated with a mammalian cell entry operon (*mce2*). Mutational studies of *mce2* operon have demonstrated its critical role, where animals infected with a *M. tuberculosis* strain bearing mutant *mce2* operon displayed a delay in granuloma formation [Gioffre *et al.*, 2005]. Another regulator Rv0792c is a novel transcriptional regulator of HutC subfamily. Identification of operator sites of *M. tuberculosis* GntRs would be a major breakthrough in deciphering genome wide targets. This systematic study addresses many questions with respect to the operator site identification and the sequence conservation across the upstream region of orthologous genes. It also strengthens the nucleotide preferences exhibited by subfamilies of GntRs. In the present study, the operator sites for these novel transcriptional regulators have been explored for the first time and which have potential implications to understand the set of regulated genes or regulons specific to each transcriptional regulator.

3.2 MATERIALS

Chemicals and reagents used for cloning, expression, purification and electrophoretic mobility shift assay were obtained from several commercial sources.

3.2.1 Reagents used in DNA cloning

Oligonucleotides were procured from Integrated DNA Technology, Inc., Sigma Genosys and MWG. DNA extraction kits were procured form Qiagen (Chatsworth CA) as well as Eppendorf AG (Germany). T4 DNA ligase, dNTPs, restriction endonucleases and protein as well as DNA marker were purchased from New England Biolabs (NEB) (Beverly, MA). AccuTaq™ LA DNA Polymerase was procured from Sigma Inc.

3.2.2 Reagents used for protein expression, purification, and electrophoretic mobility shift assay

Chemicals including Sodium phosphate, Tris, NaCl, glycerol, Imidazole, $NiSO_4$, EDTA, DTT, IPTG, X-gal, and poly dIdC were procured from either Sigma Inc. or Amersham Pharmacia. T4 Polynucleotide Kinase was purchased from New England Biolabs (NEB) (Beverly, MA). Ni-NTA-agarose was purchased from Qiagen. DNA sequencing was performed by ABI Prism automated DNA sequencer. Radioisotope- $\gamma[^{32}P]$ATP was procured from Jonaki, CCMB India.

3.2.3 Recipes used for reagents

All media and buffers used in protein expression and purification were prepared in deionized double distilled water (DDW) with low conductivity. Composition of the media and antibiotics are as in Tables 3.1 and 3.2. All solutions were prepared by standard procedures as described in the *Molecular Cloning: A Laboratory Manual* [Sambrook *et al*., 1989].

Table 3.1: Composition of culture media

Medium	Composition (1L)
Luria-Bertani (LB)	10g bactotryptone + 5g yeast extract + 10g NaCl per liter in DDW. pH was adjusted to 7.2 with NaOH. The medium was sterilized by autoclaving
Terrific-Broth (TB)	A. 12g bactotryptone + 24g yeast extract + 4ml glycerol in 900ml DDW B. 100 ml of phosphate buffer [0.17M KH_2PO4 + 0.72M K_2HPO4] A and B were autoclaved separately and mixed later

Table 3.2: Composition of antibiotics

Antibiotics	Stock solution	Working concentration
Ampicillin	100mg/ml in DDW	100µg/ml
Kanamycin	30mg/ml in DDW	30µg/ml

Antibiotics prepared in double distilled water were filter sterilized by passing though a 0.22µm filter

Table 3.3: Composition of solutions used for agarose gel electrophoresis

Reagents	Composition
50X TAE	242g Tris base + 57.1ml of glacial acetic acid + 100ml of 0.5M EDTA per litre
6X sample loading dye	0.6% Orange-G in 30% glycerol
Ethidium bromide	Stock of 10mg/ml in DDW

Table 3.4: Composition of solutions used for SDS-PAGE

Reagents	Composition
30% acrylamide	29.2% acrylamide & 0.8% bis-acrylamide in DDW
4X separation	1.5M Tris-HCl, pH8.8 in 0.4% SDS
4X stacking gel buffer	1M Tris-HCl, pH 6.8 in 0.4% SDS
1X laemmli sample buffer	10% glycerol & 1% β-mercaptoethanol & 2%SDS & 0.1% bromophenol blue in 1X separating buffer
1X Running buffer	3g Tris + 14.4g glycine + 1g SDS per liter
Destaining solution	Methanol: Acetic acid: Water :: 5:1:4
Staining solution	0.1g/l of coomassie brilliant blue R250 in de-staining solution

Table 3.5: Composition of solutions used for Non-Denaturing PAGE

Reagents	Composition
30% acrylamide	29.5% acrylamide & 0.5% bis-acrylamide
5X TBE	54 g Tris-base + 27.5 Boric acid + 20 ml of 0.5M EDTA(pH 8.0)
6X gel loading dye	0.25% (w/v) Bromophenol blue in 40 % (w/v) sucrose in H_2O

3.3 EXPERIMENTAL PROCEDURE

3.3.1 Ortholog prediction and upstream sequence analyses

In the present study, best reciprocal blast hit method was used to predict putative orthologous proteins between each of the two proteomes with the BLASTP program at an E-value cut off of 10^{-6} [Altschul *et al*., 1990; Fulton *et al*., 2006]. DNA sequences were extracted that spanned 400 bp upstream and 50 bp downstream to the translation start site of the gene of interest. In general, GntR regulators are reported to recognize palindromes and also exhibit nucleotide recognition preferences characteristic to the subfamily **[**Rigali *et al*., 2002**]**. Besides these clues, conservation across the upstream region of more than three orthologous regulators was also considered [McCue *et al*., 2001]. Using clustalX and MULTALIN, a multiple sequence alignment of the upstream sequences was carried out **[**Thompson *et al*., 1997; Corpet 1988**]**.

3.3.2 Cloning expression and purification

The transcriptional regulators, Rv0586 and Rv0792c, were amplified by PCR using forward primer with a BamHI site and reverse primer having HindIII site (Table 3.6). Both the amplified DNA fragments were cloned into the expression vector, pQE30 (Ampicillin resistance), having an N-terminal 6x His tag. Recombinant clone was identified and checked by restriction digestion followed by DNA sequencing. Recombinant vector was transformed into a suitable host strain *E. coli* M15 (Kanamycin resistance). Transformed single colony was inoculated in 5ml of growth media (LB media for Rv0586 and TB for Rv0792c) containing appropriate antibiotics (starter culture) (Table 3.1 and 3.2). This starter culture was grown overnight at 37°C with vigorous shaking. 2 ml of the above grown culture was inoculated into 200 ml growth medium containing appropriate antibiotics. The culture was grown at 37°C until the OD reached 0.6 at A_{600nm}. Control culture was maintained in parallel. The cells were kept in an incubator shaker for another twelve hours at 18˚C and 200 rpm to allow protein expression induced at 0.5 mM IPTG concentration. Next, cells were harvested by centrifugation and resuspended in 10 ml of lysis buffer (50 mM NaH_2PO_4, 300 mM NaCl and 10 mM imidazole, pH 8.0) and 1 mM PMSF and disrupted using sonicator. After a second round of centrifugation for 20 minutes at 12000 rpm supernatant was collected. The supernatant was applied to a Ni–NTA affinity column (Qiagen, USA). Both the recombinant proteins were eluted with 200 mM imidazole and analyzed by SDS–PAGE after washing the column with 5 bed-volumes of wash buffer containing 20 mM imidazole (Table 3.4). Details of cloning are given in Table 3.6 to 3.8. Purity of different plasmid vectors used and constructed was assessed by agarose gel electrophoresis on a 1% agarose gel (Table 3.3).

Table 3.6: Details of cloning primers

Gene	Primer sequence [Forward primer-FP, Reverse primer-RP]
Rv0586	FP 5`CGCGGATCCATGGCGCTGCAGCCGGTGACTCG 3` RP 5`CCCAAGCTTTCATTGCCGACTCGCCTGGCTAAC 3`
Rv0792c	FP 5` CGCGGATCCATGACATCTGTCAAGCTGGAC 3` RP 5` CCCAAGCTTTCATGCGAAATCTCGTTTCTC 3`

Table 3.7: Optimized conditions for PCR amplification

Components	Rv0586	Rv0792c
Template	40 ng	50 ng
Forward primer	10 pmol	10 pmol
Reverse primer	10 pmol	10 pmol
dNTPs	200 µM	200 µM
$MgSO_4/MgCl_2$	2.5 mM	3.0 mM
DNA Polymerase	2 unit	2 unit

Table 3.8: Optimized PCR cycling parameters

Cycle	Rv0586	Rv0792c
Pre-denaturation	95°C, 5 min	95°C, 5 min
Denaturation/	95°C, 30 sec/	95°C, 30 sec/
Annealing /	55°C, 30 sec/	55°C, 30 sec/
Extension (30 cycles)	72°C, 1 min	72°C, 1 min
Final-extension	72°C, 10 min	72°C, 10 min

3.3.3 Electrophoretic mobility shift assay

To show binding with the predicted operator site, electrophoretic mobility shift assay (EMSA) was carried out using recombinant protein. An increasing amount of protein was incubated with 10 fmol of ^{32}P-labeled DNA motif at room temperature for 40 minutes in respective reaction mix (Table 3.5, Table 3.9 and Table 3.10) and loaded onto 5% non-denaturing polyacrylamide gel containing 0.5x Tris–borate–EDTA buffer. Samples were separated using electrophoresis at 200 V for 2 h. Subsequently, gel was dried and exposed on to a storage phosphor image plate. The image plate was scanned in the storage phosphor imaging workstation.

Table 3.9: Composition of reaction mixture used for EMSA

Protein	Binding buffer
Rv0586	10 mM Tris-HCl [pH 8.0], 1 mM DTT, 10 mM NaCl, 1 mM EDTA, 10 % glycerol, 10 µg of poly (dI–dC)/ml and 5 µg of bovine serum albumin per ml
Rv0792c	20mM Tris-HCl [pH 8.0], 1 mM DTT, 20 mM NaCl, 1 mM EDTA, 10 % glycerol, 10 µg of poly (dI–dC)/ml and 5 µg of bovine serum albumin per ml

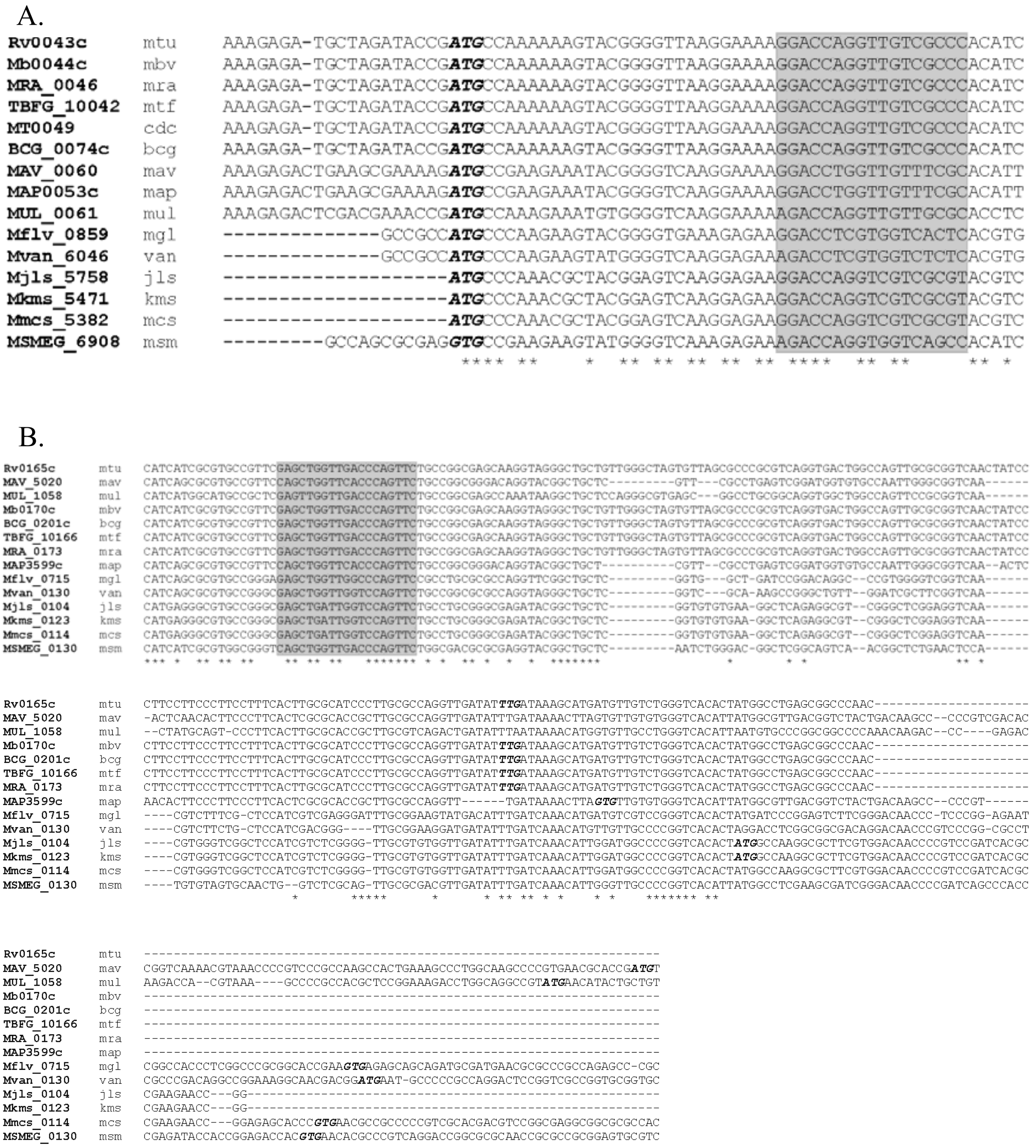
Table 3.10: List of DNA motifs used for DNA-protein interaction

Protein	Oligo sequence
Rv0586	Specific 5` GGTGTCGGTCTGACCACTTGA 3` Non-Specific 5`GTGAATGAAGATTGGTAAGAC 3`
Rv0792c	Specific 5` ATAAGACGTTTTAATACGTCTTAT 3` Non-Specific 5` CCCGGCTGCACCGCGCCACCGCGG 3`

3.4 RESULTS AND DISCUSSION

3.4.1 *In silico* identification of operator sites in the upstream regions of *M. tuberculosis* GntRs

In order to find upstream operator site for the classified *M. tuberculosis* GntR family regulators, a set of approaches have been employed. First, the most promising approach towards identifying these short DNA sequences was the comparative analyses of orthologous upstream regions for sequence conservation and second was to analyze nucleotide preferences within the conserved region as the characteristic of a particular subfamily. The second approach is an outcome of observations by Rigali and co-workers where members belonging to subfamilies of GntRs were shown to display nucleotide preferences in their DNA targets. Both the approaches in combination strengthen the validity of operator site identification. Using the best bi-directional blast hit method, putative orthologs of all seven GntRs in other mycobacterial species and the closest non-mycobacterial species *N. farciana* were identified. Further, upstream regions of these orthologs were aligned as per the method described in the methods section 3.3.1. Multiple sequence alignment of the upstream regions depicts the conserved DNA region(s) (Figure 3.1.A-G). Among these region(s), DNA sequences exhibiting known GntR nucleotide preferences were selected as a potential operator site (Table 3.11). Most of the operator sites were located upstream to the translational start site. All the selected potential operator sites were nearly palindromes.



A.

```
Rv0043c      mtu    AAAGAGA-TGCTAGATACCGATGCCAAAAAAGTACGGGGTTAAGGAAAAGGACCAGGTTGTCGCCCACATC
Mb0044c      mbv    AAAGAGA-TGCTAGATACCGATGCCAAAAAAGTACGGGGTTAAGGAAAAGGACCAGGTTGTCGCCCACATC
MRA_0046     mra    AAAGAGA-TGCTAGATACCGATGCCAAAAAAGTACGGGGTTAAGGAAAAGGACCAGGTTGTCGCCCACATC
TBFG_10042   mtf    AAAGAGA-TGCTAGATACCGATGCCAAAAAAGTACGGGGTTAAGGAAAAGGACCAGGTTGTCGCCCACATC
MT0049       cdc    AAAGAGA-TGCTAGATACCGATGCCAAAAAAGTACGGGGTTAAGGAAAAGGACCAGGTTGTCGCCCACATC
BCG_0074c    bcg    AAAGAGA-TGCTAGATACCGATGCCAAAAAAGTACGGGGTTAAGGAAAAGGACCAGGTTGTCGCCCACATC
MAV_0060     mav    AAAGAGACTGAAGCGAAAAGATGCCGAAGAAATACGGGGTCAAGGAAAAGGACCTGGTTGTTTCGCACATT
MAP0053c     map    AAAGAGACTGAAGCGAAAAGATGCCGAAGAAATACGGGGTCAAGGAAAAGGACCTGGTTGTTTCGCACATT
MUL_0061     mul    AAAGAGACTCGACGAAACCGATGCCAAAGAAATGTGGGGTCAAGGAAAAAGACCAGGTTGTTGCGCACCTC
Mflv_0859    mgl    --------------GCCGCCATGCCCAAGAAGTACGGGGTGAAAGAGAAGGACCTCGTGGTCACTCACGTG
Mvan_6046    van    --------------GCCGCCATGCCCAAGAAGTATGGGGTCAAGGAGAAAGACCTCGTGGTCTCTCACGTG
Mjls_5758    jls    --------------------ATGCCCAAACGCTACGGAGTCAAGGAGAAGGACCAGGTCGTCGCGTACGTC
Mkms_5471    kms    --------------------ATGCCCAAACGCTACGGAGTCAAGGAGAAGGACCAGGTCGTCGCGTACGTC
Mmcs_5382    mcs    --------------------ATGCCCAAACGCTACGGAGTCAAGGAGAAGGACCAGGTCGTCGCGTACGTC
MSMEG_6908   msm    ---------GCCAGCGCGAGGTGCCGAAGAAGTATGGGGTCAAAGAGAAAGACCAGGTGGTCAGCCACATC
                                        **** **    *  ** ** ** ** ** ****  ** **     ** *
```

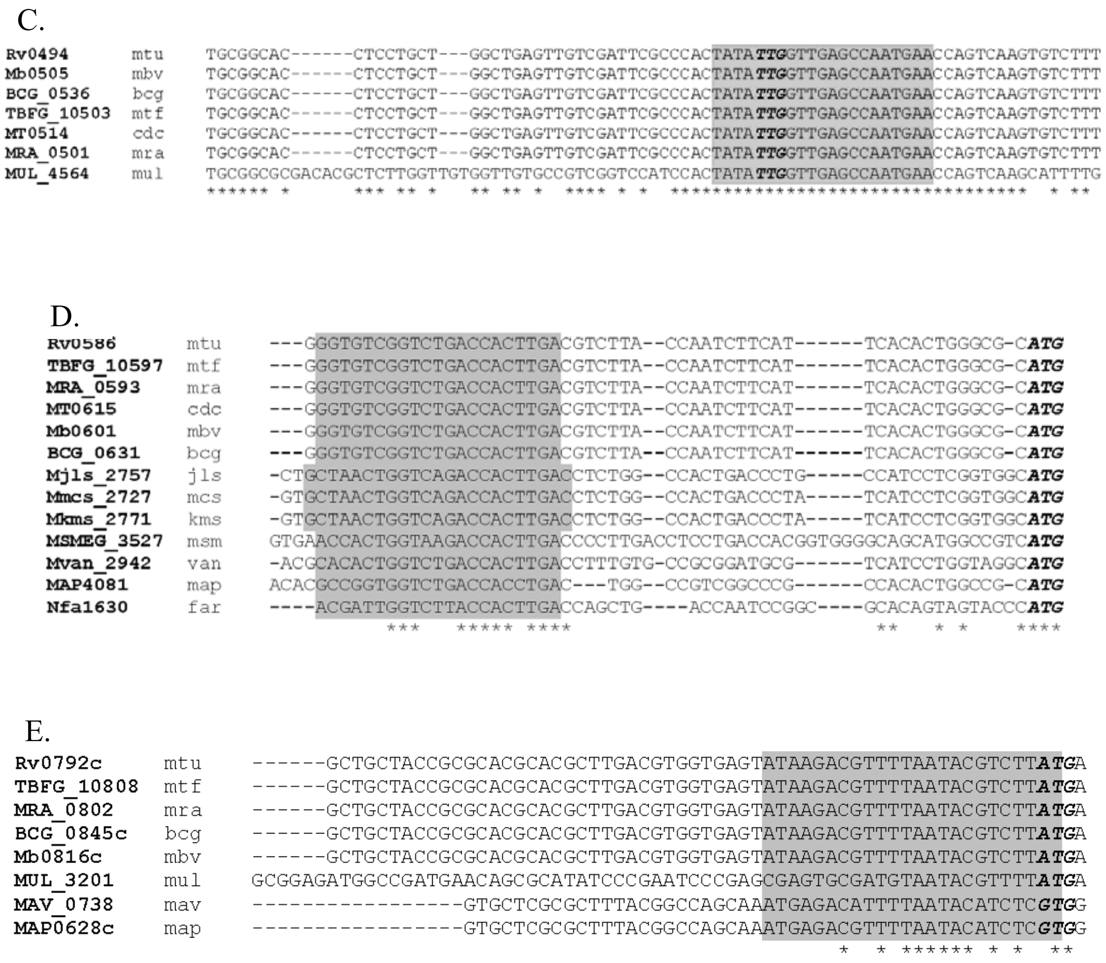
B.

```
Rv0165c      mtu    CATCATCGCGTGCCGTTCGAGCTGGTTGACCCAGTTCTGCCGGCGAGCAAGGTAGGGCTGCTGTTGGGCTAGTGTTAGCGCCCGCGTCAGGTGACTGGCCAGTTGCGCGGTCAACTATCC
MAV_5020     mav    CATCAGCGCGTGCCGTTCCAGCTGGTTCACCCAGTTCTGCCGGCGGGACAGGTACGGCTGCTC---------GTT---CGCCTGAGTCGGATGGTGTGCCAATTGGGCGGTCAA------
MUL_1058     mul    CATCATGGCATGCCGCTCGAGTTGGTTGACCCAGTTCTGCCGGCGAGCCAAATAAGGCTGCTCCAGGGCGTGAGC---GGCCTGCGGCAGGTGGCTGGCCAGTTCCGCGGTCAA------
Mb0170c      mbv    CATCATCGCGTGCCGTTCGAGCTGGTTGACCCAGTTCTGCCGGCGAGCAAGGTAGGGCTGCTGTTGGGCTAGTGTTAGCGCCCGCGTCAGGTGACTGGCCAGTTGCGCGGTCAACTATCC
BCG_0201c    bcg    CATCATCGCGTGCCGTTCGAGCTGGTTGACCCAGTTCTGCCGGCGAGCAAGGTAGGGCTGCTGTTGGGCTAGTGTTAGCGCCCGCGTCAGGTGACTGGCCAGTTGCGCGGTCAACTATCC
TBFG_10166   mtf    CATCATCGCGTGCCGTTCGAGCTGGTTGACCCAGTTCTGCCGGCGAGCAAGGTAGGGCTGCTGTTGGGCTAGTGTTAGCGCCCGCGTCAGGTGACTGGCCAGTTGCGCGGTCAACTATCC
MRA_0173     mra    CATCATCGCGTGCCGTTCGAGCTGGTTGACCCAGTTCTGCCGGCGAGCAAGGTAGGGCTGCTGTTGGGCTAGTGTTAGCGCCCGCGTCAGGTGACTGGCCAGTTGCGCGGTCAACTATCC
MAP3599c     map    CATCAGCGCGTGCCGTTCCAGCTGGTTCACCCAGTTCTGCCGGCGGGACAGGTACGGCTGCT----------CGTT--CGCCTGAGTCGGATGGTGTGCCAATTGGGCGGTCAA--ACTC
Mflv_0715    mgl    CATCAGCGCGTGCCGGGAGAGCTGGTTGGCCCAGTTCCGCCTGCGCGCCAGGTTCGGCTGCTC---------GGTG---GCT-GATCCGGACAGGC--CCGTGGGGTCGGTCAA------
Mvan_0130    van    CATCAGCGCGTGCCGGGCGAGCTGGTTGGTCCAGTTCTGCCGGCGCGCCAGGTAGGGCTGCTC---------GGTC---GCA-AAGCCGGGCTGTT--GGATCGCTTCGGTCAA------
Mjls_0104    jls    CATGAGGGCGTGCCGGGCGAGCTGATTGGTCCAGTTCTGCCTGCGGGCGAGATACGGCTGCTC---------GGTGTGTGAA-GGCTCAGAGGCGT--CGGGCTCGGAGGTCAA------
Mkms_0123    kms    CATGAGGGCGTGCCGGGCGAGCTGATTGGTCCAGTTCTGCCTGCGGGCGAGATACGGCTGCTC---------GGTGTGTGAA-GGCTCAGAGGCGT--CGGGCTCGGAGGTCAA------
Mmcs_0114    mcs    CATGAGGGCGTGCCGGGCGAGCTGATTGGTCCAGTTCTGCCTGCGGGCGAGATACGGCTGCTC---------GGTGTGTGAA-GGCTCAGAGGCGT--CGGGCTCGGAGGTCAA------
MSMEG_0130   msm    CATCATCGCGTGGCGGGTCAGCTGGTTGACCCAGTTCTGGCGACGCGCGAGGTACGGCTGCTC---------AATCTGGGAC-GGCTCGGCAGTCA--ACGGCTCTGAACTCCA------
                    *** *  ** ** **      ** ** **    ******* * *  ** *  *  * *******            *       * *                  ** *

Rv0165c      mtu    CTTCCTTCCCTTCCTTTCACTTGCGCATCCCTTGCGCCAGGTTGATATTTGATAAAGCATGATGTTGTCTGGGTCACACTATGGCCTGAGCGGCCCAAC--------------------
MAV_5020     mav    -ACTCAACACTTCCCTTCACTCGCGCACCGCTTGCGCCAGGTTGATATTTGATAAAACTTAGTGTTGTGTGGGTCACATTATGGCGTTGACGGTCTACTGACAAGCC--CCCGTCGACAC
MUL_1058     mul    --CTATGCAGT-CCCTTCACTTGCGCACCGCTTGCGTCAGACTGATATTTAATAAAACATGGTGTTGCCTGGGTCACATTAATGTGCCCGGCGGCCCAAACAAGAC--CC----GAGAC
Mb0170c      mbv    CTTCCTTCCCTTCCTTTCACTTGCGCATCCCTTGCGCCAGGTTGATATTTGATAAAGCATGATGTTGTCTGGGTCACACTATGGCCTGAGCGGCCCAAC--------------------
BCG_0201c    bcg    CTTCCTTCCCTTCCTTTCACTTGCGCATCCCTTGCGCCAGGTTGATATTTGATAAAGCATGATGTTGTCTGGGTCACACTATGGCCTGAGCGGCCCAAC--------------------
TBFG_10166   mtf    CTTCCTTCCCTTCCTTTCACTTGCGCATCCCTTGCGCCAGGTTGATATTTGATAAAGCATGATGTTGTCTGGGTCACACTATGGCCTGAGCGGCCCAAC--------------------
MRA_0173     mra    CTTCCTTCCCTTCCTTTCACTTGCGCATCCCTTGCGCCAGGTTGATATTTGATAAAGCATGATGTTGTCTGGGTCACACTATGGCCTGAGCGGCCCAAC--------------------
MAP3599c     map    AACACTTCCCTTCCCTTCACTCGCGCACCGCTTGCGCCAGGTT------TGATAAAACTTAGTGTTGTGTGGGTCACATTATGGCGTTGACGGTCTACTGACAAGCC--CCCGT------
Mflv_0715    mgl    ----CGTCTTTCG-CTCCATCGTCGAGGGATTTGCGGAAGTATGACATTTGATCAAACATGATGTCGTCCGGGTCACACTATGATCCCGGAGTCTTCGGGACAACCC-TCCCGG-AGAAT
Mvan_0130    van    ----CGTCTTCTG-CTCCATCGACGGG---TTGCGGAAGGATGATATTTGATCAAACATGTTGTTGCCCCGGTCACACTAGGACCTCGGCGGCGACAGGACAACCCGTCCCGG-CGCCT
Mjls_0104    jls    ----CGTGGGTCGGCTCCATCGTCTCGGGG-TTGCGTGTGGTTGATATTTGATCAAACATTGGATGGCCCCGGTCACACTATGGCCAAGGCGCTTCGTGGACAACCCCGTCCGATCACGC
Mkms_0123    kms    ----CGTGGGTCGGCTCCATCGTCTCGGGG-TTGCGTGTGGTTGATATTTGATCAAACATTGGATGGCCCCGGTCACACTATGGCCAAGGCGCTTCGTGGACAACCCCGTCCGATCACGC
Mmcs_0114    mcs    ----CGTGGGTCGGCTCCATCGTCTCGGGG-TTGCGTGTGGTTGATATTTGATCAAACATTGGATGGCCCCGGTCACACTATGGCCAAGGCGCTTCGTGGACAACCCCGTCCGATCACGC
MSMEG_0130   msm    ----TGTGTAGTGCAACTG--GTCTCGCAG-TTGCGCGACGTTGATATTTGATCAAACATTGGGTTGCCCCGGTCACATTATGGCCTCGAAGCGATCGGGACAACCCCGATCAGCCCACC
                                        *       *****       *      * ** ** * *    * *    ******* **

Rv0165c      mtu    ----------------------------------------------------------------------
MAV_5020     mav    CGGTCAAAACGTAAACCCCGTCCCGCCAAGCCACTGAAAGCCCTGGCAAGCCCCGTGAACGCACCGATGT
MUL_1058     mul    AAGACCA--CGTAAA----GCCCCGCCACGCTCCGGAAAGACCTGGCAGGCCGTATGAACATACTGCTGT
Mb0170c      mbv    ----------------------------------------------------------------------
BCG_0201c    bcg    ----------------------------------------------------------------------
TBFG_10166   mtf    ----------------------------------------------------------------------
MRA_0173     mra    ----------------------------------------------------------------------
MAP3599c     map    ----------------------------------------------------------------------
Mflv_0715    mgl    CGGCCACCCTCGGCCCGCGGCACCGAAGTGAGAGCAGCAGATGCGATGAACGCGCCCGCCAGAGCC-CGC
Mvan_0130    van    CGCCCGACAGGCCGGAAAGGCAACGACGGATGAAT-GCCCCCGCCAGGACTCCGGTCGCCGGTGCGGTGC
Mjls_0104    jls    CGAAGAACC---GG--------------------------------------------------------
Mkms_0123    kms    CGAAGAACC---GG--------------------------------------------------------
Mmcs_0114    mcs    CGAAGAACC---GGAGAGCACCCGTGAACGCCGCCCCGTCGCACGACGTCCGGCGAGGCGGCGCGCCAC
MSMEG_0130   msm    CGAGATACCACCGGAGACCACGTGAACACGCCCGTCAGGACCGGCGCGCAACCGCGCCGCGGAGTGCGTC
```

Figure 3.1: Diagram showing sequence alignment of the upstream sequences in relation to the translation start sites. Identified potential operator sites are highlighted in light gray background. Translational start sites in all the sequences are printed in bold. (A). Upstream DNA sequences of Rv0043c and its putative orthologs; (B) Upstream DNA sequences of Rv0165c and its putative orthologs; (C) Upstream DNA sequences of Rv0494 and its putative orthologs; (D) Upstream DNA sequences of Rv0586 and its putative orthologs; (E) Upstream DNA sequences of Rv0792c and its putative orthologs; (F) Upstream DNA sequences of Rv1152 and its putative orthologs; (G) Upstream DNA sequences of Rv3061c and its putative orthologs.



C.

```
Rv0494        mtu   TGCGGCAC------CTCCTGCT---GGCTGAGTTGTCGATTCGCCCACTATATTGGTTGAGCCAATGAACCAGTCAAGTGTCTTT
Mb0505        mbv   TGCGGCAC------CTCCTGCT---GGCTGAGTTGTCGATTCGCCCACTATATTGGTTGAGCCAATGAACCAGTCAAGTGTCTTT
BCG_0536      bcg   TGCGGCAC------CTCCTGCT---GGCTGAGTTGTCGATTCGCCCACTATATTGGTTGAGCCAATGAACCAGTCAAGTGTCTTT
TBFG_10503    mtf   TGCGGCAC------CTCCTGCT---GGCTGAGTTGTCGATTCGCCCACTATATTGGTTGAGCCAATGAACCAGTCAAGTGTCTTT
MT0514        cdc   TGCGGCAC------CTCCTGCT---GGCTGAGTTGTCGATTCGCCCACTATATTGGTTGAGCCAATGAACCAGTCAAGTGTCTTT
MRA_0501      mra   TGCGGCAC------CTCCTGCT---GGCTGAGTTGTCGATTCGCCCACTATATTGGTTGAGCCAATGAACCAGTCAAGTGTCTTT
MUL_4564      mul   TGCGGCGCGACACGCTCTTGGTTGTGGTTGTGCCGTCGGTCCATCCACTATATTGGTTGAGCCAATGAACCAGTCAAGCATTTTG
                    ****** *      *** ** *   ** ** *  **** * *  **********************************  * **
```

D.

```
Rv0586        mtu   ---GGGTGTCGGTCTGACCACTTGACGTCTTA--CCAATCTTCAT------TCACACTGGGCG-CATG
TBFG_10597    mtf   ---GGGTGTCGGTCTGACCACTTGACGTCTTA--CCAATCTTCAT------TCACACTGGGCG-CATG
MRA_0593      mra   ---GGGTGTCGGTCTGACCACTTGACGTCTTA--CCAATCTTCAT------TCACACTGGGCG-CATG
MT0615        cdc   ---GGGTGTCGGTCTGACCACTTGACGTCTTA--CCAATCTTCAT------TCACACTGGGCG-CATG
Mb0601        mbv   ---GGGTGTCGGTCTGACCACTTGACGTCTTA--CCAATCTTCAT------TCACACTGGGCG-CATG
BCG_0631      bcg   ---GGGTGTCGGTCTGACCACTTGACGTCTTA--CCAATCTTCAT------TCACACTGGGCG-CATG
Mjls_2757     jls   -CTGCTAACTGGTCAGACCACTTGACCTCTGG--CCACTGACCCTG-----CCATCCTCGGTGGCATG
Mmcs_2727     mcs   -GTGCTAACTGGTCAGACCACTTGACCTCTGG--CCACTGACCCTA-----TCATCCTCGGTGGCATG
Mkms_2771     kms   -GTGCTAACTGGTCAGACCACTTGACCTCTGG--CCACTGACCCTA-----TCATCCTCGGTGGCATG
MSMEG_3527    msm   GTGAACCACTGGTAAGACCACTTGACCCCTTGACCTCCTGACCACGGTGGGGCAGCATGGCCGTCATG
Mvan_2942     van   -ACGCACACTGGTCTGACCACTTGACCTTTGTG-CCGCGGATGCG------TCATCCTGGTAGGCATG
MAP4081       map   ACACGCCGGTGGTCTGACCACCTGAC---TGG--CCGTCGGCCCG------CCACACTGGCCG-CATG
Nfa1630       far   ----ACGATTGGTCTTACCACTTGACCAGCTG----ACCAATCCGGC----GCACAGTAGTACCCATG
                              ***   ***** ****                          **   * *    ****
```

E.

```
Rv0792c       mtu   ------GCTGCTACCGCGCACGCACGCTTGACGTGGTGAGTATAAGACGTTTTAATACGTCTTATGA
TBFG_10808    mtf   ------GCTGCTACCGCGCACGCACGCTTGACGTGGTGAGTATAAGACGTTTTAATACGTCTTATGA
MRA_0802      mra   ------GCTGCTACCGCGCACGCACGCTTGACGTGGTGAGTATAAGACGTTTTAATACGTCTTATGA
BCG_0845c     bcg   ------GCTGCTACCGCGCACGCACGCTTGACGTGGTGAGTATAAGACGTTTTAATACGTCTTATGA
Mb0816c       mbv   ------GCTGCTACCGCGCACGCACGCTTGACGTGGTGAGTATAAGACGTTTTAATACGTCTTATGA
MUL_3201      mul   GCGGAGATGGCCGATGAACAGCGCATATCCCGAATCCCGAGCGAGTGCGATGTAATACGTTTTATGA
MAV_0738      mav   -----------------GTGCTCGCGCTTTACGGCCAGCAAATGAGACATTTTAATACATCTCGTGG
MAP0628c      map   -----------------GTGCTCGCGCTTTACGGCCAGCAAATGAGACGTTTTAATACATCTCGTGG
                                                                   *  * ****** * *  **
```

F.

```
Rv1152       mtu   CAGCCAGCCTTGCGTGCTGGACAGCTCGTAAGGGTCGAATCGGGCCCACAATCCGTTCTTGTCATCGCGGAACGTCGGTACACCGCTTTCCGCGGAGATCCCCGCGCC
Mflv_2135    mgl   GCGCCAACCGTCGGCGCTGGAGATCTCGTACGGGTCGACTTTCGCCCACAGCCCCGTCTCGACGTCGCGGAACGTCGGGACGCCGCTCTCGGCGGAGATCCCGGCGCC
Mvan_4569    van   GCGCCAGCCGTCGGCGCTGGAGATCTCGTAGGGGTCGACCTTGGCCCACAGTCCGGTCTCGACGTCGCGGAACGTCGGCACACCGCTCTCGGCGGACATCCCCGCACC
Mjls_4290    jls   CCGCCATCCGTCGGCGCTGGAGATCTCGTACGGGTCGACCTTGGCCCACAGCCCGGTCTCGACGTCACGAAAGGTCGGCACCCCGCTCTCCGCGGAGATGCCGGCCCC
Mkms_4136    kms   CCGCCATCCGTCGGCGCTGGAGATCTCGTACGGATCGACCTTGGCCCACAGCCCGGTCTCGACGTCACGAAAGGTCGGCACCCCGCTCTCCGCGGAGATGCCGGCCCC
Mmcs_4061    mcs   CCGCCATCCGTCGGCGCTGGAGATCTCGTACGGATCGACCTTGGCCCACAGCCCGGTCTCGACGTCACGAAAGGTCGGCACCCCGCTCTCCGCGGAGATGCCGGCCCC
MAV_1290     mav   GCGCCACCCCTGGGTGCTGGACAGCTCGTAGGGGTCGAAGCGGGCCCACAGTCCGTTCTTGTCGTCGCGAAAGGTCGGCACCCCGCTCTCCGCGGAGATGCCCGCGCC
MAP2632c     map   GCGCCACCCCTGGGTGCTGGACAGCTCGTAGGGGTCGAAGCGGGCCCACAGTCCGTTCTTGTCGTCGCGGAAGGTCGGCACCCCGCTCTCCGCGGAGATGCCCGCGCC
Mb1183       mbv   CAGCCAGCCTTGCGTGCTGGACAGCTCGTAAGGGTCGAATCGGGCCCACAATCCGTTCTTGTCATCGCGGAACGTCGGTACACCGCTTTCCGCGGAGATCCCCGCGCC
BCG_1213     bcg   CAGCCAGCCTTGCGTGCTGGACAGCTCGTAAGGGTCGAATCGGGCCCACAATCCGTTCTTGTCATCGCGGAACGTCGGTACACCGCTTTCCGCGGAGATCCCCGCGCC
TBFG_11176   mtf   CAGCCAGCCTTGCGTGCTGGACAGCTCGTAAGGGTCGAATCGGGCCCACAATCCGTTCTTGTCATCGCGGAACGTCGGTACACCGCTTTCCGCGGAGATCCCCGCGCC
MRA_1162     mra   CAGCCAGCCTTGCGTGCTGGACAGCTCGTAAGGGTCGAATCGGGCCCACAATCCGTTCTTGTCATCGCGGAACGTCGGTACACCGCTTTCCGCGGAGATCCCCGCGCC
MT1186       cdc   CAGCCAGCCTTGCGTGCTGGACAGCTCGTAAGGGTCGAATCGGGCCCACAATCCGTTCTTGTCATCGCGGAACGTCGGTACACCGCTTTCCGCGGAGATCCCCGCGCC
MUL_0993     mul   TTCCCAGCCCTGTGTGCTGGACAGCTCGTACGGGTCGAAGCGGGCCCACAACCCGTTCTTGTCGTCACGGAACGTCGGTACTCCGCTTTCCGCGGAGATGCCCGCGCC
MSMEG_5174   msm   CTGCCAACCGTCGGTGCTGGAGATCTCGTACGGGTCGACCTGGGCCCACAGGCCGGTCTCGGCGTCGCGGAACGTGGGCACGCCGCTTTCTGCCGAGATCCCGGCGCC
                      *** ** *  * ****** * ****** ** ****     *******  **   *** * * ** ** ** ** ** ** ** ***** ** ** ** ** ** ** **
```

```
Rv1152       mtu   GCTGAGCACCGCCACTCGCAT-CCCACAAACATAGCTGTGCTTGGT--------------AGATACTGGGTA-------------------------CGTGGAGCTGC
Mflv_2135    mgl   GCTGAATACCGTCACCTGCAC-TCCACCAAGTTAGCGGCTGTAGGC---------------GATACTGGGGGC----------------------GTGGGGGATTTGG
Mvan_4569    van   GCTGAACACCGTCACCTGCACGCCCACCAAATTAGCCGCTTTGGGA---------------CATACTGGGCAC----------------------GTGGCCGAGTTGG
Mjls_4290    jls   GCTGAGCACCGTCACCTGCAC-ATGACCAATTTAGCCGGTGTGGGA---------------GATACTGGGCAC----------------------GTGGCCGGTTTGG
Mkms_4136    kms   GCTGAGCACCGTCACCTGCAC-ATGACCAATTTAGCCGGTGTGGGA---------------GATACTGGGCAC----------------------GTGGCCGGTTTGG
Mmcs_4061    mcs   GCTGAGCACCGTCACCTGCAC-ATGACCAATTTAGCCGGTGTGGGA---------------GATACTGGGCAC----------------------GTGGCCGGTTTGG
MAV_1290     mav   GCTGAGCACCGCTATGCGCAT-CCCACCAAGATAGCCGCGGCTGAG--------------CGATACTGAACG-------------------------AGTGGAGCTGG
MAP2632c     map   GCTGAGCACCGCTATGCGCAT-CCCACCAAGATAGCCGCGGCTGAG--------------CGATACTGAACG-------------------------AGTGGAGCTGG
Mb1183       mbv   GCTGAGCACCGCCACTCGCAT-CCCACAAACATAGCTGTGCTTGGT--------------AGATACTGGGTA-------------------------CGTGGAGCTGC
BCG_1213     bcg   GCTGAGCACCGCCACTCGCAT-CCCACAAACATAGCTGTGCTTGGT--------------AGATACTGGGTA-------------------------CGTGGAGCTGC
TBFG_11176   mtf   GCTGAGCACCGCCACTCGCAT-CCCACAAACATAGCTGTGCTTGGT--------------AGATACTGGGTA-------------------------CGTGGAGCTGC
MRA_1162     mra   GCTGAGCACCGCCACTCGCAT-CCCACAAACATAGCTGTGCTTGGT--------------AGATACTGGGTA-------------------------CGTGGAGCTGC
MT1186       cdc   GCTGAGCACCGCCACTCGCAT-CCCACAAACATAGCTGTGCTTGGT--------------AGATACTGGGTA-------------------------CGTGGAGCTGC
MUL_0993     mul   GCTGAGCACCGTCACTCGCAT-CCCACAAAGATAGCCGTGCTTGACC-------------GGATACTGGGCAATACTGGGTGA-----TACTGGGATTGTGGAGCTTC
MSMEG_5174   msm   GCTGAGCACAGTAACTTGCACGTCAACCAAGGTAGCGGGTCTTCGGGACGGCGGGTCAGGAGGACCTGGACGGATAAAGGAGCATACTCAAGGGTGTGACCGAATTGG
                   *****  ** *  *   ***      ** **  **** *                          ***                                 *   *
```

G.

```
Rv3060c      mtu   -------------------------------------GCCCTTGACTCAGCCAATACAGTACACTCTATTGAAATGAGCACCGAG
TBFG_13077   mtf   AAAATCTTGGATTTTGGGCGATTCTGCGTCTGCTCGCGCCCTTGACTCAGCCAATACAGTACACTCTATTGAAATGAGCACCGAG
BCG_3085c    bcg   AAAATCTTGGATTTTGGGCGATTCTGCGTCTGCTCGCGCCCTTGACTCAGCCAATACAGTACACTCTATTGAAATGAGCACCGAG
MRA_3092     mra   -------------------------------------GCCCTTGACTCAGCCAATACAGTACACTCTATTGAAATGAGCACCGAG
Mb3086c      mbv   -------------------------------------GCCCTTGACTCAGCCAATACAGTACACTCTATTGAAATGAGCACCGAG
MT3146       cdc   AAAATCTTGGATTTTGGCGATTCTGCGTCTGCTCGCGCCCTTGACTCAGCCAATACAGTACAC---------------------
                                                        **************************
```

Table 3.11: List of identified potential operator sites for *M. tuberculosis* GntRs

Gene	Potential operator site
Rv0043c	GGACCAGGTTGTCGCCC
Rv0165c	GAGCTGGTTGACCCAGTTC
Rv0494	TATATTGGTTGAGCCAATGAA
Rv0586	GGTGTCGGTCTGACCACTTGA
Rv0792c	ATAAGACGTTTTAATACGTCTTAT
Rv1152	CCTTGCGTGCTGGACAGCTCGTAAGG
Rv3060c	CCTTGACTCAGCCAATA

3.4.2 Experimental validation of operator sites

Analyses of the upstream regions of all the classified *M. tuberculosis* GntRs has revealed potential operator sites. To strengthen the computational predictions, two of the transcriptional regulators were subjected to experimental validation as an *in vitro* model system.

3.4.2.1 *In vitro* validation of identified operator site specific to Rv0586

The ORF encoding Rv0586 was amplified, cloned and expressed in the bacterial expression system (Experimental procedure section 3.3.2). Recombinant protein was purified using metal affinity chromatography (Figure 3.2 and Figure 3.3). Purified protein was subjected to electrophoretic mobility shift assay with the identified operator sites. The DNA-protein interaction in Figure 3.4 showed clear binding with increasing concentration of protein to the synthesized double stranded DNA motif corresponding to the operator site sequence (Table 3.11). This binding was abolished gradually with increasing molar concentration of the unlabeled DNA as a specific competitor (10x, 25x and 50x molar excess), but same fold molar excess of the non-specific DNA did not affect the DNA protein complex (Figure 3.4). It clearly demonstrates that Rv0586 binds specifically to the operator site. This non-specific DNA was also chosen from the upstream region of Rv0586.

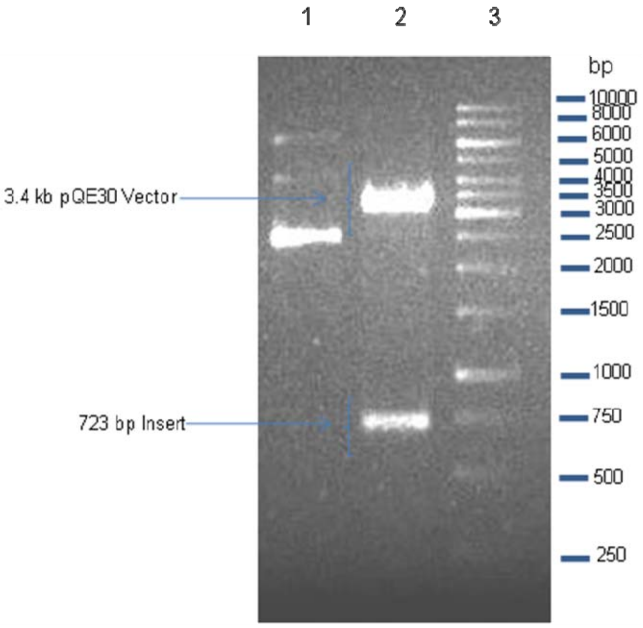


Figure 3.2 Agarose gel showing the cloned ORF Rv0586 in the expression vector pQE30. Lane no. 1 shows recombinant plasmid, lane 2 shows 723 bp fallout of Rv0586 gene digested with BamHI and HindIII and lane 3 shows the 1Kb DNA ladders.

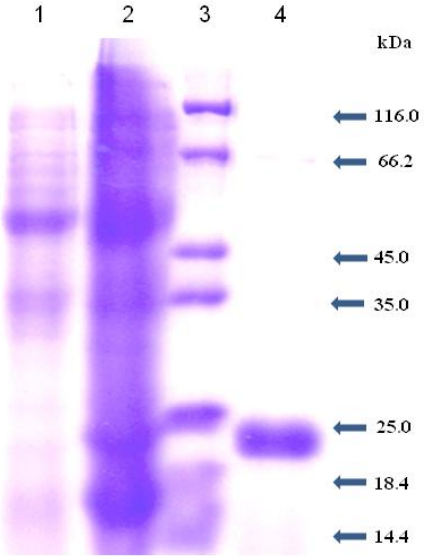


Figure 3.3 SDS-PAGE scan showing the expressed and purified Rv0586. Lane 1, IPTG induced *E. coli* M15 cell lysate harboring pQE30 as a control; lane 2, IPTG induced *E. coli* M15 cell lysate harboring recombinant pQE30 vector cloned with ORF Rv0586; lane 3, protein marker; lane 4, Rv0586 purified protein. All samples were loaded on 12% SDS–PAGE followed by Coomassie blue staining.

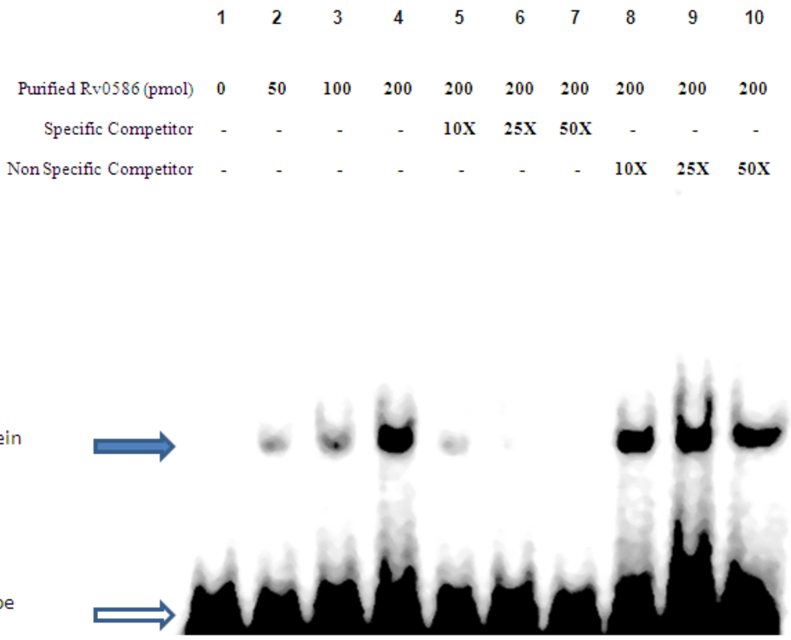

1 2 3 4 5 6 7 8 9 10
Purified Rv0586 (pmol) 0 50 100 200 200 200 200 200 200 200
Specific Competitor - - - - 10X 25X 50X - - -
Non Specific Competitor - - - - - - - 10X 25X 50X
DNA-protein complex
Free Probe

Figure 3.4 Binding of the Rv0586 protein to the identified operator DNA from the upstream region of Rv0586. Lane 1, labeled fragment; lanes 2–4, labeled fragment with 50, 100, 200 pmol of purified Rv0586 protein per reaction (30µl); lanes 5–7 contain an increasing amount of cold specific dsDNA oligonucleotide competitor (10-, 25- and 50-fold molar excess); lanes 8–10 contain an increasing amount of cold non-specific competitor (10-, 25- and 50-fold molar excess). The positions of DNA–protein complex and free probe are shown with solid and open arrows, respectively.

3.4.2.2 *In vitro* validation of identified operator site specific to Rv0792c

Rv0792c was another protein chosen for *in vitro* model experimental validation. It is a novel transcriptional regulator classified as a HutC subfamily of transcriptional regulator. The ORF, encoding protein Rv0792c, was amplified using PCR and subsequently cloned and expressed in *E. coli* expression system (Figure 3.5 and Figure 3.6). The recombinant protein was purified using metal affinity chromatography. Recombinant protein was subjected to electrophoretic mobility shift assay to validate the identified operator site within the upstream region. Increasing concentration of purified protein showed clear binding with the synthesized double stranded DNA motif corresponding to the identified operator site (Table 3.11). This binding was abolished gradually with increasing molar concentration of the unlabeled DNA used as a specific competitor (10x, 25x and 50x molar excess), but same fold excess of the non-specific DNA did not affect the labeled DNA protein complex (Figure 3.7).

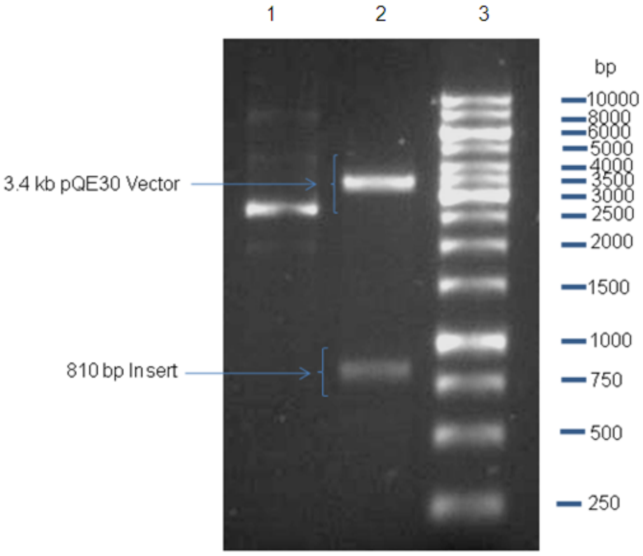


Figure 3.5 Agarose gel showing the cloned ORF Rv0792c in the expression vector pQE30. Lane 1 shows recombinant plasmid, lane 2 shows 810 bp fallout of Rv0792c gene digested with BamHI and HindIII and lane 3 shows the 1Kb DNA ladders.

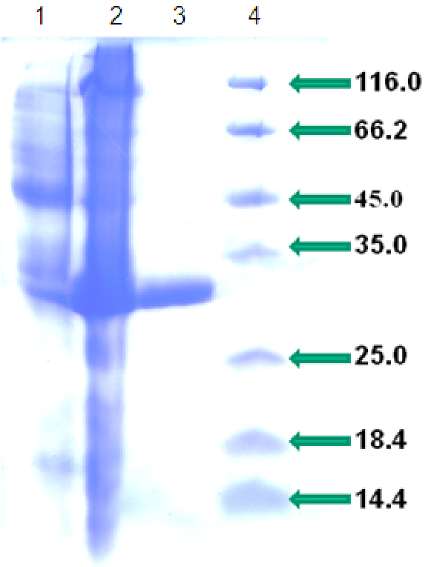


Figure 3.6 SDS-PAGE scan showing the expressed and purified Rv0792c. Lane 1, IPTG induced *E. coli* M15 cell lysate harboring pQE30 as a control; lane 2, IPTG induced *E. coli* M15 cell lysate harboring recombinant pQE30 vector cloned with ORF Rv0792c; lane 3, Rv0792c purified protein; lane 4, protein marker. All samples were loaded on 12% SDS–PAGE followed by Coomassie blue staining.

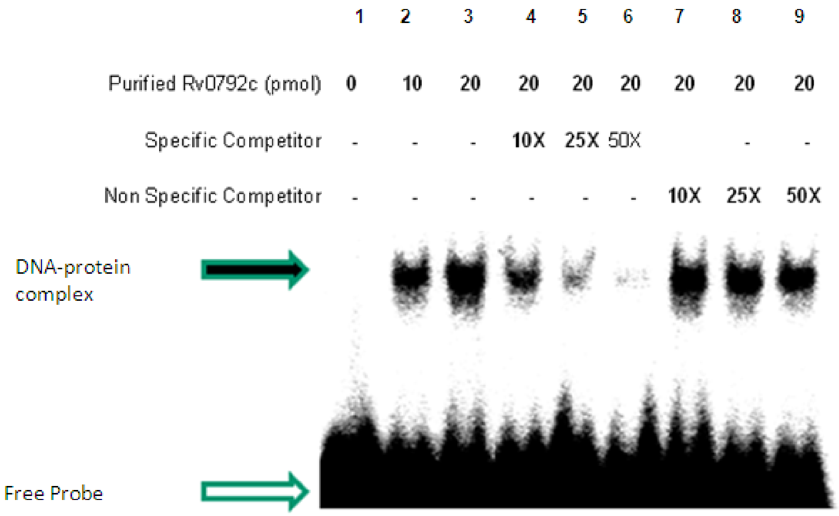

1
2
3
4
5
6
7
8
9
Purified Rv0792c (pmol) 0 10 20 20 20 20 20 20 20
Specific Competitor - - - 10X 25X 50X - -
Non Specific Competitor - - - - - - 10X 25X 50X
DNA-protein complex
Free Probe

Figure 3.7 Binding of the Rv0792c protein to the identified operator DNA from the upstream region of Rv0792c. Lane 1, labeled fragment; lanes 2–3, labeled fragment with 10, 20 pmol of purified Rv0792c protein per reaction (30µl); lanes 4–6 contain an increasing amount of cold specific dsDNA oligonucleotide competitor (10-, 25- and 50-fold molar excess); lanes 7–9 contain an increasing amount of cold non-specific competitor (10-, 25- and 50-fold molar excess). The positions of DNA–protein complex and free probe are shown with solid and open arrows, respectively.

3.5 CONCLUSION

Present chapter makes use of the information achieved from the classification of the transcriptional regulator carried out in the previous chapter. The study has been extended further to understand operator site features associated with specific subfamilies of GntR. It systematically describes the set of DNA sequences that are likely to serve as operator sites for Rv0043c, Rv0165c, Rv0494, Rv0586, Rv0792c, Rv1152 and Rv3060c. Additionally, operator sites were identified for some of the putative orthologs of these regulators across the mycobacterial species and closely related non-mycobacterial species like *Nocardia farciana.* These identified DNA targets could serve as initiating points towards further characterization of these transcriptional regulators in mycobacteria.

Among all classified GntR regulators, DNA-protein interaction of two of the regulators Rv0586 and Rv0792c was evaluated as proof of concept to identify the operator sites. These two regulators represent two major subfamilies of the GntR family of regulators. Both the regulators were subjected to experimental validation using electrophoretic mobility shift assay. This assay demonstrated that identified operator sites can indeed bind to the specific transcriptional regulator. This analysis has also provided valuable insights into the general architecture of the operator site for the respective subfamilies of GntRs and may prove useful in elucidating DNA targets of other regulators in the genome.

Chapter 4

Operator site recognition: Insight from DNA binding domain

Part of the work is published as:

Vindal V, Kumar E A and Ranjan A (2008) Identification of operator sites within the upstream region of the putative mce2R from mycobacteria. FEBS Lett., 582:1117-1122.

4.1 INTRODUCTION

Transcriptional regulators have a central role in regulating the transcriptional profile of the cell in various conditions. They are part of molecular switches which modulate the RNA polymerase activity in response to the environmental signal. They bind to specific sequences of DNA and thereby control the transfer of genetic information from DNA to RNA. The DNA recognition by any transcriptional regulator depends upon the precise interaction between the amino acids of the DNA binding domain and the nucleotides in the target DNA. It is observed that besides possessing a similar DNA binding domain, amino acid residues that interact with the DNA backbone are usually conserved across the members of a protein family [Luscombe and Thornton, 2002]. Therefore, DNA-protein interaction may be better understood in the context of individual protein family [Kaplan *et al.*, 2005].

Last decade has witnessed a great influx of high-quality crystal structures of DNA-binding proteins in the Protein Data Bank. These structures, especially those having DNA as part of complexes, have provided invaluable insights into the molecular basis of DNA-protein interaction. However, crystal structures of none of the *M. tuberculosis* GntR family proteins have been reported yet. Among GntR family regulators, *E. coli* FadR is well characterized and the structure of its co-crystallized DNA-protein complex is also available [van Aalten *et al.*, 2001; Xu *et al.*, 2001]. Using molecular modeling studies, it is possible to obtain structural insights from this structure.

The present chapter explores the sequence conservation in one of the classified GntR regulator, Rv0586, and its putative orthologs. Using molecular modeling, the structural insight into DNA-protein interaction was obtained. The model was used systematically to identify the amino acid residues that are likely to be involved in DNA-protein interaction. Attempts were made to analyze the conservation of these amino acids across the orthologs. Further, *in vitro* assays were also carried out to address the binding ability of the regulator towards a set of identified operator sites in the upstream regions of the orthologs.

4.2 EXPERIMENTAL PROCEDURE

4.2.1 Molecular modeling

Three-dimensional (3D) model of Rv0586 was built by comparative protein structure modeling using the program MODELLER 9v5 [Sali and Blundell, 1993]. For this, the input consisted of the template structure and the sequence alignment of the target sequence with the sequence of template structure. The closest suitable template structure was of *E. coli* FadR-DNA complex [PDB code: 1H9T]. Further, the assessment of the Ramachandran plot of the model was carried out using a program RAMPAGE [Lovell *et al.*, 2003]. The structures were visualized and analyzed using PYMOL and SPDBviewer.

4.2.2 Consensus logo and sequence alignment

To identify the frequencies of each nucleotide within the set of DNA sequences, consensus sequence logo was generated using the web server at http://weblogo.berkeley.edu/logo.cgi. [Crooks *et al.*, 2004]. Sequence alignment was carried out as per the method explained in Chapter 2, Section 2.2.4.

4.2.3 Electrophoretic mobility shift assay

This assay was carried out as per the method described in Chapter 3, Section 3.3.3.

4.3 RESULTS AND DISCUSSION

Conservation in DNA binding domain (DBD) is the key to recognize the similar DNA targets within a family of transcriptional regulators. In order to address systematically the conservation of DBD of GntR family of regulators and its influence towards the recognition of DNA targets, one of the classified GntR regulators, Rv0586 was studied as an example.

4.3.1 Conservation of DNA binding domain

To study the conservation in case of Rv0586 and its orthologs, multiple sequence alignment of all the protein sequences was carried out according to the method described in Chapter 2, Section 2.2.4. It is clear from Figure 4.1 that the N-terminal region, known as the DNA binding domain, is conserved in comparison to the C-terminal ligand-binding domain in all the proteins. This sequence conservation in DBD gave an interesting possibility that Rv0586 could recognize target DNA motifs specific to the orthologs in closely related species as well [Yellaboina *et al*., 2006]. Moreover, these orthologs showed conservation in the pattern of secondary structural elements known for FadR subfamily of transcriptional regulators (Figure 4.1) [Rigali *et al*., 2002].

4.3.2 Rv0586-DNA interaction: Structural insight

Generally in DNA binding protein, amino acid residues in contact with the DNA are more conserved than the rest of the residues [Luscombe and Thornton, 2002]. Therefore, in order to identify the amino acids that play a vital role in the formation of DNA-protein complex, investigation of the amino acids that are in close proximity with the DNA would be useful. Thus, although Rv0586 protein-DNA complex structure is not determined, molecular modeling with the closest structural homolog, *E.coli* FadR-DNA complex, would throw light on the amino acids that play a major role in forming DNA-protein complex.

4.3.2.1 Homology model of Rv0586

To obtain structural insight of DNA-protein interaction comparative protein structure modeling was carried out. Within GntR family, FadR, a well-studied transcriptional regulator from *E. coli,* was observed to be the closest available structural homolog of Rv0586. It is a well-known protein of the same subfamily involved in the regulation of fatty acid metabolism. Besides native protein structure, DNA-protein complex of this protein has also been determined at a resolution of 3.25 Å. Structure of this complex possesses two-protein monomers and the 19mer double stranded DNA chain (CATCTGGTACGACCAGATC) [van Aalten *et al*., 2001]. A model for Rv0586 was also built having protein dimer with the same double stranded DNA sequence (Figure 4.2). Further the model was evaluated using program RAMPAGE. It was observed that in the model that in the main chain conformations, 92.2% of amino acid residues were within the favored region, 5.2 % of amino acid residues were in allowed region of the Ramachandran plot (Figure 4.3), and only twelve amino acids (2.6%) were found in outlier region. Homology model of this protein with *E. coli* was obtained at a low RMS deviation (0.38Å over 452 structurally equivalent atoms).

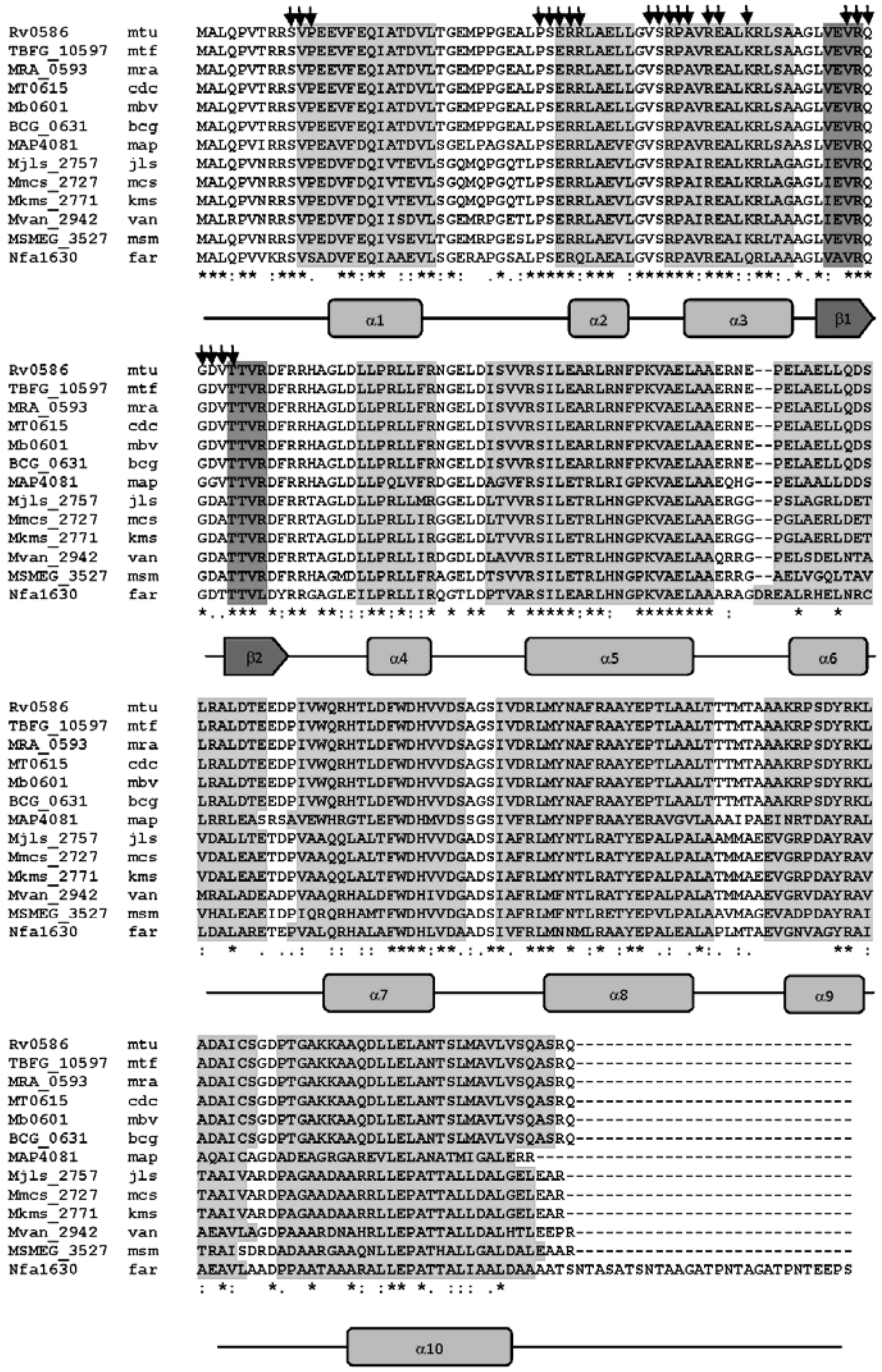


Figure 4.1: Multiple sequence alignment of Rv0586 and its orthologous proteins from mycobacteria and *N. farciana*. The graphical representations of α-helix and β-strand regions are highlighted with light and dark gray background, respectively. Amino acids, observed in the close vicinity (4Å) of DNA chain in the model of DNA-Protein complex are marked with arrow symbol. (Note: abbreviations: *mtu – M. tuberculosis; mtf – M. tuberculosis F11; mra - M. tuberculosis H37Ra; cdc - M. tuberculosis CDC1551; mbv – M. bovis; bcg – M. bovis BCG Pasteur 1173P2; map – M. avium subsp. paratuberculosis; jls – M. sp. JLS; mcs – M. sp. MCS; kms – M. sp KMS; van – M. vanbaalenii PYR; msm – M. smegmatis; far – N. farciana*).

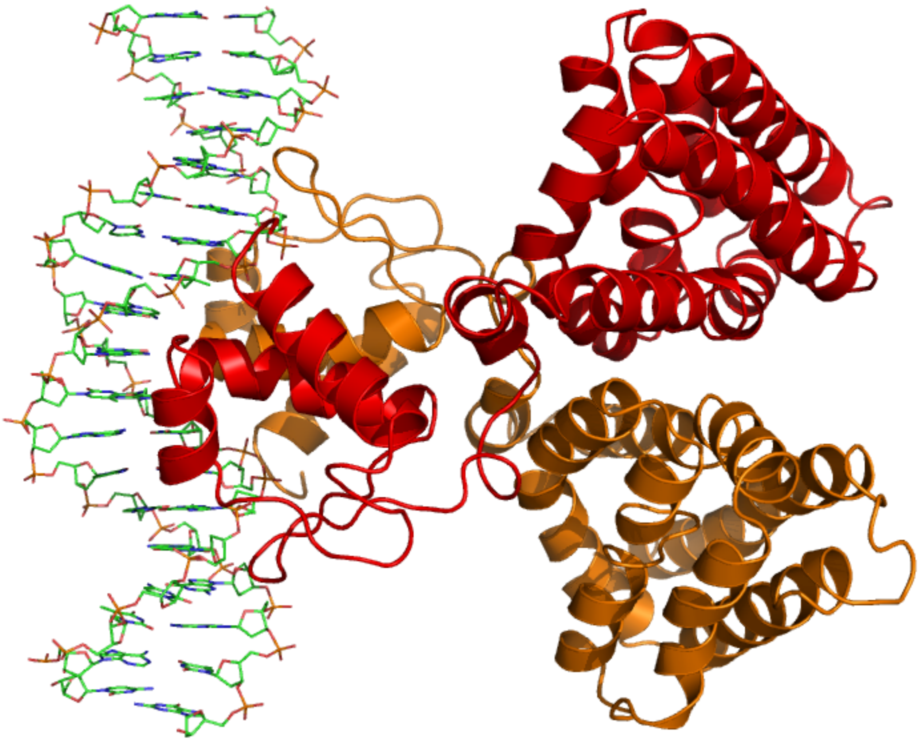


Figure 4.2: Model of Rv0586 dimer with DNA chain. The two monomers are highlighted in red and orange colour. The polypeptide is shown in ribbon and DNA is shown in stick representation.

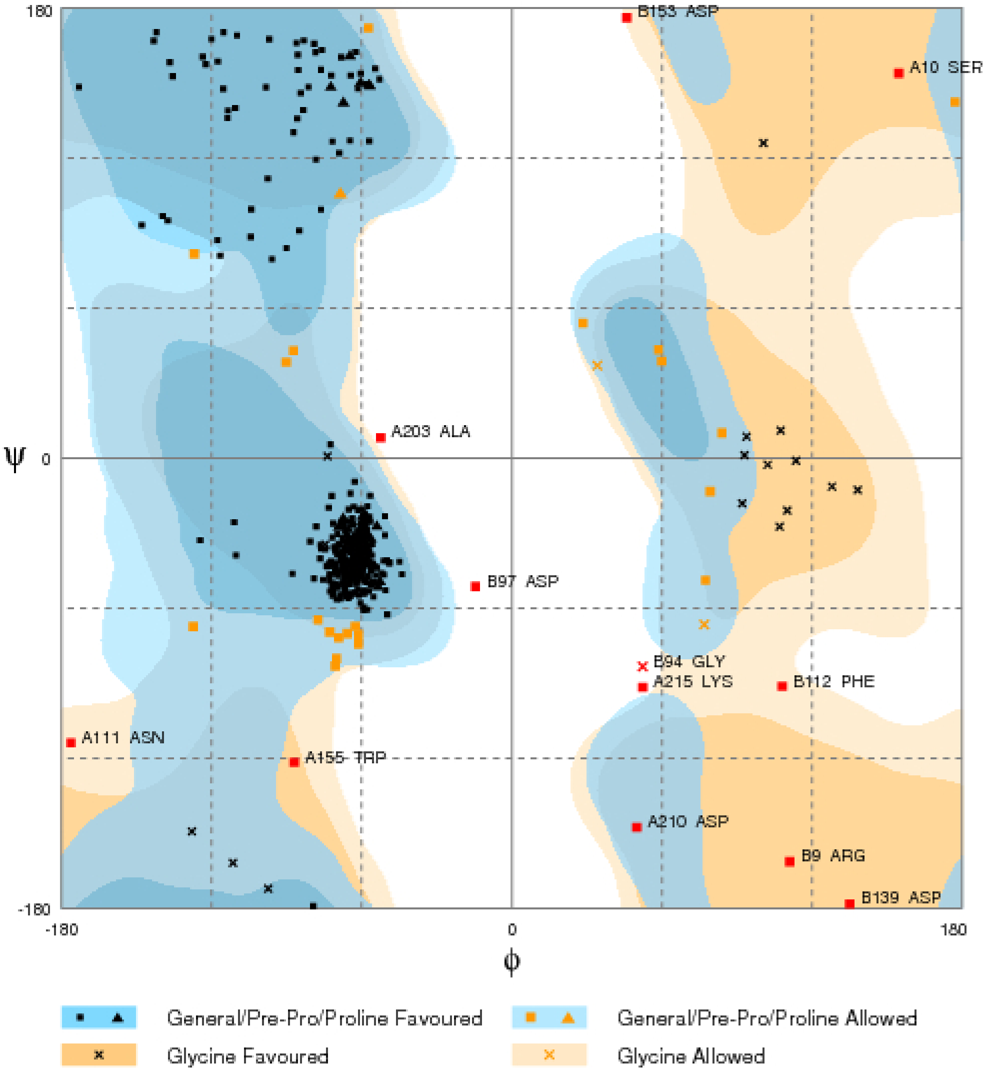


Figure 4.3: The Ramachandran plot for the proposed model of Rv0586.

4.3.2.2 Critical residue conservation across the Rv0586 and its orthologs

In view of the importance of critical residues in the operator site recognition, amino acid residues were selected which were within 4Å radius from the DNA chain in the Rv0586 model (Figure 4.4). All these residues were analyzed for their conservation across the orthologs of Rv0586. Most of the residues were observed to be conserved in all the identified orthologs of Rv0586 and these residues are marked with an arrow in Figure 4.1.

4.3.3 Consensus of DNA targets for the orthologs of Rv0586

Besides amino acid conservation, base-specific contacts within DNA targets also play major role to DNA-protein interaction [Siggers *et al.*, 2005]. Therefore, consensus logo of the identified operator sites was drawn (Figure 4.5). Generated sequence logos provided more precise description of sequence similarity and the conservation of nucleotide within the set of DNA targets. It displayed the sequence conservation among the identified DNA targets of the Rv0586. This nucleotide conservation was logical and in line with amino acid conservation across the orthologs, which are likely to play an important role in DNA contact. The heights of each stack among the logos clearly indicated preferences in terms of nucleotide, having higher frequencies, within the operator sites. The nucleotide positions 4–9 in the consensus are nearly an inverted palindrome of positions 13–18 (Figure 4.5).

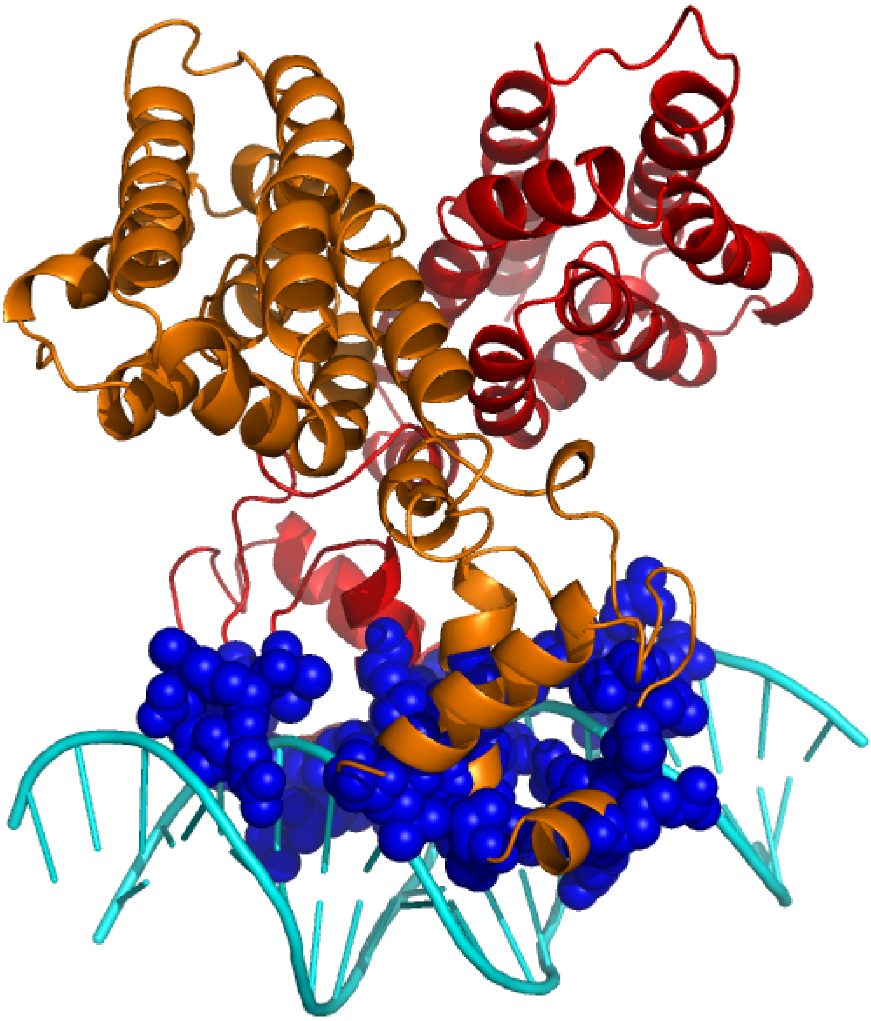


Figure 4.4: Amino acid residues of Rv0586 in the close proximity of DNA chain. The two-polypeptide backbones are shown in Orange and Red colours. The DNA is shown in cyan. Amino acid residues that are in close contact (within 4Å) with DNA molecule are shown as spheres in blue colour.

Table 4.1: List of potential operator sites identified for *M. tuberculosis* Rv0586 and its closest orthologs.

Gene/ORF	Organism	Potential operator site
Rv0586	*M. tuberculosis*	GGTGTCGGTCTGACCACTTGA
TBFG_10597	*M. tuberculosis F11*	GGTGTCGGTCTGACCACTTGA
MRA_0593	*M. tuberculosis H37Ra*	GGTGTCGGTCTGACCACTTGA
MT0615	*M. tuberculosis CDC1551*	GGTGTCGGTCTGACCACTTGA
Mb0601	*M. bovis*	GGTGTCGGTCTGACCACTTGA
BCG_0631	*M. bovis BCG Pasteur 1173P2*	GGTGTCGGTCTGACCACTTGA
Mjls_2757	*M. JLS*	CTAACTGGTCAGACCACTTGA
Mmcs_2727	*M. MCS*	CTAACTGGTCAGACCACTTGA
Mkms_2771	*M. KMS*	CTAACTGGTCAGACCACTTGA
MSMEG_3527	*M. smegmatis MC2 155*	ACCACTGGTAAGACCACTTGA
Mvan_2942	*M. vanbaalenii PYR-1*	CACACTGGTCTGACCACTTGA
MAP4081	*M. avium paratuberculosis*	GCCGGTGGTCTGACCACCTGA
Nfa1630	*N. farciana*	ACGATTGGTCTTACCACTTGA

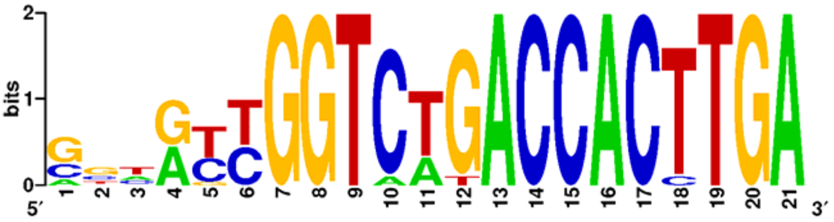


Figure 4.5: Sequence logo showing consensus in the upstream sequences of *M. tuberculosis* Rv0586 and its orthologs.

4.3.4 *In vitro* experimental validation

The conservation among the identified operator sites, and the similarity of DNA binding domain having similar amino acids which are likely to be associated into DNA-protein interaction, gave a possibility of DNA recognition by Rv0586 to the identified operator site across the orthologs. To test the possibility of *in vitro* DNA binding ability, purified Rv0586 protein was subjected to electrophoretic mobility shift assays with set of DNA sequences (Table 4.1). The DNA-protein interaction (Figure 4.6.A-E) showed clear binding with increasing concentration of protein to the synthesized double stranded DNA motifs corresponding to the operator site sequences. The DNA binding was abolished with 50x molar excess concentration of the unlabeled DNA as a specific competitor, but the same fold excess of the non-specific DNA did not affect the DNA protein complex (Figure 4.6.A-E). It clearly depicts the specific DNA-protein interaction of Rv0586 to the operator site. The knowledge of the DNA-binding specificity will add up to as a tool for the search of new physiologically relevant binding sites for GntRs.

4.3.5 Consensus in DNA targets for other GntRs

In accordance to the homology model of Rv0586, *in vitro* results suggested that within a protein family critical residues are more conserved. Indeed, this conservation also constrains the nucleotide conservation within an operator site. Therefore, consensus sequences for the DNA targets identified for the GntR regulators in the previous chapter were drawn (Table 4.2 to 4.7, Figure 4.7).

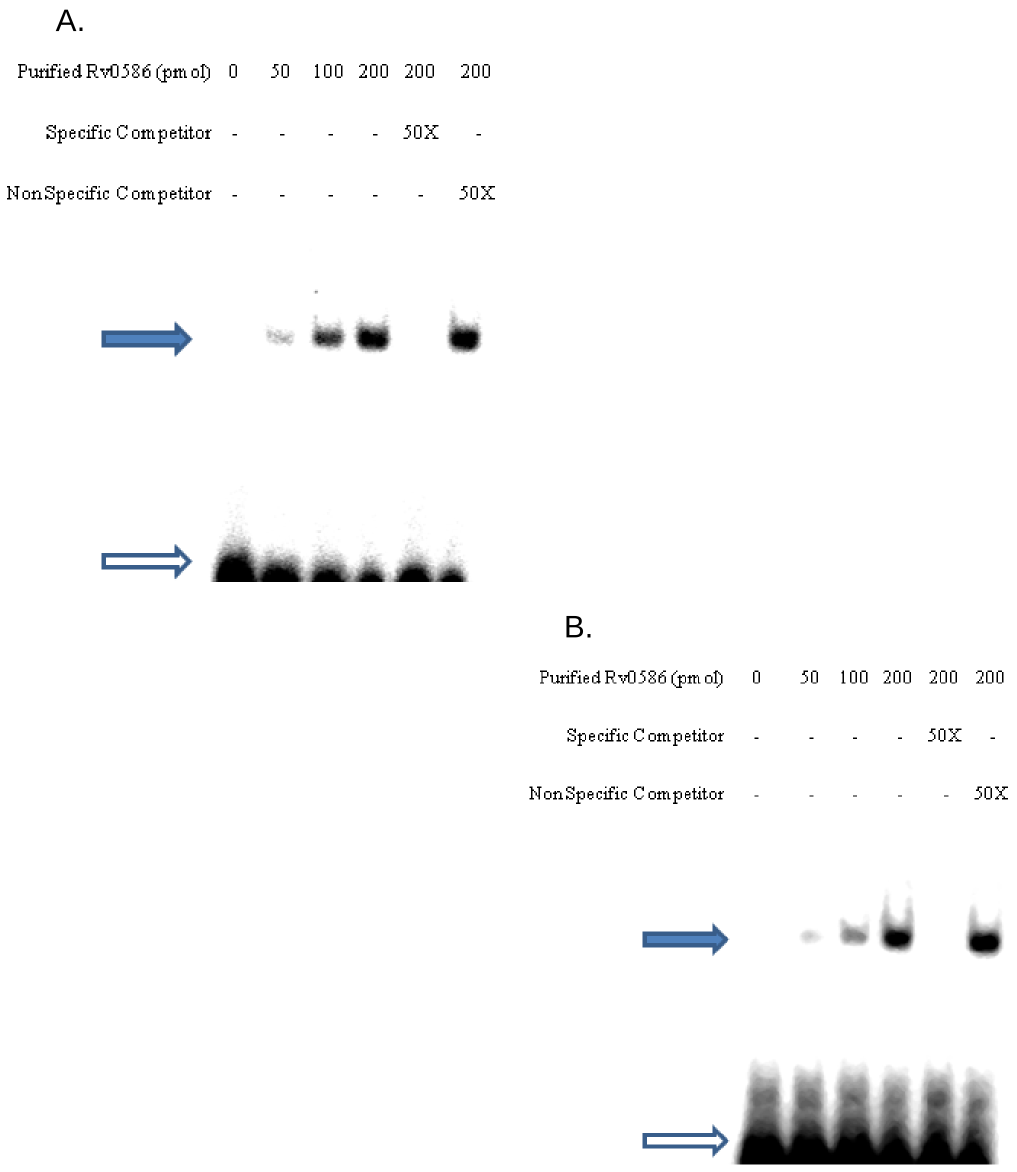
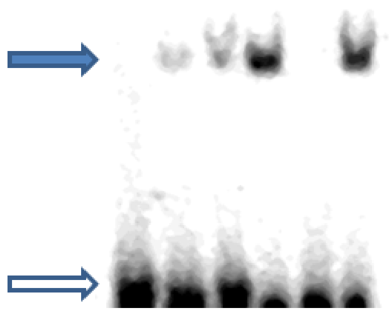

A.
Purified Rv0586 (pmol) 0 50 100 200 200 200
Specific Competitor - - - - 50X -
NonSpecific Competitor - - - - - 50X
B.
Purified Rv0586 (pmol) 0 50 100 200 200 200
Specific Competitor - - - - 50X -
NonSpecific Competitor - - - - - 50X

Figure 4.6: EMSA for the predicted operator sites from other mycobacterial species. Lane 1, labeled fragment; lanes 2–4, labeled fragment with 50, 100, 200 pmol of purified His6-Rv0586 protein per reaction (30µl); lane 5 contains 50-fold molar excess cold specific dsDNA competitor; lane 6 contains the same concentration of cold non-specific competitor. (A) MSMEG_3527 (*M. smegmatis MC2 155*); (B) MAP4081 (*M. avium paratuberculosis*); (C) Mvan_2942 (*M. vanbaalenii PYR*); (D) Mkms_2771/Mjls_2757/Mmcs_2727 (*M. KMS/M. JLS/M. MCS*); (E) Nfa1630 (*N. farciana*).

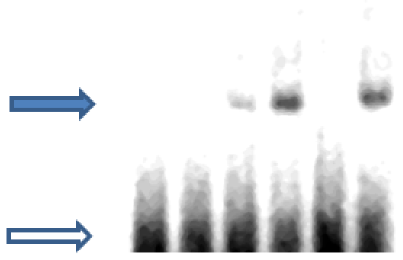
C.

Purified Rv0586 (pmol)	0	50	100	200	200	200
Specific Competitor	-	-	-	-	50X	-
NonSpecific Competitor	-	-	-	-	-	50X



D.

Purified Rv0586 (pmol)	0	50	100	200	200	200
Specific Competitor	-	-	-	-	50X	-
NonSpecific Competitor	-	-	-	-	-	50X



E.

Purified Rv0586 (pmol)	0	50	100	200	200	200
Specific Competitor	-	-	-	-	50X	-
NonSpecific Competitor	-	-	-	-	-	50X

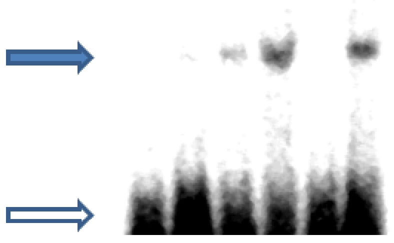


Table 4.2: List of potential operator sites identified for *M. tuberculosis* Rv0043c and its closest orthologs.

Gene/ORF	Organism	Potential operator site
Rv0043c	*M. tuberculosis*	GGACCAGGTTGTCGCCC
Mb0044c	*M. bovis*	GGACCAGGTTGTCGCCC
MRA_0046	*M. tuberculosis H37Ra*	GGACCAGGTTGTCGCCC
TBFG_10042	*M. tuberculosis F11*	GGACCAGGTTGTCGCCC
MT0049	*M. tuberculosis CDC1551*	GGACCAGGTTGTCGCCC
BCG_0074c	*M. bovis BCG Pasteur 1173P2*	GGACCAGGTTGTCGCCC
MAV_0060	*M. avium 104*	GGACCTGGTTGTTTCGC
MAP0053c	*M. avium paratuberculosis*	GGACCTGGTTGTTTCGC
MUL_0061	*M. ulcerans Agy99*	AGACCAGGTTGTTGCGC
Mflv_0859	*M. gilvum PYR-GCK*	GGACCTCGTGGTCACTC
Mvan_6046	*M. vanbaalenii PYR-1*	AGACCTCGTGGTCTCTC
Mjls_5758	*M. JLS*	GGACCAGGTCGTCGCGT
Mkms_5471	*M. KMS*	GGACCAGGTCGTCGCGT
Mmcs_5382	*M. MCS*	GGACCAGGTCGTCGCGT
MSMEG_6908	*M. smegmatis MC2 155*	AGACCAGGTGGTCAGCC

Table 4.3: List of potential operator sites identified for *M. tuberculosis* Rv0165c and its closest orthologs.

Gene/ORF	Organism	Potential operator site
Rv0165c	*M. tuberculosis*	GAGCTGGTTGACCCAGTTC
MAV_5020	*M. avium_104*	CAGCTGGTTCACCCAGTTC
MUL_1058	*M. ulcerans Agy99*	GAGTTGGTTGACCCAGTTC
Mb0170c	*M. bovis*	GAGCTGGTTGACCCAGTTC
BCG_0201c	*M. bovis BCG Pasteur 1173P2*	GAGCTGGTTGACCCAGTTC
TBFG_10166	*M. tuberculosis F11*	GAGCTGGTTGACCCAGTTC
MRA_0173	*M. tuberculosis H37Ra*	GAGCTGGTTGACCCAGTTC
MAP3599c	*M. avium paratuberculosis*	CAGCTGGTTCACCCAGTTC
Mflv_0715	*M. gilvum PYR-GCK*	GAGCTGGTTGGTCCAGTTC
Mvan_0130	*M. vanbaalenii PYR-1*	GAGCTGGTTGGCCCAGTTC
Mjls_0104	*M. JLS*	GAGCTGATTGGTCCAGTTC
Mkms_0123	*M. KMS*	GAGCTGATTGGTCCAGTTC
Mmcs_0114	*M. MCS*	CAGCTGGTTGACCCAGTTC
MSMEG_0130	*M. smegmatis MC2 155*	GAGCTGATTGGTCCAGTTC

Table 4.4: List of potential operator sites identified for *M. tuberculosis* Rv0494 and its closest orthologs.

Gene/ORF	Organism	Potential operator site
Rv0494	*M. tuberculosis*	TATATTGGTTGAGCCAATGAA
Mb0505	*M. bovis*	TATATTGGTTGAGCCAATGAA
BCG_0536	*M. bovis BCG Pasteur 1173P2*	TATATTGGTTGAGCCAATGAA
TBFG_10503	*M. tuberculosis F11*	TATATTGGTTGAGCCAATGAA
MT0514	*M. tuberculosis CDC1551*	TATATTGGTTGAGCCAATGAA
MRA_0501	*M. tuberculosis H37Ra*	TATATTGGTTGAGCCAATGAA
MUL_4564	*M. ulcerans Agy99*	TATATTGGTTGAGCCAATGAA

Table 4.5: List of potential operator sites identified for *M. tuberculosis* Rv0792c and its closest orthologs.

Gene/ORF	Organism	Potential operator site
Rv0792c	*M. tuberculosis*	ATAAGACGTTTTAATACGTCTTAT
TBFG_10808	*M. tuberculosis F11*	ATAAGACGTTTTAATACGTCTTAT
MRA_0802	*M. tuberculosis H37Ra*	ATAAGACGTTTTAATACGTCTTAT
BCG_0845c	*M. bovis BCG Pasteur 1173P2*	ATAAGACGTTTTAATACGTCTTAT
Mb0816c	*M. bovis*	ATAAGACGTTTTAATACGTCTTAT
MUL_3201	*M. ulcerans Agy99*	CGAGTGCGATGTAATACGTTTTAT
MAV_0738	*M. avium 104*	ATGAGACATTTTAATACATCTCGT
MAP0628c	*M. avium paratuberculosis*	ATGAGACGTTTTAATACATCTCGT

Table 4.6: List of potential operator sites identified for *M. tuberculosis* Rv1152 and its closest orthologs.

Gene/ORF	Organism	Potential operator site
Rv1152	*M. tuberculosis*	CCTTGCGTGCTGGACAGCTCGTAAGG
Mflv_2135	*M. gilvum PYR-GCK*	CCGTCGGCGCTGGAGATCTCGTACGG
Mvan_4569	*M. vanbaalenii PYR-1*	CCGTCGGCGCTGGAGATCTCGTAGGG
Mjls_4290	*M. JLS*	CCGTCGGCGCTGGAGATCTCGTACGG
Mkms_4136	*M. KMS*	CCGTCGGCGCTGGAGATCTCGTACGG
Mmcs_4061	*M. MCS*	CCGTCGGCGCTGGAGATCTCGTACGG
MAV_1290	*M. avium 104*	CCCTGGGTGCTGGACAGCTCGTAGGG
MAP2632c	*M. avium paratuberculosis*	CCCTGGGTGCTGGACAGCTCGTAGGG
Mb1183	*M. bovis*	CCTTGCGTGCTGGACAGCTCGTAAGG
BCG_1213	*M. bovis BCG Pasteur 1173P2*	CCTTGCGTGCTGGACAGCTCGTAAGG
TBFG_11176	*M. tuberculosis F11*	CCTTGCGTGCTGGACAGCTCGTAAGG
MRA_1162	*M. tuberculosis H37Ra*	CCTTGCGTGCTGGACAGCTCGTAAGG
MT1186	*M. tuberculosis CDC1551*	CCTTGCGTGCTGGACAGCTCGTAAGG
MUL_0993	*M. ulcerans Agy99*	CCCTGTGTGCTGGACAGCTCGTACGG
MSMEG_5174	*M. smegmatis MC2 155*	CCGTCGGTGCTGGAGATCTCGTACGG

Table 4.7: List of potential operator sites identified for *M. tuberculosis* Rv3060c and its closest orthologs.

Gene/ORF	Organism	Potential operator site
Rv3060c	*M. tuberculosis*	CCTTGACTCAGCCAATA
TBFG_13077	*M. tuberculosis F11*	CCTTGACTCAGCCAATA
MRA_3092	*M. tuberculosis H37Ra*	CCTTGACTCAGCCAATA
MT3146	*M. tuberculosis CDC1551*	CCTTGACTCAGCCAATA
BCG_3085c	*M. bovis BCG Pasteur 1173P2*	CCTTGACTCAGCCAATA
Mb3086c	*M. avium 104*	CCTTGACTCAGCCAATA

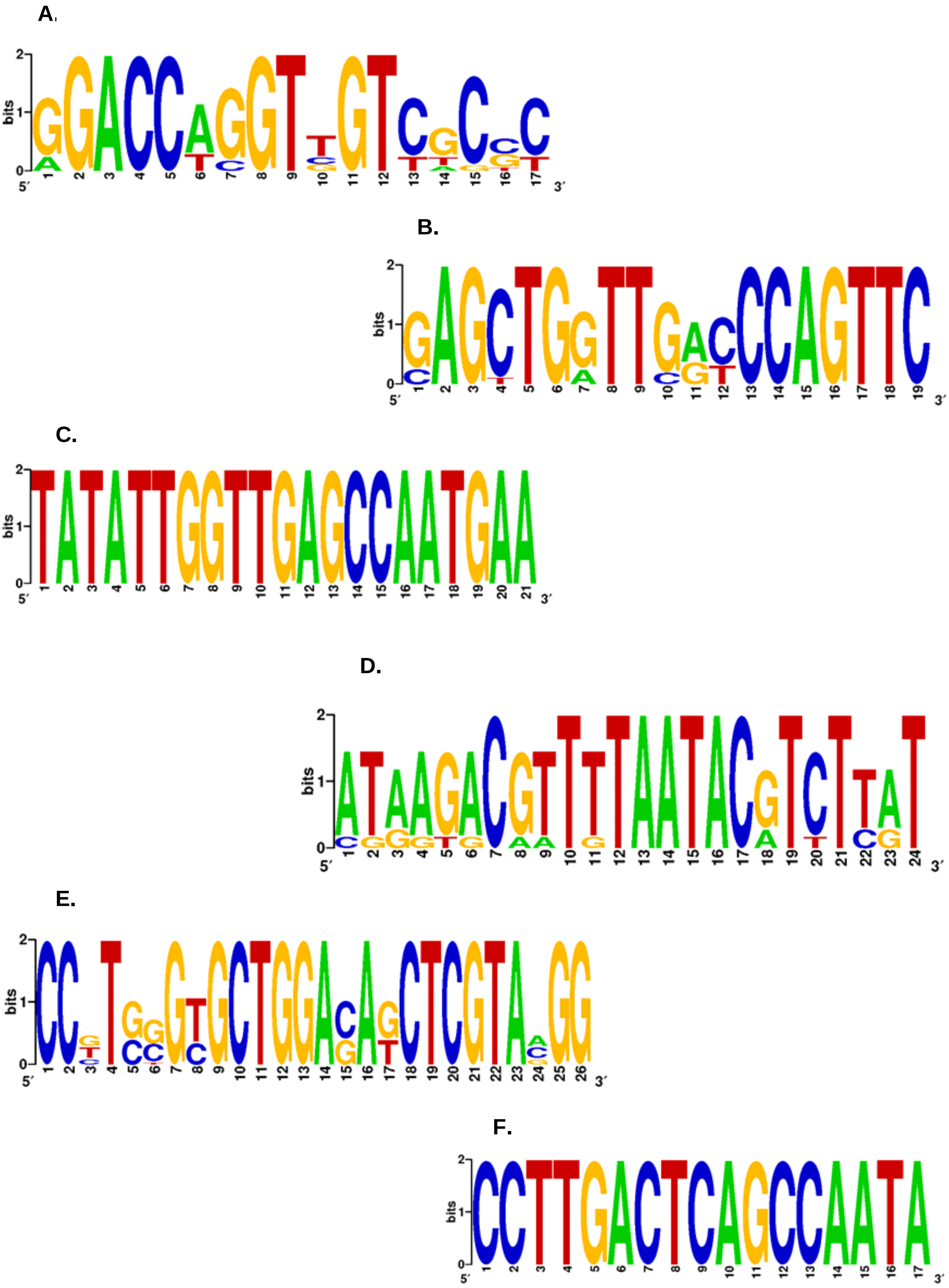


Figure 4.7: Sequence logo showing consensus in the upstream sequences of *M. tuberculosis* GntRs with their respective orthologs. (A). Rv0043c, (B). Rv0165c, (C). Rv0494, (D). Rv0792c, (E). Rv1152, (F). Rv3060c.

4.4 CONCLUSION

Present chapter explores the conservation in DNA binding domain (DBD) across the orthologs of Rv0586. It shows the influence of DBD conservation on the operator site recognition for similar transcriptional regulators. One of the classified GntR transcriptional regulators, Rv0586, was shown as an example. It also displays the nucleotide conservation within the identified operator sites across *M. tuberculosis* GntRs and their orthologous transcriptional regulators. The generated consensus sequence logo clearly shows the sharing of preferred nucleotides associated with subfamilies. Further, in view of the solved structure of FadR-DNA co-crystallized complex, homology model of Rv0586 dimer molecule with the DNA chain was built. This helped in identifying the specific amino acid residues that are in the close proximity of the DNA chain. Additionally, Rv0586 was assessed for its DNA binding ability towards the identified DNA targets from upstream regions of its orthologous transcriptional regulators.

It is worth mentioning that such interactions strengthen the idea of sharing nucleotide preferences among the transcriptional regulators belonging to the same family. It also reveals the influence of the conservation of DNA binding domain upon the specificity of the DNA targets. An outcome of this study is to provide an insight for operator site recognition. These results support the importance of the critical residue analyses in the operator site recognition. In summary, the study highlights the molecular basis of DNA binding to the transcriptional regulators and addresses the possibility of cross-species recognition of targets if the DNA binding domain is similar.

Chapter 5

GntR family of transcriptional regulators from non-infectious Mycobacterium smegmatis

Part of the work is published as:

Vindal V, Suma K and Ranjan A (2007) GntR family of regulators in *Mycobacterium smegmatis*: a sequence and structure based characterization *BMC Genomics*, 8(1):289.

5.1 INTRODUCTION

Being a fast growing, non-pathogenic mycobacteria, *Mycobacterium smegmatis* has been widely used as a model organism to study the biology of other virulent and extremely slow growing species like *Mycobacterium tuberculosis* [Wang and Marcotte, 2008; Chaturvedi *et al*., 2007]. The recently completed genome sequence of *M. smegmatis*, available at TIGR, contains a large number of putative GntR-like regulators. This completed genome sequence provides an opportunity to study the GntR family of transcriptional regulators in this bacterium as well. Earlier chapters have outlined the studies undertaken to describe this family of transcriptional regulators from *M. tuberculosis*. Present study explores these regulators in *M. smegmatis*, which can potentially have major implications for *in vivo* model studies. This study employs the comprehensive sequence based analyses of *M. smegmatis* GntRs. Primarily, besides the annotated GntR-like members at TIGR website, whole *M. smegmatis* proteome was scanned with GntR protein family profile. All identified GntRs were subjected to distance based phylogenetic analyses to classify them into functionally meaningful subfamilies. This analyses was further extended to the identification of cis-regulatory elements in the upstream region of the corresponding regulator. This study will help towards extending the annotation of *M. smegmatis* GntR proteins. Suitable orthologs of the *M. smegmatis* GntRs were also investigated in *M. tuberculosis*, *M. avium, M. bovis*, *M. ulcerans*, *M. sp. KMS*, *M. sp. MCS*, *M. vanbaalenii PYR-1* and *B. subtilis* that has implications for *in vivo* model studies for orthologous regulators from virulent as well as other saprophytic mycobacteria.

5.2 EXPERIMENTAL PROCEDURE

5.2.1 Selection of GntR family members

The sequences of *M. smegmatis MC2* were downloaded from the Institute for Genomic Research Comprehensive Microbial Resource web portal [http://cmr.tigr.org/tigr-scripts/CMR/CmrHomePage.cgi]. Apart from classified GntR regulators or proteins annotated as GntR-like regulator, other putative GntRs from *M. smegmatis* proteome were selected using GntR Pfam profile [Eddy, 1998]. Among the 63 predicted GntR regulators, one protein (MSMEG_3400) was excluded from this study because of its unusual size (741 amino acid) and its annotation as glutamyl-tRNA(Gln) amidotransferase subunit A. The complete protein sequences for rest of the GntR regulators were retrieved from the SWISS-PROT/TrEMBL sequence database as per their Swiss-Prot ID (Table 5.1). [Note: When the study was carried out, complete genome sequence was not available at NCBI ftp site (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/)]

5.2.2 Secondary structure prediction

Secondary structure predictions were made as described in Chapter 2, Section 2.2.3.

5.2.3 Multiple sequence alignments and phylogenetic tree construction

Multiple sequence alignments and phylogenetic tree construction were carried out as per the method described in Chapter 2, Section 2.2.4.

5.2.4 Operator site analyses

To study the upstream region of GntR family regulators, Upstream DNA sequences were extracted that spanned 400 bp upstream to 50 bp downstream in relation to the translation start site of the gene of interest. These upstream regions were analyzed for DNA palindrome exhibiting nucleotide preferences as per their associated subfamilies.

5.2.5 Ortholog prediction

A reciprocal BLAST hit method was utilized to identify the orthologs in two species. It has been described in Chapter 3, Section 3.2.1 in detail. This study sought the best reciprocal BLAST hit for *M. smegmatis* GntR proteins with *Mycobacterium tuberculosis, Mycobacterium avium, Mycobacterium bovis, Mycobacterium ulcerans, Mycobacterium sp* KMS, *Mycobacterium sp.* MCS, *Mycobacterium vanbaalenii* PYR-1 and *Bacillus subtilis*.

Table 5.1: Details of GntR regulators used as representative from all subfamilies in present study

Sub family	Organism	Protein	A. Acid	S P ID
FadR (FadR Gr)	*Escherichia coli O157:H7*	FadR	238	P0A8V8
FadR (VanR Gr)	*Rhizobium leguminosarum*	MatR	222	Q9JP74
MocR	*Rhizobium meliloti*	MocR	493	P49309
HutC	*Pseudomonas putida*	HutC	248	P22773
YtrA	*Bacillus halodurans*	BH0651	123	Q9KF35
	Bacillus halodurans	BH2647	123	Q9K9J9
	Staphylococcus aureus	SAV1934	126	Q99SV4
	Bacillus subtilis	YhcF	121	P54590
	Bacillus subtilis	YtrA	130	O34712
AraR	*Bacillus subtilis*	P96711	362	P96711
	Bacillus halodurans	Q9KBQ0	375	Q9KBQ0
	Bacillus stearothermophilus	Q9S470	364	Q9S470
PlmA	*Synechocystis sp. strain PCC 6803*	sll1961	388	P73804
	Anabaena sp. strain PCC 7120	Q8YXY0	328	Q8YXY0
	Synechococcus elongatus	Q8DH43	367	Q8DH43
	Trichodesmium erythraeum IMS101	Q3HFX5	327	Q3HFX5

5.3 RESULTS AND DISCUSSION

5.3.1 Classification of the putative *M. smegmatis* GntRs into subfamilies

Unrooted tree of the *M. smegmatis* GntRs was constructed with the classified representatives of all subfamilies (Table 5.1). Among all putative *M. smegmatis* GntRs, two proteins (MSMEG_1043 and MSMEG_2323) were found to be identical in sequence, hence only one of them (MSMEG_1043) was taken for the classification. Each branch of the constructed tree represents a subfamily. Bootstrapping, involving 1000 replicates, shows all subfamily branches clustered with high bootstrap values. FadR subfamily is divided into two groups, FadR and VanR (Figure 5.1).

5.3.2 FadR subfamily regulators of *M. smegmatis*

Of all the putative GntRs, 45 proteins were classified as the FadR subfamily regulators. These subfamily members were further classified into two groups’ viz., FadR and VanR groups, where the C-terminal effector binding and/or oligomerization domain length is about 170 and 150 amino acid residues respectively comprising all α-helices [Rigali *et al*., 2002]. Among all FadR subfamily regulators, 19 regulators were clustered as members of the FadR group and 26 of the VanR group (Table 5.2). Due to a large number of these regulators, secondary structural features of members of both group were studied separately. C-terminal domains of all the members of FadR group were predicted with seven α-helices except MSMEG_2599 (Figure 5.2 and Figure 5.3) [Rigali *et al*., 2002]. Secondary structural patterns of the regulator MSMEG_3959 revealed an extra secondary structural element, which could be significant in studying protein family evolution. FadR subfamily regulators are known to be involved in the regulation of gene expression in response to oxidized substrates related either to amino acid metabolism or in various metabolic pathways. One of the classified FadR subfamily transcriptional regulators, MSMEG_6700 is known to be involved in the regulation of piperidine and pyrrolidine metabolism [Poupin *et al.*, 1999].

Table 5.2: List of classified *M. smegmatis* GntR regulators

Gene	SF	A. Acid	Gene	SF	A. Acid
MSMEG_0124	FadR	227	MSMEG_2546	FadR	239
MSMEG_0130	FadR	230	MSMEG_2599	FadR	224
MSMEG_0166	FadR	242	MSMEG_2605	FadR	255
MSMEG_0179	FadR	223	MSMEG_2682	FadR	262
MSMEG_0268	HutC	292	MSMEG_2794	FadR	225
MSMEG_0286	HutC	228	MSMEG_2910	FadR	235
MSMEG_0426	MocR	469	MSMEG_3345	FadR	258
MSMEG_0454	FadR	245	MSMEG_3822	FadR	267
MSMEG_0480	FadR	219	MSMEG_3527	FadR	240
MSMEG_0535	FadR	212	MSMEG_3959	FadR	290
MSMEG_0596	FadR	228	MSMEG_3980	FadR	214
MSMEG_0650	HutC	244	MSMEG_4042	FadR	252
MSMEG_0778	HutC	246	MSMEG_4057	FadR	221
MSMEG_0874	FadR	234	MSMEG_4121	FadR	229
MSMEG_0895	FadR	247	MSMEG_4140	MocR	508
MSMEG_2323	MocR	534	MSMEG_4659	HutC	245
MSMEG_1117	FadR	239	MSMEG_5174	YtrA	121
MSMEG_1227	HutC	274	MSMEG_5201	FadR	254
MSMEG_1317	FadR	229	MSMEG_5375	FadR	230
MSMEG_1572	MocR	470	MSMEG_5630	HutC	245
MSMEG_1995	FadR	241	MSMEG_5731	FadR	240
MSMEG_2009	FadR	226	MSMEG_5760	MocR	463
MSMEG_2104	MocR	449	MSMEG_6300	FadR	224
MSMEG_2164	FadR	262	MSMEG_6371	MocR	488
MSMEG_2173	FadR	230	MSMEG_6639	FadR	222
MSMEG_2209	FadR	222	MSMEG_6700	FadR	245
MSMEG_1043	MocR	534	MSMEG_6738	FadR	227
MSMEG_2453	FadR	244	MSMEG_6745	HutC	247
MSMEG_2480	FadR	246	MSMEG_6789	FadR	246
MSMEG_2489	FadR	240	MSMEG_6881	FadR	209
MSMEG_2531	FadR	253	MSMEG_6908	FadR	221

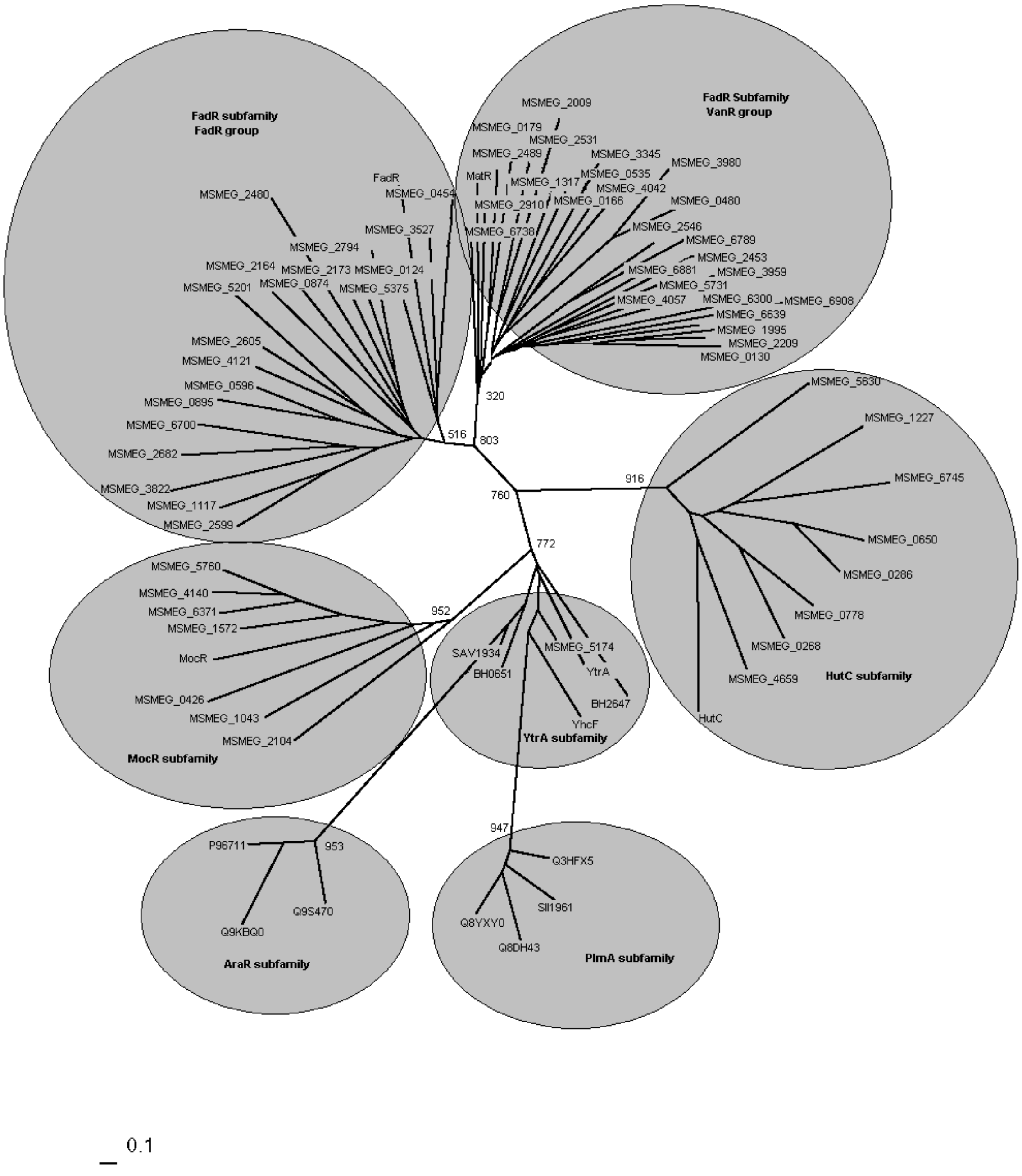
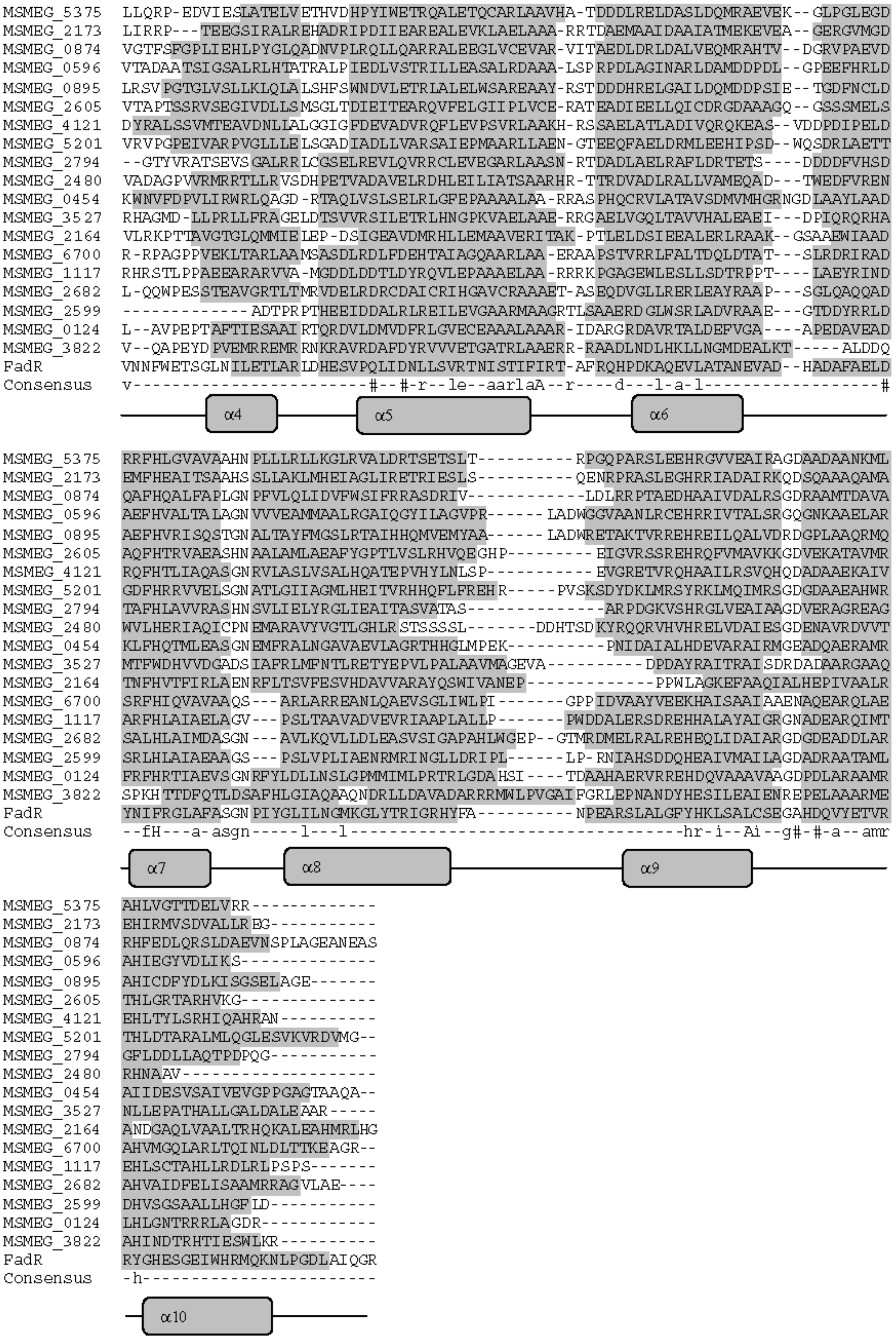


Figure 5.1: Unrooted tree of the proteins of GntR family regulators of *M. smegmatis* including representatives of all subfamily regulators from different bacterial genomes with 1000 bootstrap replicates. All the GntR regulators are clustered into six subfamilies. FadR subfamily is branched again into two groups (FadR and VanR). (Note: abbreviations are as indicated in Table 5.1 and Table 5.2).



```
MSMEG_5375   LLQRP-EDVIESLATELVETHVDHPYIWETRQALETQCARLAAVHA-TDDDLRELDASLDQMRAEVEK--GLPGLEGD
MSMEG_2173   LIRRP---TEEGSIRALREHADRIPDIIEAREALEVKLAELAAA-RRTDAEMAAIDAAIATMEKEVEA--GERGVMGD
MSMEG_0874   VGTFSFGPLIEHLPYGLQADNVPLRQLLQARRALEEGLVCEVAR-VITAEDLDRLDALVEQMRAHTV--DGRVPAEVD
MSMEG_0596   VTADAATSIGSALRLHTATRALPIEDLVSTRILLEASALRDAAA-LSPRPDLAGINARLDAMDDPDL--GPEEFHRLD
MSMEG_0895   LRSVPGTGLVSLLKLQLALSHFSWNDVLETRLALELWSAREAAY-RSTDDDHRELGAILDQMDDPSIE--TGDFNCLD
MSMEG_2605   VTAPTSSRVSEGIVDLLSMSGLTDIEITEARQVFELGIIPLVCE-RATEADIEELLQICDRGDAAAGQ--GSSSMELS
MSMEG_4121   DYRALSSVMTEAVDNLLALGGIGFDEVADVRQFLEVPSVRLAAKH-RSSAELATLADIVQRQKEAS--VDDPDIPELD
MSMEG_5201   VRVPGPEIVARPVGLLLELSGADIADLLVARSAIEPMAARLLAEN-GTEEQFAELDRMLEEHIPSD--WQSDRLAETT
MSMEG_2794   --GTYVRATSEVSGALRRLCGSELREVLQVRRCLEVEGARLAASN-RTDADLAELRAFLDRTETS----DDDDFVHSD
MSMEG_2480   VADAGPVVRMRRTLLRVSDHPETVADAVELRDHLEILIATSAARHR-TTRDVADLRALLVAMEQAD---TWEDFVREN
MSMEG_0454   KWNVFDPVLIRWRLQAGD-RTAQLVSLSELRLGFEPAAAALAA-RRASPHQCRVLATAVSDMVMHGRNGDLAAYLAAD
MSMEG_3527   RHAGMD-LLPRLLFRAGELDTSVVRSILETRLHNGPKVAELAAE-RRGAELVGQLTAVVHALEAEI---DPIQRQRHA
MSMEG_2164   VLRKPTTAVGTGLQMMIELEP-DSIGEAVDMRHLLEMAAVERITAK-PTLELDSIEEALERLRAAK--GSAAEWIAAD
MSMEG_6700   R-RPAGPPVEKLTARLAAMSASDLRDLFDEHTAIAGQAARLAA-ERAAPSTVRRLFALTDQLDTAT---SLRDRIRAD
MSMEG_1117   RHRSTLPPAEEARARVVA-MGDDLDDTLDYRQVLEPAAAELAA-RRRKPGAGEWLESLLSDTRPPT----LAEYRIND
MSMEG_2682   L-QQWPESSTEAVGRTLTMRVDELRDRCDAICRIHGAVCRAAAET-ASEQDVGLLRERLEAYRAAP---SGLQAQQAD
MSMEG_2599   ------------ADTPRPTHEEIDDALRLREILEVGAARMAAGRTLSAAERDGLWSRLADVRAAE---GTDDYRRLD
MSMEG_0124   L--AVPEPTAFTIESAAIRTQRDVLDMVDFRLGVECEAAALAAAR-IDARGRDAVRTALDEFVGA----APEDAVEAD
MSMEG_3822   V--QAPEYDPVEMRREMRRNKRAVRDAFDYRVVVETGATRLAAERR-RAADLNDLHKLLNGMDEALKT-----ALDDQ
FadR         VNNFWETSGLNILETLARLDHESVPQLIDNLLSVRTNISTIFIRT-AFRQHPDKAQEVLATANEVAD--HADAFAELD
Consensus    v------------------------#--#-r--le--aarlaA--r----d---l-a-l------------------#

MSMEG_5375   RRFHLGVAVAAHNPLLLRLLKGLRVALDRTSETSLT----------RPGQPARSLEEHRGVVEAIRAGDAADAANKML
MSMEG_2173   EMFHEAITSAAHSSLLAKLMHEIAGLIRETRIESLS----------QENRPRASLEGHRRIADAIRKQDSQAAAQAMA
MSMEG_0874   QAFHQALFAPLGNPFVLQLIDVFWSIFRRASDRIV------------LDLRRPTAEDHAAIVDALRSGDRAAMTDAVA
MSMEG_0596   AEFHVALTALAGNVVVEAMMAALRGAIQGYILAGVPR------LADWGGVAANLRCEHRRIVTALSRGQGNKAAELAR
MSMEG_0895   AEFHVRISQSTGNALTAYFMGSLRTAIHHQMVEMYAA------LADWRETAKTVRREHREILQALVDRDGPLAAQRMQ
MSMEG_2605   AQFHTRVAEASHNAALAMLAEAFYGPTLVSLRHVQEGHP---------EIGVRSSREHRQFVMAVKKGDVEKATAVMR
MSMEG_4121   RQFHTLIAQASGNRVLASLVSALHQATEPVHYLNLSP-----------EVGRETVRQHAAILRSVQHQDADAAEKAIV
MSMEG_5201   GDFHRRVVELSGNATLGIIAGMLHEITVRHHQFLFREHR-----PVSKSDYDKLMRSYRKLMQIMRSGDGDAAEAHWR
MSMEG_2794   TAFHLAVVRASHNSVLIELYRGLIEAITASVATAS--------------ARPDGKVSHRGLVEAIAAGDVERAGREAG
MSMEG_2480   WVLHERIAQICPNEMARAVYVGTLGHLRSTSSSSL-------DDHTSDKYRQQRVHVHRELVDAIESGDENAVRDVVT
MSMEG_0454   KLFHQTMLEASGNEMFRALNGAVAEVLAGRTHHGLMPEK----------PNIDAIALHDEVARAIRMGEADQAERAMR
MSMEG_3527   MTFWDHVVDGADSIAFRLMFNTLRETYEPVLPALAAVMAGEVA----------DPDAYRAITRAISDRDADAARGAAQ
MSMEG_2164   TNFHVTFIRLAENRFLTSVFESVHDAVVARAYQSWIVANEP-------------PPWLAGKEFAAQIALHEPIVAALR
MSMEG_6700   SRFHIQVAVAAQS---ARLARREANLQAEVSGLIWLPI-------GPPIDVAAYVEEKHAISAAIAAENAQEARQLAE
MSMEG_1117   ARFHLAIAELAGV---PSLTAAVADVEVRIAAPLALLP-------PWDDALERSDREHHALAYAIGRGNADEARQIMT
MSMEG_2682   SALHLAIMDASGN---AVLKQVLLDLEASVSIGAPAHLWGEP--GTMRDMELRALREHEQLIDAIARGDGDEADDLAR
MSMEG_2599   SRLHLAIAEAAGS---PSLVPLIAENRMRINGLLDRIPL------LP-RNIAHSDDQHEAIVMAILAGDADRAATAML
MSMEG_0124   FRFHRTIAEVSGNRFYLDLLNSLGPMMIMLPRTRLGDAHSI----TDAAHAERVRREHDQVAAAVAAGDPDLARAAMR
MSMEG_3822   SPKHTTDFQTLDSAFHLGIAQAAQNDRLLDAVADARRRMWLPVGAIFGRLEPNANDYHESILEAIENREPELAAARME
FadR         YNIFRGLAFASGNPIYGLILNGMKGLYTRIGRHYFA----------NPEARSLALGFYHKLSALCSEGAHDQVYETVR
Consensus    --fH---a-asgn-----l---l----------------------------------hr-i--Ai--g#-#-a--amr

MSMEG_5375   AHLVGTTDELVRR-------------
MSMEG_2173   EHIRMVSDVALLREG-----------
MSMEG_0874   RHFEDLQRSLDAEVNSPLAGEANEAS
MSMEG_0596   AHIEGYVDLIKS--------------
MSMEG_0895   AHICDFYDLKISGSELAGE-------
MSMEG_2605   THLGRTARHVKG--------------
MSMEG_4121   EHLTYLSRHIQAHRAN----------
MSMEG_5201   THLDTARALMLQGLESVKVRDVMG--
MSMEG_2794   GFLDDLLAQTPDPQG-----------
MSMEG_2480   RHNAAV--------------------
MSMEG_0454   AIIDESVSAIVEVGPPGAGTAAQA--
MSMEG_3527   NLLEPATHALLGALDALEAAR-----
MSMEG_2164   ANDGAQLVAALTRHQKALEAHMRLHG
MSMEG_6700   AHVMGQLARLTQINLDLTTKEAGR--
MSMEG_1117   EHLSCTAHLLRDLRLPSPS-------
MSMEG_2682   AHVAIDFELISAAMRRAGVLAE----
MSMEG_2599   DHVSGSAALLHGFLD-----------
MSMEG_0124   LHLGNTRRRLAGDR------------
MSMEG_3822   AHINDTRHTIESWLKR----------
FadR         RYGHESGEIWHRMQKNLPGDLAIQGR
Consensus    -h------------------------
```

Figure 5.2: Structure based sequence analyses of *M. smegmatis* GntR family of regulators by the multiple sequence alignment of the C-terminal domains of GntR regulators belonging to FadR Subfamily (FadR group). (Note: gene abbreviations are as indicated in Table 5.1; details of consensus symbol are given in Chapter 2, method Section 2.2.4)

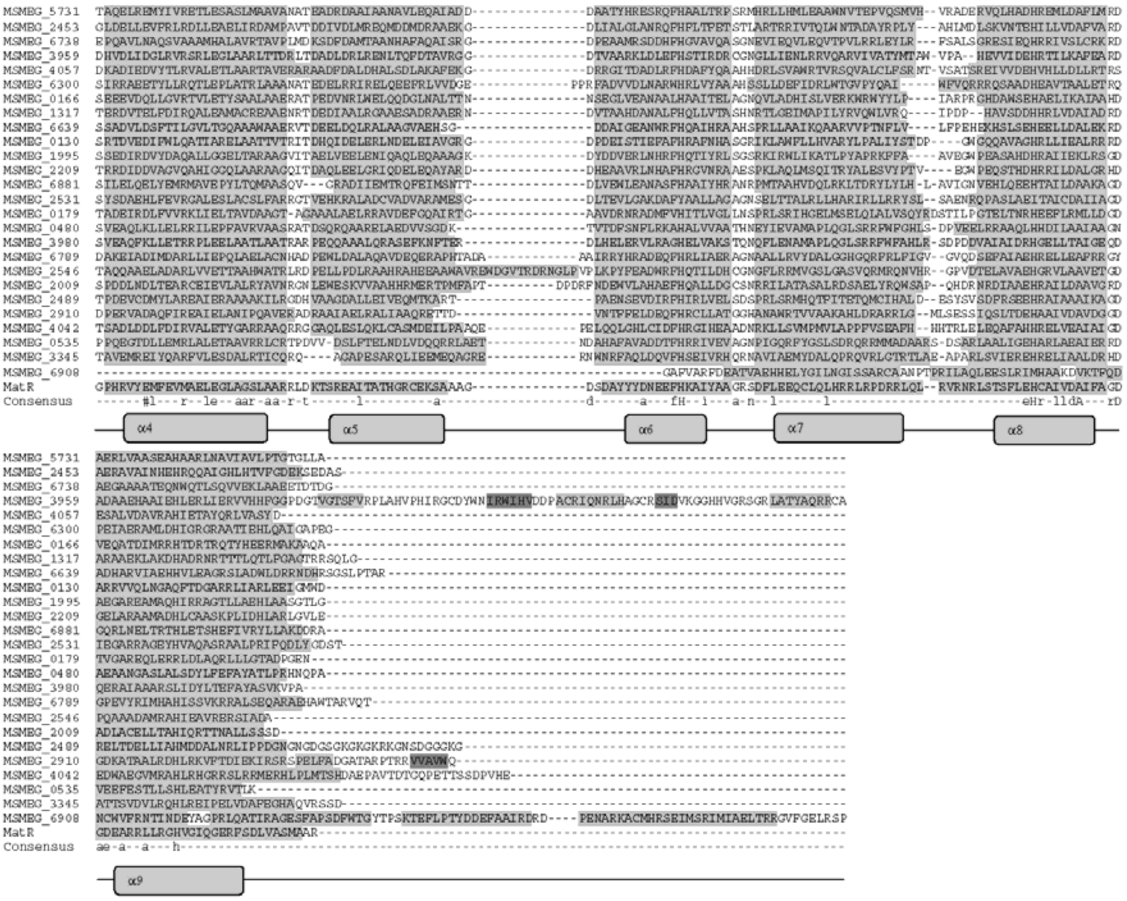


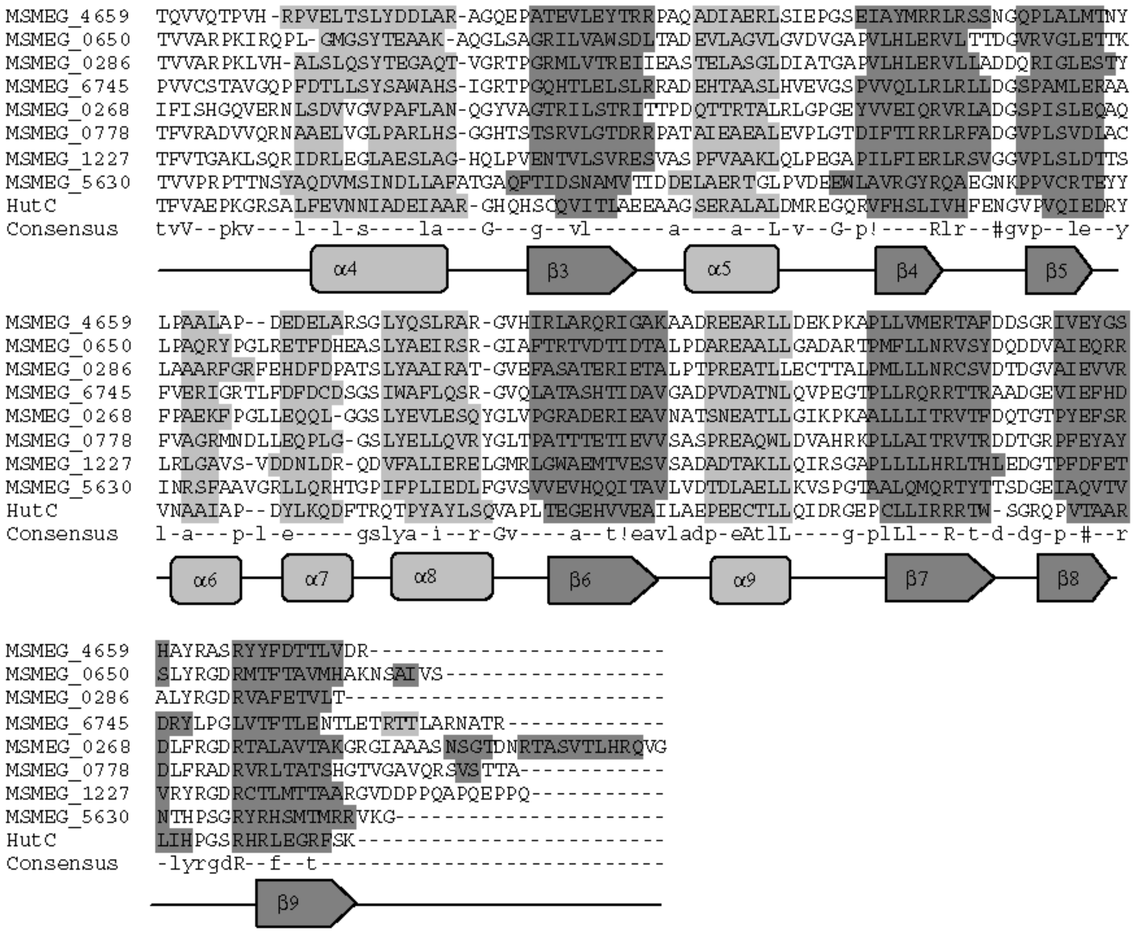
Figure 5.3: Structure based sequence analyses of *M. smegmatis* GntR family regulators by the multiple sequence alignment of C-terminal domains of GntR regulators belonging to FadR Subfamily (VanR group). (Note: gene abbreviations are as indicated in Table 5.1; details of consensus symbol are given in Chapter 2, method Section 2.2.4)

5.3.3 HutC subfamily regulators of *M. smegmatis*

Contrary to the FadR subfamily regulators, the regulators of this subfamily consist of both α-helices and β-strand structures in the C-terminal domain. Eight GntRs were identified as members of this subfamily (Table 5.2). All these members showed distinguishable predicted secondary structural features specific to this subfamily (Figure 5.4) [Rigali *et al*., 2002]. These regulators are known to acquire the same protein fold as *Escherichia coli* UbiC protein [Aravind and Anantharaman, 2003]. These effector-binding domains responds to various ligands like histidine (HutC) [Allison and Phillips, 1990], long chain fatty acids [Quail *et al*., 1994], trehalose 6-phosphate [Matthijs *et al*., 2000; Schock and Dahl, 1996] or alkylphosphonate [Chen *et al*., 1990]. A range of known ligands, specific to many HutC subfamily regulators, will help in characterizing the classified *M. smegmatis* regulators.

5.3.4 MocR subfamily regulators of *M. smegmatis*

Among all the putative GntR regulators, eight were classified as members of the MocR subfamily (Table 5.2). All the eight regulators showed distinguishable predicted secondary structural features specific to this subfamily (Figure 5.5) [Rigali *et al*., 2002]. MocR subfamily regulators show homology to the class I aminotransferase proteins [Sung *et al*., 1991], which requires pyridoxal 5’-phosphate (PLP) as a co-factor. All MocR subfamily regulators exhibit a PLP attachment site with a conserved lysine residue, which is also evident in the classified MocR subfamily regulators (Figure 5.5).



```
MSMEG_4659   TQVVQTPVH-RPVELTSLYDDLAR-AGQEPATEVLEYTRRPAQADIAERLSIEPGSEIAYMRRLRSSNGQPLALMTNY
MSMEG_0650   TVVARPKIRQPL-GMGSYTEAAK-AQGLSAGRILVAWSDLTADEVLAGVLGVDVGAPVLHLERVLTTDGVRVGLETTK
MSMEG_0286   TVVARPKLVH-ALSLQSYTEGAQT-VGRTPGRMLVTREIIEASTELASGLDIATGAPVLHLERVLLADDQRIGLESTY
MSMEG_6745   PVVCSTAVGQPFDTLLSYSAWAHS-IGRTPGQHTLELSLRRADEHTAASLHVEVGSPVVQLLRLRLLDGSPAMLERAA
MSMEG_0268   IFISHGQVERNLSDVVGVPAFLAN-QGYVAGTRILSTRITTPDQTTRTALRLGPGEYVVEIQRVRLADGSPISLEQAQ
MSMEG_0778   TFVRADVVQRNAAELVGLPARLHS-GGHTSTSRVLGTDRRPATAIEAEALEVPLGTDIFTIRRLRFADGVPLSVDLAC
MSMEG_1227   TFVTGAKLSQRIDRLEGLAESLAG-HQLPVENTVLSVRESVASPFVAAKLQLPEGAPILFIERLRSVGGVPLSLDTTS
MSMEG_5630   TVVPRPTTNSYAQDVMSINDLLAFATGAQFTIDSNAMVTIDDELAERTGLPVDEEWLAVRGYRQAEGNKPPVCRTEYY
HutC         TFVAEPKGRSALFEVNNIADEIAAR-GHQHSCQVITLAEEAAGSERALALDMREGQRVFHSLIVHFENGVPVQIEDRY
Consensus    tvV--pkv---l--l-s----la---G---g--vl------a----a--L-v--G-p!----Rlr--#gvp--le--y

MSMEG_4659   LPAALAP--DEDELARSGLYQSLRAR-GVHIRLARQRIGAKAADREEARLLDEKPKAPLLVMERTAFDDSGRIVEYGS
MSMEG_0650   LPAQRYPGLRETFDHEASLYAEIRSR-GIAFTRTVDTIDTALPDAREAALLGADARTPMFLLNRVSYDQDDVAIEQRR
MSMEG_0286   LAAARFGRFEHDFDPATSLYAAIRAT-GVEFASATERIETALPTPREATLLECTTALPMLLLNRCSVDTDGVAIEVVR
MSMEG_6745   FVERIGRTLFDFDCDSGSIWAFLQSR-GVQLATASHTIDAVGADPVDATNLQVPEGTPLLRQRRTTRAADGEVIEFHD
MSMEG_0268   FPAEKFPGLLEQQL-GGSLYEVLESQYGLVPGRADERIEAVNATSNEATLLGIKPKAALLLITRVTFDQTGTPYEFSR
MSMEG_0778   FVAGRMNDLLEQPLG-GSLYELLQVRYGLTPATTTETIEVVSASPREAQWLDVAHRKPLLAITRVTRDDTGRPFEYAY
MSMEG_1227   LRLGAVS-VDDNLDR-QDVFALIERELGMRLGWAEMTVESVSADADTAKLLQIRSGAPLLLLHRLTHLEDGTPFDFET
MSMEG_5630   INRSFAAVGRLLQRHTGPIFPLIEDLFGVSVVEVHQQITAVLVDTDLAELLKVSPGTAALQMQRTYTTSDGEIAQVTV
HutC         VNAAIAP--DYLKQDFTRQTPYAYLSQVAPLTEGEHVVEAILAEPEECTLLQIDRGEPCLLIRRRTW-SGRQPVTAAR
Consensus    l-a---p-l-e-----gslya-i--r-Gv----a--t!eavladp-eAtlL----g-plLl--R-t-d-dg-p-#--r

MSMEG_4659   HAYRASRYYFDTTLVDR------------------------
MSMEG_0650   SLYRGDRMTFTAVMHAKNSAIVS------------------
MSMEG_0286   ALYRGDRVAFETVLT--------------------------
MSMEG_6745   DRYLPGLVTFTLENTLETRTTLARNATR-------------
MSMEG_0268   DLFRGDRTALAVTAKGRGIAAASNSGTDNRTASVTLHRQVG
MSMEG_0778   DLFRADRVRLTATSHGTVGAVQRSVSTTA------------
MSMEG_1227   VRYRGDRCTLMTTAARGVDDPPQAPQEPPQ-----------
MSMEG_5630   NTHPSGRYRHSMTMRRVKG----------------------
HutC         LIHPGSRHRLEGRFSK-------------------------
Consensus    -lyrgdR--f--t----------------------------
```

Figure 5.4: Structure based sequence analyses of *M. smegmatis* GntR family regulators by the multiple sequence alignment of C-terminal domains of GntR regulators belonging to the HutC Subfamily. (Note: abbreviations are as indicated in Table 5.1; details of consensus symbol are given in method section 2.2.4)

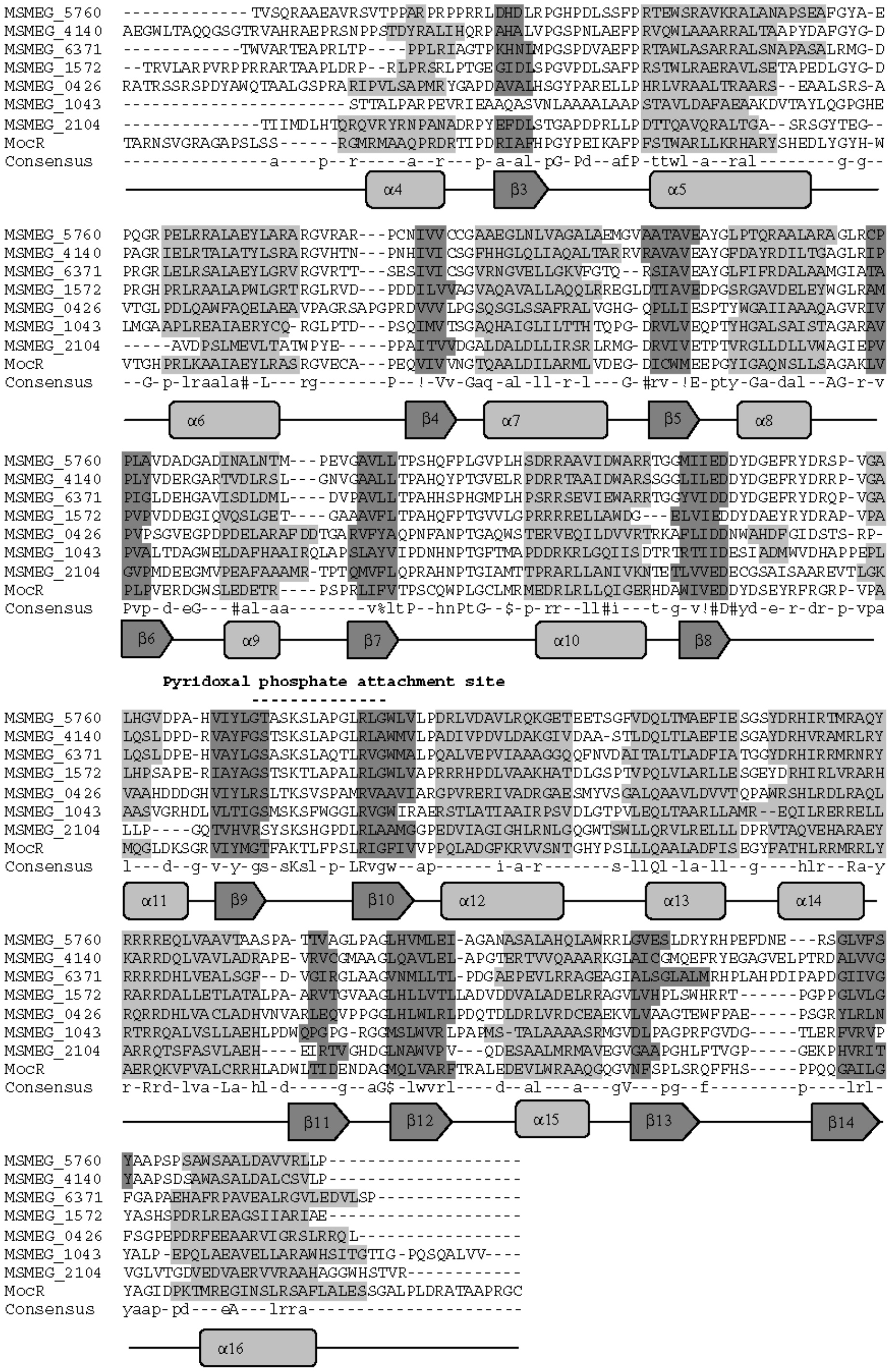


Figure 5.5: Structure based sequence analyses of *M. smegmatis* GntR family regulators by the multiple sequence alignment of C-terminal domains of GntR regulators belonging to the MocR Subfamily. (Note: abbreviations are as indicated in Table 5.1; details of consensus symbol are given in Chapter 2, method Section 2.2.4)

5.3.5 YtrA subfamily regulator of *M. smegmatis*

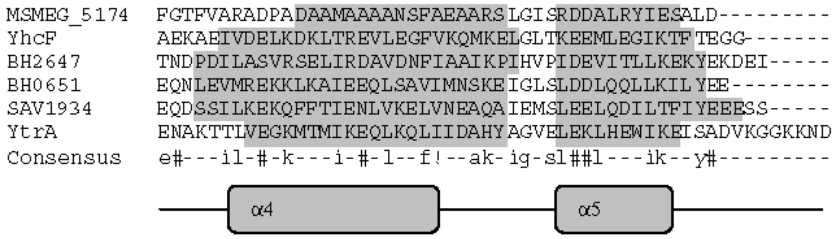
The YtrA subfamily is the least represented GntR family regulator in bacterial genomes, which was also observed in the *M. smegmatis* genome. Among all *M. smegmatis* GntR regulators, only one regulator, *i.e.* MSMEG_5174 showed the signatures of theYtrA subfamily member (Table 5.2, Figure 5.6). YtrA possesses a reduced C-terminal domain with only two α-helices. The average length of the putative effector binding and/or oligomerization domain is about 50 amino acids [Rigali *et al*., 2002]. YtrA from *B. subtilis* is an experimentally explored regulator, which is part of a large self-regulated operon. This operon consists of genes encoding the ATP binding cassette (ABC) transport systems in addition to the YtrA [Yoshida *et al.,* 2000].

5.3.6 Operator/binding site analyses

Potential operator sites with near perfect palindrome sequences with conserved residues, which are found to be specific for most of the subfamily members were listed (Table 5.3). However operator sites in the upstream sequences of all the remaining regulators were not found. All the predicted sites were found to be in the upstream region from the translation start site except MSMEG_2599. Identification of these sites is an important step to understand the GntR associated regulon or the gene regulatory network in the genome **[**Yellaboina *et al*., 2004; Ranjan *et al*., 2006**].**

5.3.7 Ortholog prediction

A large number of identified *M. smegmatis* GntR regulators were annotated as putative orthologs of proteins from other species of mycobacteria and *B. subtilis* (Table 5.4). As orthologs typically share the same function, these regulators could serve as a model to study homologues from other species of mycobacteria. These characterized orthologs may provide clues for initiating detailed biochemical characterization of *M. smegmatis* proteins. Many putative orthologs were experimentally known like Rv0165c that is involved in regulation of the *mce1* operon [Casali *et al*., 2006]; GntR, a transcriptional repressor of gluconate operon [Fujita *et al*., 1986; Reizer *et al*., 1991]; YcbG, involved in utilization of D-glucarate and D-galactarate [Hosoya *et al*., 2002]; YcnF, involved in utilization of gamma-aminobutyrate [Belitsky and Sonenshein, 2002]. However, orthologs for all *M. smegmatis* GntRs were not found in pathogenic species.



```
MSMEG_5174   FGTFVARADPADAAMAAAANSFAEAARSLGISRDDALRYIESALD---------
YhcF         AEKAEIVDELKDKLTREVLEGFVKQMKELGLTKEEMLEGIKTFTEGG-------
BH2647       TNDPDILASVRSELIRDAVDNFIAAIKPIHVPIDEVITLLKEKYEKDEI-----
BH0651       EQNLEVMREKKLKAIEEQLSAVIMNSKEIGLSLDDLQQLLKILYEE--------
SAV1934      EQDSSILKEKQFFTIENLVKELVNEAQAIEMSLEELQDILTFIYEEESS-----
YtrA         ENAKTTLVEGKMTMIKEQLKQLIIDAHYAGVELEKLHEWIKEISADVKGGKKND
Consensus    e#---il-#-k---i-#-l--f!--ak-ig-sl##l---ik--y#---------
```

Figure 5.6: Structure based sequence analyses of *M. smegmatis* GntR family regulators by the multiple sequence alignment of the C-terminal domains of GntR regulators belonging to YtrA Subfamily. (Note: abbreviations are as indicated in Table 5.1.; details of consensus symbol are given in Chapter 2, method Section 2.2.4)

Table 5.3: List of predicted potential operator sites

Subfamily	Regulator	Potential Operator sequence
FadR	MSMEG_0124	CCACTGTTCAACGAGCG
	MSMEG_0179	AAGATCGTCCGACAATT
	MSMEG_0454	CAATCGTCATACGATTG
	MSMEG_0596	GTGTGGTCAGACCACAC
	MSMEG_0895	TCGTGGGACGA
	MSMEG_2164	CCGTTGAACGG
	MSMEG_2480	ACCGGTGGCACCAGGGT
	MSMEG_2599	ACCGTGGGACGGT
	MSMEG_2682	TGGCAAGACCA
	MSMEG_2910	CCTTGATGTCCCACAACG
	MSMEG_3527	TGGTAAGACCA
	MSMEG_3822	TTGTTACTCAA
	MSMEG_3959	TTGCCGCGCGACAA
	MSMEG_3980	TGGTGATACACCA
	MSMEG_4057	TTCGTGTCACAAGTCGAA
	MSMEG_6789	TTTGTGTCACAAA
HutC	MSMEG_0268	ACCGTCTACATCGT
	MSMEG_0650	TGGTCTATACCA
YtrA	MSMEG_5174	GCCATCATGTAGTGC

Table 5.4: Orthologs of *M. smegmatis* GntR family regulators in other bacterial species

M.smeg	***M.tub***	***M.aviump***	***M.bov***	***M.van***	***M.spMCS***	***M.spKMS***	***M.ulc***	***B.sub***
MSMEG_0130	Rv0165c	MAP3599c	Mb0170c	Mvan_0130	Mmcs_0114	Mkms_0123	MUL_1058	-
MSMEG_0179	-	-	-	-	-	-	MUL_1833	-
MSMEG_0268	-	-	-	Mvan_5574	Mmcs_0189	Mkms_0198	-	-
MSMEG_0286	-	-	-	Mvan_0056	-	-	-	-
MSMEG_0454	-	-	-	Mvan_5910	-	Mkms_5416	-	-
MSMEG_0535	-	-	-	-	-	-	-	GntR
MSMEG_0596	-	-	-	-	-	Mkms_4471	-	-
MSMEG_1043	-	-	-	Mvan_2084	-	Mkms_1901	-	-
MSMEG_1227	-	MAP1105	-	-	-	-	-	-
MSMEG_1317	-	-	-	Mvan_3051	-	-	-	-
MSMEG_2104	-	MAP1267	-	-	-	-	MUL_1552	-
MSMEG_2173	-	-	-	Mvan_0294	-	-	-	YcbG
MSMEG_2209	-	MAP2404c	-	Mvan_1978	-	Mkms_1807	MUL_3894	-
MSMEG_2599	-	-	-	Mvan_2282	-	Mkms_2107	-	-
MSMEG_2794	-	-	-	Mvan_0952	-	Mkms_0349	MUL_1381	-
MSMEG_3527	Rv0586	-	Mb0601	Mvan_2942	-	Mkms_2771	MUL_4564	-
MSMEG_3822	-	-	-	Mvan_0606	-	Mkms_0519	-	-
MSMEG_4057	-	-	-	-	-	-	-	YdhC
MSMEG_4140	-	-	-	-	-	-	-	YcnF
MSMEG_4659	Rv0792c	MAP0628c	Mb0816c	Mvan_4015	-	-	MUL_0525	YvoA
MSMEG_5174	Rv1152	MAP2632c	Mb1183	Mvan_4569	-	-	MUL_0993	YtrA
MSMEG_5201	Rv3060c	MAP2347	Mb3086c	Mvan_4590	-	Mkms_4157	MUL_3832	-
MSMEG_5630	-	MAP3505c	-	Mvan_4965	-	Mkms_4496	MUL_4818	-
MSMEG_5731	-	-	-	Mvan_0931	-	Mkms_4957	-	-
MSMEG_6371	-	-	-	Mvan_5625	-	Mkms_5086	-	YhdI
MSMEG_6700	-	-	-	Mvan_1846	-	-	-	-
MSMEG_6908	Rv0043c	MAP0053c	Mb0044c	Mvan_6046	-	Mkms_5471	MUL_0061	-

Note: '-' Represents, corresponding orthologs are not present in the genome. *M.smeg – M. smegmatis; M.tub – M. tuberculosis; M.aviump. – M. avium para.; M.bov – M. bovis; M.van – M. vanbaalenii PYR-1; M.spMCS – M. sp. MCS;M.spKMS – M. sp. KMS; M.ulc – M. ulcerans; B.sub – B. subtilis.*

5.4 CONCLUSION

This chapter described sequence based analyses of GntR family transcriptional regulators from *M. smegmatis*. It was observed that in comparison to *M. tuberculosis, M. smegmatis* is equipped with large number of GntR family transcriptional regulators, belonging to four subfamilies (FadR, HutC, MocR, and YtrA). It suggests that the GntR regulatory repertoire of *M. smegmatis* is far more complex than that of *M. tuberculosis*. Indeed, additional GntR regulators possibly control a subset of genes required for adapting to a range of environmental conditions. One of the FadR subfamily regulators shows additional secondary structural element, suggesting a possible origin of a new group within the FadR subfamily. This subfamily of regulators was further divided into two groups' viz., FadR and VanR. Interestingly this analyses also observed that among the GntR proteins two transcriptional regulators were identical in sequence. It is also worth mentioning that in comparison to the lack of MocR subfamily regulators from *M. tuberculosis*, *M. smegmatis* proteome was identified with eight transcriptional regulators belonging to MocR subfamily of transcriptional regulators.

Using subfamily associated features, potential operator sites were identified for many GntR family transcriptional regulators. These DNA sites will be useful for further analyses of cis-regulatory elements belonging to GntR family regulators. These identified regulatory elements may have major implications in identifying DNA targets of these regulators in the whole genome. Besides operator site prediction, identified potential orthologs of *M. smegmatis* GntR in other bacterial genomes will be useful to carryout *in vivo* study of pathogenic species as model organism.

SUMMARY

The availability of whole genome sequences, in addition to many ongoing sequencing projects of many mycobacterial species, makes mycobacterium one of the highly sequenced genera. This wealth of sequence data provides unique opportunity to extract the genome information in order to address better intervention strategies. Although over the past three decades, a number of techniques have been applied to analyze genome sequences. Yet the present genome annotation comprises a large number of uncharacterized or poorly characterized genes. This includes genes encoding GntR family of transcriptional regulators. A major question regarding these GntR regulators is to determine their DNA targets. Since DNA targets of regulatory proteins are short and fuzzy in sequence, identification of these DNA targets is difficult and still remains a quest in biology.

Work presented in the thesis is a combination of computational and experimental approaches which successfully explore the *M. tuberculosis* GntRs and their DNA targets. It extends the knowledge of the repertoire of GntR family of transcriptional regulators and in particular their upstream DNA targets. Identification of the DNA targets could help in understanding how the genome encodes the required proteins responsible for the successful survival of the pathogen in various environmental conditions.

Considering the importance of family/subfamily classification, this work was initiated with the classification of *M. tuberculosis* GntRs into meaningful subfamilies. Among six subfamilies, *M. tuberculosis* genome was observed to encode regulators belonging to only three subfamilies, FadR, HutC and YtrA. All the classified members were observed to exhibit the secondary structural features associated with respective subfamilies. Besides this, an interesting outcome of this classification was the insight into

the operator sites for GntR regulators. It provided clues to search the upstream DNA motifs of these GntRs carrying nucleotide preference characteristics of a subfamily, besides being conserved across the orthologous upstream region. Additionally, in the course of sequence analysis, one of the regulators was observed to be an outcome of a gene duplication and fusion event. As the FadR subfamily of regulators are known to work as dimers, perhaps this regulator might work as a monomer instead of a dimer. Apart from assigning the subfamilies, this appears to be the first report of describing gene duplication and fusion events in GntRs.

After identifying the operator sites for GntRs, my main concern was to carryout *in vitro* assays as an experimental proof for my *in silico* findings. Hence, two of the transcriptional regulators, Rv0586 and Rv0792c, were taken as a model to conduct *in vitro* validations. The corresponding ORFs were cloned and expressed in *E. coli* and the recombinant proteins were purified using metal affinity chromatography. Pure proteins were subjected to electrophoretic mobility shift assay. *In vitro* DNA binding study with both the proteins has shown their ability to recognize specific DNA sequence.

Ortholog identification plays a major role in comparative genomics. It is used as a tool to impart the functional role as well as to compare the upstream conserved regions. Present study identifies orthologs of *M. tuberculosis* GntRs in other mycobacterial species. Further in order to assess the effect of sequence conservation in DNA binding domain, mobility shift assays were carried out. The binding ability of *M. tuberculosis* Rv0586 to the operator sites identified in the upstream region of orthologs was also examined. Successful binding study generated sufficient evidence that the transcriptional regulators bearing similar DNA binding domain may interact with similar DNA targets.

Additionally, this regulator was also compared with its well-characterized structural homolog, FadR, from *E. coli*. Probable critical residues playing role in DNA sequence recognition were identified. Further, recombinant protein Rv0586, was found to possess DNA binding affinity to the known operator site of FadR from *E. coli*. This assay further provided the importance of critical residues within the operator site DNA.

Besides analyzing the GntRs from *M. tuberculosis*, *M. smegmatis* genome was scanned for GntR family of transcriptional regulators. All the GntRs were classified into their respective subfamilies. Interestingly, in the whole *M. smegmatis* proteome, one of the GntR regulators was found in two copies. In comparison to *M. tuberculosis*, *M. smegmatis* was observed to be equipped with a relatively large number of transcriptional regulators belonging to GntR family. This suggests that the GntR regulatory repertoire of *M. smegmatis* is far more complex than in *M. tuberculosis.* Indeed, additional GntR regulators possibly control a subset of genes required for adapting to a range of environmental conditions. One of the FadR subfamily regulator shows additional secondary structural elements suggesting a possible origin of a new group within the FadR subfamily. Identified orthologs from *M. smegmatis* could serve as a model to decipher molecular regulation events taking place in the pathogenic mycobacteria. Potential operator sites were also identified based on the nucleotide recognition preferences of GntR regulators.

In brief, present study takes the first systematic effort to analyze GntR family of transcriptional regulators and their upstream DNA targets in mycobacterial species, particularly *M. tuberculosis*. It also describes for the first time, the operator site conservation in the upstream region of ORFs belonging to GntR family of regulators.

While earlier reports have demonstrated the essentiality of *mce2* operon in the virulence of *M. tuberculosis*, this study has been able to identify the operator site within its upstream region. To strengthen the operator site prediction, two of the novel regulators were subjected to experimental validation. Among all the *M. tuberculosis* GntRs, one of the regulators was also proposed as a product of gene duplication event. Clearly, besides the *M. tuberculosis* GntRs repertoire, this study enhances the overall knowledge of the bacterial transcriptional regulators belonging to GntR family. The present work has made significant advancements towards analyzing the GntRs and their upstream DNA targets in many mycobacterial genomes and particularly it has undertaken detailed analyses of GntR family of transcriptional regulators from *M. tuberculosis*.

BIBLIOGRAPHY

Afanas'ev, M.V., Ikryannikova, L.N., Il'ina, E.N., Sidorenko, S.V., Kuz'min, A.V., Larionova, E.E., Smirnova, T.G., Chernousova, L.N., Kamaev, E.Y., Skorniakov, S.N., Kinsht, V.N., Cherednichenko, A.G. and Govorun, V.M. (2007) Molecular characteristics of rifampicin- and isoniazid-resistant *Mycobacterium tuberculosis* isolates from the Russian Federation. ***J Antimicrob Chemother***. **59:** 1057-1064.

Alekshun, M.N. and Levy, S.B. (1999) Alteration of the repressor activity of MarR, the negative regulator of the *Escherichia coli marRAB* locus, by multiple chemicals in vitro. ***J Bacteriol.*** **181**: 4669-4672.

Allison, S.L. and Phillips, A.T. (1990) Nucleotide sequence of the gene encoding the repressor for the histidine utilization genes of *Pseudomonas putida*. ***J Bacteriol***. **172**: 5470-5476.

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.

Aramaki, H., Yagi, N. and Suzuki, M. (1995) Residues important for the function of a multihelical DNA binding domain in the new transcription factor family of Cam and Tet repressors. ***Protein Eng***. **8**: 1259-1266.

Aravind, L. and Anantharaman, V. (2003) HutC/FarR-like bacterial transcription factors of the GntR family contain a small molecule-binding domain of the chorismate lyase fold. ***FEMS Microbiol Lett.*** **222**:17-23.

Aravind, L., Anantharaman, V., Balaji, S., Babu, M.M. and Iyer, L.M. (2005) The many faces of the helix-turn-helix domain: transcription regulation and beyond. ***FEMS Microbiol Rev***. **29**: 231-262.

Bailey, T.L., Williams, N., Misleh, C. and Li, W.W. (2006) MEME: discovering and analyzing DNA and protein sequence motifs. ***Nucleic Acids Res.*** **34**: W369-373.

Bashton, M. and Chothia, C. (2007) The generation of new protein functions by the combination of domains. ***Structure.*** **15**: 85-99.

Belitsky, B.R. and Sonenshein, A.L. (2002) GabR, a member of a novel protein family, regulates the utilization of gamma-aminobutyrate in *Bacillus subtilis*. ***Mol Microbiol***. **45**: 569-583.

Blondal, K. (2007) Barriers to reaching the targets for tuberculosis control: multidrug-resistant tuberculosis. ***Bull World Health Organ***. **85**: 387-390.

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

Brown, N.L., Stoyanov, J.V., Kidd, S.P. and Hobman, J.L. (2003) The MerR family of transcriptional regulators. ***FEMS Microbiol Rev***. **27**: 145-63.

Bruggemann, H., Hagman, A., Jules, M., Sismeiro, O., Dillies, M.A., Gouyette, C., Kunst, F., Steinert, M., Heuner, K., Coppee, J.Y. and Buchrieser, C. (2006) Virulence strategies for infecting phagocytes deduced from the in vivo transcriptional program of *Legionella pneumophila*. ***Cell Microbiol.*** **8**: 1228-1240.

Buck, D. and Guest, J.R. (1989) Overexpression and site-directed mutagenesis of the succinyl-CoA synthetase of *Escherichia coli* and nucleotide sequence of a gene (g30) that is adjacent to the suc operon. ***Biochem J.*** **260:** 737-47.

Bulyk, M.L., McGuire, A.M., Masuda, N. and Church, G.M. (2004) A motif co-occurrence approach for genome-wide prediction of transcription-factor-binding sites in *Escherichia coli*. ***Genome Res.*** **14**: 201-208.

Busenlehner, L.S., Pennella, M.A. and Giedroc, D.P. (2003) The SmtB/ArsR family of metalloregulatory transcriptional repressors: Structural insights into prokaryotic metal resistance. ***FEMS Microbiol Rev.*** **27**: 131-143.

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

Casali, N., White, A.M. and Riley, L.W. (2006) Regulation of the *Mycobacterium tuberculosis mce1* operon. ***J Bacteriol.*** **188**: 441-449.

Chakhaiyar, P. and Hasnain, S.E. (2004) Defining the mandate of tuberculosis research in a postgenomic era. ***Med Princ Pract.*** **13**: 177-184

Chaturvedi, V., Dwivedi, N., Tripathi, R.P. and Sinha, S. (2007) Evaluation of *Mycobacterium smegmatis* as a possible surrogate screen for selecting molecules active against multi-drug resistant *Mycobacterium tuberculosis*. ***J Gen Appl Microbiol***. **53**: 333-337.

Chen, C.M., Ye, Q.Z., Zhu, Z.M., Wanner, B.L. and Walsh, C.T. (1990) Molecular biology of carbon-phosphorus bond cleavage. Cloning and sequencing of the *phn* (*psiD*) genes involved in alkylphosphonate uptake and C-P lyase activity in *Escherichia coli* B. ***J Biol Chem.*** **265**: 4461-4471.

Cheng, J., Randall, A.Z., Sweredoski, M.J. and Baldi, P. (2005) SCRATCH: a protein structure and structural feature prediction server. ***Nucleic Acids Res***. **33**: W72-76.

Chintu, C. and Mwaba, P. (2005) Tuberculosis in children with human immunodeficiency virus infection. ***Int J Tuberc Lung Dis***. **9**: 477-484.

Cole, S.T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S.V., Eiglmeier, K., Gas, S., Barry, C.E., 3rd, Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Krogh, A., McLean, J., Moule, S., Murphy, L., Oliver, K., Osborne, J., Quail, M.A., Rajandream, M.A., Rogers, J., Rutter, S., Seeger, K., Skelton, J., Squares, R., Squares, S., Sulston, J.E., Taylor, K., Whitehead, S. and Barrell, B.G. (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. ***Nature.*** **393**: 537-544.

Corpet, F. (1988) Multiple sequence alignment with hierarchical clustering. ***Nucleic Acids Res***. **16**: 10881-10890.

Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. (2004) WebLogo: a sequence logo generator. ***Genome Res.*** **14:** 1188-1190.

Cuff, J.A., Clamp, M.E., Siddiqui, A.S., Finlay, M. and Barton, G.J. (1998) ***Bioinformatics.*** **14**: 892-893.

Eddy, S.R. (1998) Profile hidden Markov models. ***Bioinformatics.*** **14**: 755-763

Elofsson, A. and Sonnhammer, E.L. (1999) A comparison of sequence and structure protein domain families as a basis for structural genomics. ***Bioinformatics.*** 15: 480-500

Ernst, J.D., Trevejo-Nunez, G. and Banaiee, N. (2007) Genomics and the evolution, pathogenesis, and diagnosis of tuberculosis. ***Clin Invest***. **117**: 1738-1745

Felsenstein, J. (1989) PHYLIP-phylogeny interface package. ***Cladistics.*** **5**: 164–166.

Fitch, W.M. and Margoliash, E. (1967) Construction of phylogenetic trees. ***Science.*** **155**: 279-284

Friedberg, D., Midkiff, M. and Calvo, J.M. (2001) Global versus local regulatory roles for Lrp-related proteins: *Haemophilus influenzae* as a case study. ***J Bacteriol.*** **183**: 4004-4011.

Frieden, T.R., Sterling, T.R., Munsiff, S.S., Watt, C.J. and Dye, C. (2003) Tuberculosis. ***Lancet*** **362**: 887-899.

Fujita, Y., Fujita, T., Miwa, Y., Nihashi, J. and Aratani, Y. (1986) Organization and transcription of the gluconate operon, *gnt*, of *Bacillus subtilis*. ***J Biol Chem.*** **261**: 13744-13753.

Fulton, D.L., Li, Y.Y., Laird, M.R., Horsman, B.G., Roche, F.M. and Brinkman, F.S. (2006) Improving the specificity of high-throughput ortholog prediction. ***BMC Bioinformatics***. **7**: 270.

Fuqua, W.C., Winans, S.C. and Greenberg, E.P. (1994) Quorum sensing in bacteria: the LuxR-LuxI family of cell density-responsive transcriptional regulators. ***J Bacteriol***. **176**:269-275.

Gao, Y.G., Yao, M., Itou, H., Zhou, Y. and Tanaka, I. (2007) The structures of transcription factor CGL2947 from *Corynebacterium glutamicum* in two crystal forms: a novel homodimer assembling and the implication for effector-binding mode. ***Protein Sci.*** **16**: 1878-86.

Gebhard, S. and Cook, G.M. (2008) Differential regulation of high-affinity phosphate transport systems of *Mycobacterium smegmatis*: identification of PhnF, a repressor of the *phnDCE* operon. ***J Bacteriol.*** **190**: 1335-1343.

Gelfand, M.S. (1999) Recognition of regulatory sites by genomic comparison. ***Res Microbiol***. **150**: 755-771.

Gioffre, A., Infante, E., Aguilar, D., Santangelo, M.P., Klepp, L., Amadio, A., Meikle, V., Etchechoury, I., Romano, M.I., Cataldi, A., Hernandez, R.P. and Bigi, F. (2005) Mutation in *mce* operons attenuates *Mycobacterium tuberculosis* virulence. ***Microbes Infect.*** **7:** 325-334.

Gomez, J.E. and McKinney, J.D. (2004) *M. tuberculosis* persistence, latency, and drug tolerance. ***Tuberculosis*** (Edinb) **84**: 29-44.

Haydon, D.J. and Guest, J.R. (1991) A new family of bacterial regulatory proteins. ***FEMS Microbiol Lett***. **63**: 291-295.

Hillerich, B. and Westpheling, J. (2006) A new GntR family transcriptional regulator in streptomyces coelicolor is required for morphogenesis and antibiotic production and controls transcription of an ABC transporter in response to carbon source. ***J Bacteriol.*** **188**: 7477-7487.

Hosoya, S., Yamane, K., Takeuchi, M. and Sato, T. (2002) Identification and characterization of the Bacillus subtilis D-glucarate/galactarate utilization operon *ycbCDEFGHJ*. ***FEMS Microbiol Lett.*** **210**: 193-199.

Jacob, F., Perrin, D., Sanchez, C. and Monod, J. (1960) Operon: a group of genes with the expression coordinated by an operator. ***C R Hebd Seances Acad Sci***. **250:** 1727-1729.

Jacob, F., Ullman, A. and Monod, J. (1964) The promotor, a genetic element necessary to the expression of an operon. ***C R Hebd Seances Acad Sci***. **258**: 3125-3128.

Jain, A. and Mondal, R. (2008) Extensively drug-resistant tuberculosis: current challenges and threats. ***FEMS Immunol Med Microbiol***. **53**: 145-150.

Kaplan, T., Friedman, N. and Margalit, H. (2005) Ab initio prediction of transcription factor targets using structural knowledge. ***PLoS Comput Biol***. **1**: e1.

Karin, M. (1990) Too many transcription factors: positive and negative interactions. ***New Biol***. **2**: 126-131.

Karmirantzou, M. and Hamodrakas, S.J. (2001) A Web-based classification system of DNA-binding protein families. ***Protein Eng.*** **14**:465-472.

Keane, J. and Bresnihan, B. (2008) Tuberculosis reactivation during immunosuppressive therapy in rheumatic diseases: diagnostic and therapeutic strategies. ***Curr Opin Rheumatol.*** 20: 443-449.

Kelley, L.A., MacCallum, R.M. and Sternberg, M.J. (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. ***J Mol Biol***. **299**: 499-520.

Korner, H., Sofia, H.J. and Zumft, W.G. (2003) Phylogeny of the bacterial superfamily of Crp-Fnr transcription regulators: exploiting the metabolic spectrum by controlling alternative gene programs. ***FEMS Microbiol Rev.*** **27**: 559-592.

Lalloo, U.G. and Pillay, S. (2008) Managing tuberculosis and HIV in sub-Sahara Africa. ***Curr HIV/AIDS Rep***. 5:132-9.

Latchman, D.S. (1997) Transcription factors: an overview. ***Int J Biochem Cell Biol.*** **29**: 1305-1312.

Lee, H.Y., An, J.H. and Kim, Y.S. (2000) Identification and characterization of a novel transcriptional regulator, MatR, for malonate metabolism in *Rhizobium leguminosarum bv. trifolii*. ***Eur J Biochem***. **267**:7224-7230.

Lee, M.H., Scherer, M., Rigali, S. and Golden, J.W. (2003) PlmA, a new member of the GntR family, has plasmid maintenance functions in *Anabaena sp. strain PCC 7120*. ***J Bacteriol.*** **185**: 4315-4325.

Locht, C., Rouanet, C., Hougardy, J.M. and Mascart, F. (2007) How a different look at latency can help to develop novel diagnostics and vaccines against tuberculosis. ***Expert Opin Biol Ther.*** **7**: 1665-1677.

Lovell, S.C., Davis, I.W., Arendall ,W.B. 3rd., de Bakker, P.I., Word, J.M., Prisant, M.G., Richardson, J.S. and Richardson, D.C. (2003) Structure validation by Calpha geometry: phi,psi and Cbeta deviation. ***Proteins.*** **50**: 437-450.

Luscombe, N.M. and Thornton, J.M. (2002) Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. ***J Mol Biol***. **320**: 991-1009.

Ly, L.H. and McMurray, D.N. (2008) Tuberculosis: vaccines in the pipeline. ***Expert Rev Vaccines***. **7**: 635-650.

Mackintosh, C.G., de Lisle, G.W., Collins, D.M. and Griffin, J.F. (2004) Mycobacterial diseases of deer. ***N Z Vet J***. **52:** 163-174.

Magarvey, N., He, J., Aidoo, K.A. and Vining, L.C. (2001) The *pdx* genetic marker adjacent to the chloramphenicol biosynthesis gene cluster in *Streptomyces venezuelae ISP5230*: functional characterization. ***Microbiology***. **147**: 2103-2112.

Marsden, R.L., Ranea, J.A., Sillero,, A., Redfern, O., Yeats, C., Maibaum, M, Lee, D., Addou, S., Reeves, G.A., Dallman, T.J. and Orengo, C.A. (2006) Exploiting protein structure data to explore the evolution of protein function and biological complexity. ***Philos Trans R Soc Lond B Biol Sci.*** **361:** 425-440.

Martin, R.G. and Rosner, J.L. (2001) The AraC transcriptional activators. ***Curr Opin Microbiol***. **4:**132-137.

Martinez-Hackert, E. and Stock, A.M. (1997) Structural relationships in the OmpR family of winged-helix transcription factors. ***J Mol Biol.*** **269**: 301-12.

Martins, M., Viveiros, M. and Amaral, L. (2008) The TB laboratory of the future: macrophage-based selection of XDR-TB therapeutics. ***Future Microbiol.*** **3:** 135-144.

Matthijs, S., Koedam, N., Cornelis, P. and De Greve, H. (2000) The trehalose operon of *Pseudomonas fluorescens ATCC 17400*. ***Res Microbiol***. **151**: 845-851.

McCue, L., Thompson, W., Carmack, C., Ryan, M.P., Liu, J.S., Derbyshire, V. and Lawrence, C.E. (2001) Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. ***Nucleic Acids Res***. **29**: 774-782.

McCue, L.A., Thompson, W., Carmack, C.S. and Lawrence, C.E. (2002) Factors influencing the identification of transcription factor binding sites by cross-species comparison. ***Genome Res***. **12**: 1523-1532.

Minezaki, Y., Homma, K. and Nishikawa, K. (2005) Genome-wide survey of transcription factors in prokaryotes reveals many bacteria-specific families not found in archaea. ***DNA Res.*** **12**: 269-280.

Miziara, M.N., Riggs, P.K. and Amaral, M.E. (2004) Comparative analysis of noncoding sequences of orthologous bovine and human gene pairs. ***Genet Mol Res***. **3:** 465-473.

Morett, E. and Segovia, L. (1993) The sigma 54 bacterial enhancer-binding protein family: mechanism of action and phylogenetic relationship of their functional domains. ***J Bacteriol.*** **175**: 6067-6074.

Mota, L.J., Tavares, P. and Sa-Nogueira, I. (1999) Mode of action of AraR, the key regulator of L-arabinose metabolism in *Bacillus subtilis****.*** ***Mol Microbiol.*** **33**: 476-489.

Nerlich, A.G., Haas, C.J., Zink, A., Szeimies, U. and Hagedorn, H.G. (1997) Molecular evidence for tuberculosis in an ancient Egyptian mummy. ***Lancet.*** **350**: 1404.

Nguyen, C.C. and Saier, M.H., Jr. (1995) Phylogenetic, structural and functional analyses of the LacI-GalR family of bacterial transcription factors. ***FEBS Lett.*** **377**: 98-102.

Nunez, M.F., Pellicer, M.T., Badia, J., Aguilar, J. and Baldoma, L. (2001) The gene *yghK* linked to the *glc* operon of *Escherichia coli* encodes a permease for glycolate that is structurally and functionally similar to L-lactate permease. ***Microbiology***. **147**:1069-1077.

Okunade, A.L., Elvin-Lewis, M.P. and Lewis, W.H. (2004) Natural antimycobacterial metabolites: current status. ***Phytochemistry.*** 65: 1017-1032.

Orengo, C.A. and Thornton, J.M. (2005) Protein families and their evolution-a structural perspective. ***Annu Rev Biochem.*** **74**: 867-900.

Page, R.D. **(**1996**)** TreeView: an application to display phylogenetic trees on personal computers. ***Comput Appl Biosci.*** **12**: 357-358.

Pellicer, M.T., Fernandez, C., Badia, J., Aguilar, J., Lin, E.C. and Baldom, L. (1999) Cross-induction of glc and ace operons of *Escherichia coli* attributable to pathway intersection. Characterization of the *glc* promoter. ***J Biol Chem.*** **274:** 1745-1752.

Pelly, T., Moore, D.A., Gilman, R. and Evans, C. (2004) Recent tuberculosis advances in Latin America. ***Curr Opin Infect Dis.*** **17**: 397-403.

Perez-Rueda, E. and Collado-Vides, J. (2000) The repertoire of DNA-binding transcriptional regulators in *Escherichia coli K-12*. ***Nucleic Acids Res***. **28**: 1838-1847.

Perez-Rueda, E. and Collado-Vides, J. (2001) Common history at the origin of the position-function correlation in transcriptional regulators in archaea and bacteria. ***J Mol Evol.*** **53:** 172-179.

Poupin, P., Ducrocq, V., Hallier-Soulier, S. and Truffaut, N. (1999) Cloning and characterization of the genes encoding a cytochrome P450 (PipA) involved in piperidine and pyrrolidine utilization and its regulatory protein (PipR) in *Mycobacterium smegmatis mc2155*. ***J Bacteriol***. **181**: 3419-3426.

Pritsker, M., Liu, Y.C., Beer, M.A. and Tavazoie, S. (2004) Whole-genome discovery of transcription factor binding sites by network-level conservation. ***Genome Res.*** **14**: 99-108.

Ptashne, M. and Gann, A. (1997) Transcriptional activation by recruitment. ***Nature***. **386**: 569-577.

Quail, M.A. and Guest, J.R. (1995) Purification, characterization and mode of action of PdhR, the transcriptional repressor of the pdhR-aceEF-lpd operon of *Escherichia coli*. ***Mol Microbiol.*** **15**: 519-529.

Quail, M.A., Dempsey, C.E. and Guest, J.R. (1994) Identification of a fatty acyl responsive regulator (FarR) in *Escherichia coli*. ***FEBS Lett***. **356**: 183-187.

Raman, N., Black, P.N. and DiRusso, C.C. (1997) Characterization of the fatty acid-responsive transcription factor FadR. Biochemical and genetic analyses of the native conformation and functional domains. ***J Biol Chem.*** **272**: 30645-30650.

Ramos, J.L., Martinez-Bueno, M., Molina-Henares, A.J., Teran, W., Watanabe, K., Zhang, X., Gallegos, M.T., Brennan, R. and Tobes, R. (2005) The TetR family of transcriptional repressors. ***Microbiol Mol Biol Rev.*** **69**: 326-356.

Ranjan, S., Seshadri, J., Vindal, V., Yellaboina, S. and Ranjan, A. (2006) iCR: a web tool to identify conserved targets of a regulatory protein across the multiple related prokaryotic species. ***Nucleic Acids Res.*** **34**: W584-587.

Reizer, A., Deutscher, J., Saier, M.H., Jr. and Reizer, J. (1991) Analysis of the gluconate (*gnt*) operon of *Bacillus subtilis*. ***Mol Microbiol.*** **5**: 1081-1089.

Rigali, S., Derouaux, A., Giannotta, F. and Dusart, J. (2002) Subdivision of the helix-turn-helix GntR family of bacterial regulators in the FadR, HutC, MocR, and YtrA subfamilies. ***J Biol Chem***. **277**: 12507-12515.

Rigali, S., Schlicht, M., Hoskisson, P., Nothaft, H., Merzbacher, M., Joris, B. and Titgemeyer, F. (2004) Extending the classification of bacterial transcription factors beyond the helix-turn-helix motif as an alternative approach to discover new cis/trans relationships. ***Nucleic Acids Res***. **32**: 3418-3426.

Rosinski, J.A. and Atchley, W.R. (1999) Molecular evolution of helix-turn-helix proteins. ***J Mol Evol.*** **49**: 301-309.

Saghatelian, A. and Cravatt, B.F. (2005) Assignment of protein function in the postgenomic era. ***Nat Chem Biol***. **1**: 130-142.

Saier, M.H., Jr. (1996) Phylogenetic approaches to the identification and characterization of protein families and superfamilies. ***Microb Comp Genomics.*** **1:** 129-150.

Sajduda, A., Brzostek, A., Poplawska, M., Augustynowicz-Kopec, E., Zwolska, Z., Niemann, S., Dziadek, J. and Hillemann, D. (2004) Molecular characterization of rifampin- and isoniazid-resistant *Mycobacterium tuberculosis* strains isolated in Poland. ***J Clin Microbiol***. **42**: 2425-2431.

Sali, A. and Blundell, T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. ***J Mol Biol.*** **234**: 779-815.

Sambrook, J., Fritsch, E. F. and Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual, 2nd Ed.,* Cold Spring Harbor Laboratory, Cold Spring Harbor, NY

Sandelin, A. and Wasserman, W.W. (2004) Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. ***J Mol Biol.*** **338**: 207-215.

Saqi, M.A. and Wild, D.L. (2005) Expectations from structural genomics revisited: an analysis of structural genomics targets. ***Am J Pharmacogenomics***. **5**: 339-342.

Schell, M.A. (1993) Molecular biology of the LysR family of transcriptional regulators. ***Annu Rev Microbiol***. **47**: 597-626.

Schmidt, V., Schneider, S., Schlomer, J., Krautwald-Junghanns, M.E. and Richter, E. (2008) Transmission of tuberculosis between men and pet birds: a case report. ***Avian Pathol***. 1-4.

Schock, F. and Dahl, M.K. (1996) Expression of the *tre* operon of *Bacillus subtilis* 168 is regulated by the repressor TreR. ***J Bacteriol.*** **178**: 4576-4581.

Shimono, N., Morici, L., Casali, N., Cantrell, S., Sidders, B., Ehrt, S. and Riley, L.W. (2003) Hypervirulent mutant of *Mycobacterium tuberculosis* resulting from disruption of the *mce1* operon. ***Proc Natl Acad Sci U S A***. **100**: 15918-15923.

Siddharthan, R., Siggia, E.D. and van Nimwegen, E. (2005) PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. ***PLoS Comput Biol.*** 1(7).

Siggers, T.W., Silkov, A. and Honig, B. (2005) Structural alignment of protein--DNA interfaces: insights into the determinants of binding specificity. ***J Mol Biol.*** **345**:1027-4105.

Singla, R., Al-Sharif, N., Al-Sayegh, M., Osman, M. and Shaikh, M.A. (2003) Prevalence of resistance to antituberculosis drugs in Riyadh and a review of previous reports. ***Ann Saudi Med***. **23**: 143-147.

Stormo, G.D. and Tan, K. (2002) Mining genome databases to identify and understand new gene regulatory systems. ***Curr Opin Microbiol.*** **5**: 149-153.

Sulochana, S., Narayanan, S., Paramasivan, C.N., Suganthi, C. and Narayanan, P.R. (2007) Analysis of fluoroquinolone resistance in clinical isolates of *Mycobacterium tuberculosis* from India. ***J Chemother.*** **19**: 166-171.

Sung, M.H., Tanizawa, K., Tanaka, H., Kuramitsu, S., Kagamiyama, H., Hirotsu, K., Okamoto, A., Higuchi, T. and Soda, K. (1991) Thermostable aspartate aminotransferase from a thermophilic *Bacillus species*. Gene cloning, sequence determination, and preliminary x-ray characterization. ***J Biol Chem.*** **266**: 2567-2572.

Sunnarborg, A., Klumpp, D., Chung, T. and LaPorte, D.C. (1990) Regulation of the glyoxylate bypass operon: cloning and characterization of iclR. ***J Bacteriol.*** **172**: 2642-2649.

Tatusov, R.L., Koonin, E.V. and Lipman, D.J. (1997) A genomic perspective on protein families. ***Science***. **278**: 631-637.

Taylor, J.S. and Raes, J. (2004) Duplication and divergence: the evolution of new genes and old ideas. ***Annu Rev Genet.*** **38**: 615-643.

Thieffry, D., Salgado, H., Huerta, A.M. and Collado-Vides, J. (1998) Prediction of transcriptional regulatory sites in the complete genome sequence of *Escherichia coli K-12*. ***Bioinformatics*****. 14:** 391-400.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgins, D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. ***Nucleic Acids Res.*** **25**: 4876-4882.

Todd, A.E., Marsden, R.L., Thornton, J.M. and Orengo, C.A. (2005) Progress of structural genomics initiatives: an analysis of solved target structures. ***J Mol Biol***. **348:** 1235-1260.

Tsoka, S. and Ouzounis, C.A. (2000) Recent developments and future directions in computational genomics. ***FEBS Lett.*** **480**: 42-48.

Umubyeyi, A., Rigouts, L., Shamputa, I.C., Dediste, A., Struelens, M. and Portaels, F. (2008) Low levels of second-line drug resistance among multidrug-resistant *Mycobacterium tuberculosis* isolates from Rwanda. ***Int J Infect Dis.*** **12**: 152-156.

van Aalten, D.M., DiRusso, C.C. and Knudsen, J. (2001) The structural basis of acyl coenzyme A-dependent regulation of the transcription factor FadR. ***Embo J***. **20**: 2041-2050.

van Rooijen, R.J. and de Vos, W.M. (1990) Molecular cloning, transcriptional analysis, and nucleotide sequence of lacR, a gene encoding the repressor of the lactose phosphotransferase system of *Lactococcus lactis*. ***J Biol Chem.*** **265**: 18499-18503.

Vogel, C. and Chothia, C. (2006) Protein family expansions and biological complexity. ***PLoS Comput Biol.*** **2:** e48.

Wang, R. and Marcotte, E.M. (2008) The proteomic response of *Mycobacterium smegmatis* to anti-tuberculosis drugs suggests targeted pathways. ***J Proteome Res.*** 7: 855-865.

Wu, C.H., Huang, H., Yeh, L.S. and Barker, W.C. (2003) Protein family classification and functional annotation. ***Comput Biol Chem.*** **27:** 37-47.

Xu, Y., Heath, R.J., Li, Z., Rock, C.O. and White, S.W. (2001) The FadR.DNA complex. Transcriptional control of fatty acid metabolism in *Escherichia coli*. ***J Biol Chem***. **276**: 17373-17379.

Yellaboina, S., Ranjan, S., Vindal, V. and Ranjan, A. (2006) Comparative analysis of iron regulated genes in mycobacteria. ***FEBS Lett.*** **580:** 2567-2576.

Yellaboina, S., Seshadri, J., Kumar, M.S. and Ranjan, A. (2004) PredictRegulon: a web server for the prediction of the regulatory protein binding sites and operons in prokaryote genomes. ***Nucleic Acids Res.*** **32**: W318-320.

Yoshida, K.I, Fujita, Y. and Ehrlich, S.D. (2000) An operon for a putative ATP-binding cassette transport system involved in acetoin utilization of *Bacillus subtilis*. ***Bacteriol.*** **182**: 5454-5461.

Young, C.L., Barker, W.C., Tomaselli, C.M. and Dayhoff, M.O. (1979) From Atlas of Protein Sequence and Structure Volume 5. Issue Suppl 3 Edited by: Dayhoff MO. ***National Biochemical Foundation, Silver Spring, MD.*** 73-93.

Young, D.B. (2001) A post-genomic perspective. ***Nat Med.*** **7**: 11-13.

Zhou, D. and Yang, R. (2006) Global analysis of gene transcription regulation in prokaryotes. ***Cell Mol Life Sci***. **63**: 2260-2290.

Zimmerman, M.R. (1979) Pulmonary and osseous tuberculosis in an Egyptian mummy. ***Bull N Y Acad Med***. **55:** 604-608.

Zink, A.R., Sola, C., Reischl, U., Grabner, W., Rastogi, N., Wolf, H. and Nerlich, A.G. (2003) Characterization of *Mycobacterium tuberculosis* complex DNAs from Egyptian mummies by spoligotyping. ***J Clin Microbiol***. **41**: 359-367.

Ziskind, B. and Halioua, B. (2007) Tuberculosis in ancient Egypt. ***Rev Mal Respir.*** **24**: 1277-1283.

APPENDIX

BioSuite: A comprehensive bioinformatics software package (A unique industry–academia collaboration)

*The NMITLI-BioSuite Team**

Keywords: Bioinformatics, BioSuite, industry–academia collaboration, software.

THE last decade has witnessed an exponential growth of information in the field of biological macromolecules such as proteins and nucleic acids and their interactions with other molecules. Computational analysis and predictions based on such information are increasingly becoming an essential and integral part of modern biology. With rapid advances in the area, there is a growing need to develop versatile bioinformatics software packages, which are efficient and incorporate the latest developments in this field. In view of this, the Council of Scientific and Industrial Research, India, undertook an initiative to promote a unique industry–academia collaboration, to develop a comprehensive bioinformatics software package, under its New Millennium Initiative for Technology Leadership in India programme. BioSuite, a product of that effort, has been developed by Tata Consultancy Services who took the primary coding responsibility with significant backing from a large academic community who participated on advisory roles through the project period.

BioSuite integrates the functions of macromolecular sequence and structural analysis, chemoinformatics and algorithms for aiding drug discovery. The suite organized into four major modules, contains 79 different programs, making it one of the few comprehensive suites that caters to a major part of the spectrum of bioinformatics applications. The four major modules, (a) Genome and proteome sequence analysis, (b) 3D modelling and structural analysis, (c) Molecular dynamics simulations and (d) Drug design, are made available through a convenient graphics-user interface along with adequate documentation and tutorials. The unique partnership with academia has also ensured that the best available methodology has been adopted for each of the 79 programs, which has been thoroughly evaluated in several stages, leading to high scientific value of the suite. The software, apart from having the advantage of running on a Linux platform on a personal computer, is also flexible, modular, and allows for newer algorithms to be plugged into the overall framework. The package will be valuable for high quality academic research, industrial research and development and for teaching purposes, both locally within the country as well as in the international arena. A full list of the programs as well as their example usage can be found at http://www.atc.tcs.co.in/bioinfo/publications/biosuite_paper.pdf.

Background

Genesis of BioSuite

The Council of Scientific and Industrial Research (CSIR), Government of India, proposed a new millennium initiative for technology leadership in India (NMITLI), in 2000, wherein India could acquire leadership positions in key technology areas (NMITLI). Development of versatile,

The team consists of *Tata Consultancy Services*: M. Vidyasagar, S. Mande, S. Rajgopal, B. Gopalkrishnan, S. T. P. T. Srinivas, C. Uma Maheswara Rao, T. Kathiravan, K. Mastanarao, S. Narendranath, S. Rohini, A. Irshad, T. Murali, C. Subrahmanyam, T. Mona, S. Sankha, V. Priya, D. Suman, V. V. Raja Rao, P. Nageswara Rao, R. Issaac, H. Yashodeep, B. Arundhoti, G. Nishant, S Jignesh, K. S. Chaitanya, S. P. V. Prasad Reddy; *Bose Institute*: P. Chakraborty; *Centre for DNA Fingerprinting and Diagnosis*: S. E. Hasnain, S. Mande, A. Nagarajaram, A. Ranjan, M. S. Acharya, M. Anwaruddin, S. K. Arun, Gyanrajkumar, D. Kumar, S. Priya, S. Ranjan, B. R. Reddi, J. Seshadri, P. Sravan Kumar, S. Swaminathan, P. Umadevi, V. Vindal, S. Vijaykrishnan; *Central Drug Research Institute*: A. K. Saxena, A. Dixit, P. Prathipati, S. K. Kashaw; *Indian Institute of Chemical Biology*: C. Mandal, S. Bag; *Indian Institute of Science*: N. Balakrishnan, M. Bansal, N. R. Chandra*, M. R. N. Murthy, S. Ramakumar, K. Sekar, N. Srinivasan, K. Suguna, S. Vishveshwara*, R. Anandhi, Bhadra, S. Das, P. Hansia, S. Hariharaputran, J. Jeyakani, R. Karthikeyan, R. K. Pandey, C. S. Swamy, B. Vasanthakumar; *Indian Institute of Technology Bombay*: P. V. Balaji, R. Y. Patel; *Indian Institute of Technology Delhi*: B. Jayaram, S. A. Shaikh; *Indian Institute of Technology Kharagpur*: P. P. Chakrabarti, A. Banerjee, A. Chakrabarti; *Indian Statistical Institute*: R. L. Karandikar, Delhi and P. Chaudhuri, Kolkata; *Institute of Microbial Technology*: G. P. S. Raghava, A. Ghosh; *Institute of Bioinformatics and Applied Biotechnology*: M. Bansal, N. Paramsivam; *Institute of Genomics and Integrative Biology*: S. K. Brahmachari, D. Dash, C. Balasubramaniam, A. Basu, P. Biswas, M. Hariharan, R. Mathur, K. S. Sandhu, V. Scaria, R. Shankar; *International Institute of Information Technology*: P. J. Narayanan, V. Jain, Nirnimesh; *Madurai Kamaraj University*: S. Krishnaswamy, V. Alaguraj, R. Marikkannu, A. V. S. K. Mohan Katta, N. Krishnan, K. V. Srividhya, P. J. Eswari; *National Institute of Pharmaceutical Education and Research*: P. V. Bharatam, P. Iqbal; *Saha Institute of Nuclear Physics*: D. Bhattacharyya; *University of Hyderabad*: G. R. Desiraju, J. J. Kumar, M. Ravikumar; *University of Madras*: M. Gautham, P. A. Prasad and D. Bharanidharan. *For correspondence. (nchandra@physics.iisc.ernet.in; sv@mbu.iisc.ernet.in)

Table 1. Roles played by different groups for ensuring successful development of BioSuite

Algorithm design, Code writing, Coding quality checks, Graphic-user interfaces and performance benchmarking	Tata Consultancy Services, team led by M. Vidyasagar Sharmila Mande and Rajagopal Srinivasan
Algorithm/module design suggestions and scientific evaluations	Academic partners
Project monitoring committee	R. Narasimha, G. Padmanaban, G. R. Desiraju, D. Balasubramanian
Project co-ordination	Yogeswara Rao and Vibha Sawhney, CSIR
Project funding	CSIR, NMITLI Scheme, Govt of India
Manuscript preparation	Coordinated by Nagasuma Chandra and Saraswathi Vishveshwara, IISc

portable bioinformatics software was recognized as one such area, taking into account the expertise available in the Indian academic community. Such a project, promoted by CSIR, was therefore flagged off in partnership with the industry, where Tata Consultancy Services (TCS) took the major responsibility of developing the BioSuite software with significant scientific support from the major academic institutions in the country (Table 1). The objectives of the project have been to develop indigenously, a set of software tools, that would assist the academic research, R&D and applications in industry, in the rapidly emerging field of bioinformatics and rational drug design.

The need for such a software suite is exemplified by two main factors: (a) increase in bioinformatics activities at all levels – education, research, industry, rapid growth of primary data and methods in computational biology and (b) limitations of existing suites – such as very high cost and not being comprehensive under a single framework, as discussed later. A team of 35 members from TCS worked on this project.

Mode of operation

To ensure the smooth functioning of the project, the following management structure was put in place: (a) *A Monitoring Committee*, monitored the progress of the project through periodic meetings with TCS and the academic partners providing timely focus, (b) *A Steering Committee*, consisting of scientists from academic institutions and TCS, coordinated the activities of the group, (c) *Domain experts and consultants*, consisting of all academic partners, helped in arriving at a basic structure for the suite. Given the large size of the group and the involvement of 18 institutions, the efforts from CSIR and the monitoring committees have played a significant role in fostering the unique partnership to ensure success of this project. The domain experts have advised TCS on the individual modules and individual programs required in each module, identified appropriate algorithms at each step, as also the features required for each program, as per the current research trends and requirements. Further, (d) *a team of pseudo-code developers of six people at* TCS, have interacted with domain experts and directed their (e) in-house *team of code developers*, consisting of 27 software engineers, who have written the actual code. The (f) *Software Project Management Committee* from TCS has ensured the overall activities at that end and ensured appropriate benchmarking and in-house quality checks from the software perspective. The scientific performance of the codes developed has been further evaluated by the academic partners, who have tested and reported bugs to Project Management Committee, after which the codes have been improved/modified where required. Further, an autonomous assessment of the suite has been obtained by an independent expert in the area.

Operational schedules

A glimpse of the schedules and the various milestones reached are given below: (a) Identification of the modules, the required programs in each module and the appropriate algorithm(s) for each program, was completed in the first four months, following which a (b) Software Requirement Specification (SRS) document was developed and reviewed in the next two months. Next, the pseudo-codes were developed in about five months and converted into final code in the next 12 months. In parallel with alpha-testing that was carried out simultaneously with code development, the documentation and creation of a user guide took about seven months. Bug reporting and bug fixes were carried out in iterations through the testing phases and a beta-version was produced by June 2004, taking a total of 24 months. Evaluation and bug fixing of this version was carried out in five months, leading to the first full version, soft-launched in July 2004 and product released in December 2004.

Overview of the organization of the suite

The entire package, consisting of 79 different programs is organized into four major modules, all linked through three common graphics-user interface (GUI) workbenches, as illustrated in Figure 1. The four modules are: (a) Genome and sequence analysis, (b) 3D modelling and structure analysis, (c) Molecular dynamics simulations and (d) Drug design. They are accessible through central GUIs for file handling, sequence and structure windows.

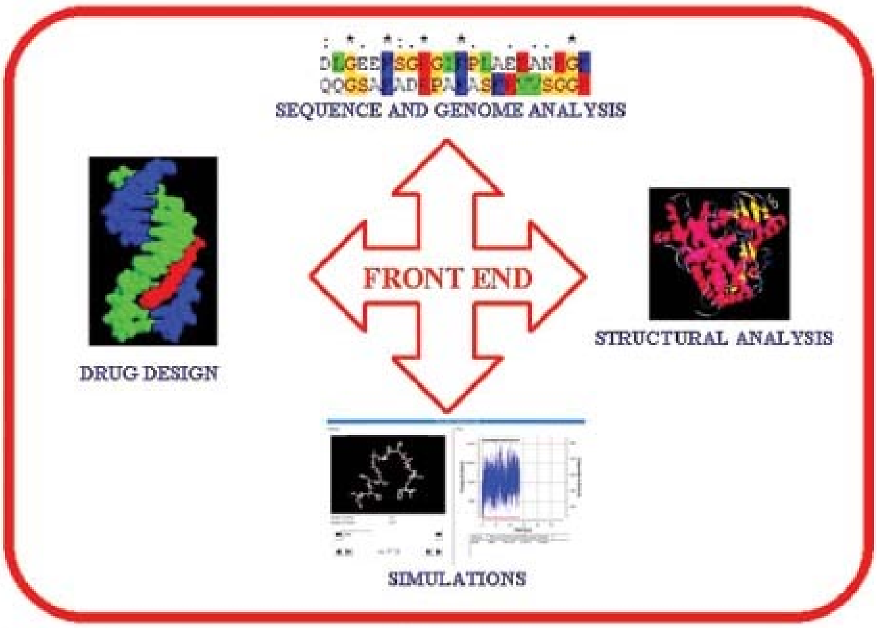


Figure 1. Modular organization of BioSuite.

Table 2. Examples of programs contained in the modules

Sequence and genome analysis
Genome sequence assembly and EST mapping[1], ePCR[2], ORF prediction[3], Intron–exon boundary[4], Database search[5] and sequence alignments (pairwise[6,7]; multiple[8]; whole genome alignment[9]); Motifs and patterns (restriction sites[10], motif building and searching[11]; primer and probe design[12]); RNA and protein secondary structure and transmembrane prediction[13–15]; Domain building and searching[16], gene order[17], unique genes[18]; Phylogenetic analysis, tree construction, evolutionary distance estimation and profiling[19–21].

Structural analysis
Nucleic acid analysis[22], protein structure quality check[23], symmetry-related molecules, structural superposition[24], interactions[25], homology modelling and threading[26]; Fold classification[27]; Molecular surface area, solvent accessible surface area and volume[28]; Binding site detection (PASS[29]; ET[30]).

Simulations
Energy minimizations (steepest descent[31] and conjugate gradient minimizers[32]; forcefields[33]); Electrostatic potential maps[34,35]; Molecular dynamics[36,37]; MD analysis of various trajectories, RMSD, average position and plots of system properties.

Drug design
Structure-based design using protein–ligand docking[38]; Conformation search[39]; Steric and electrostatic ligand alignment[40]; QSAR with over 80 descriptors and regression analysis; Pharmacophore identification and pharmacophore-based search[41,42].

Table 2 lists the important programs in each module. A full list of the modules as well as example outputs of the individual programs can be found at http://www.atc.tcs.co.in/bioinfo/publications/biosuite_paper.pdf. Combination of the four modules makes BioSuite a comprehensive package, covering much of the activities of the bioinformatics spectrum, starting from genome sequences to individual and multiple protein sequences, different levels of structure prediction, analysis of the structures, molecular mechanics calculations, molecular dynamics simulations, chemoinformatics and finally integration with the application of the sequence and structural analyses in rational drug design through algorithms for QSAR, pharmacophore identification and docking processes, for facilitating rational drug design.

Choice of algorithms and coding methods

Choice of algorithms was discussed extensively with academic partners and the latest concepts available in the literature have been adopted wherever possible. For some programs, more than one algorithm has also been implemented, to suit the current research trends of using multiple methods and studying consensus predictions. In general, about two scientists have analysed and chosen a particular algorithm for a particular purpose. Table 2 indicates the algorithms chosen for each of the programs. The knowledge and description of each of the algorithms have been captured into detailed SRS documents by the pseudo-code development team at TCS through extensive interactions with the academic partners as well as with a detailed study of the appropriate literature. The pseudo-code generated for each algorithm and its linkages have been developed using formal software engineering methods, so as to guarantee correctness. The pseudo-code was then converted into actual code by another set of programers who have ensured strict adherence to well-established quality processes such as CMMi Level 5.

All codes have been written in C^{++}. A total of 170 algorithms and about 100 QSAR descriptor calculators have been implemented in 79 programs, with about 700,000 lines of code. The suite is modular, which not only facilitates seamless updation of the modules but also enables integration of new programs by the end users.

Description of the modules

The functionalities of the programs contained within the four major modules are briefly described below.

Genome and proteome sequence analysis

This module deals with the applications relating to the analysis of the nucleic acid and protein sequences, not only of individual molecules, but also of complete genome and proteome sequences. This module would enable researchers to annotate genomes, predict protein secondary structures, derive a phylogenetic relationship among organisms and compare two genomes for similarities at the gene or protein level, along with a range of other applications. This module is further divided into four sub-modules: Sequence analysis, genome analysis, Comparative genomics and Utilities.

Sequence analysis of individual molecules is enabled through the sequence analysis modules, while the programs in the 'Genome analysis' sub-module enable comparison and analysis of full genomes and proteomes. Two database searching tools, BLAST and PSI-BLAST are interfaced with the suite, that will enable searching databases to identify a given sequence or find conserved domains or even find distantly related homologues from some other species. An option of building custom-made databases is also provided. Alignment of sequences, a crucial task in sequence analysis, is provided for, through two well-established algorithms for global and local alignments using dynamic programing algorithms (Needleman–Wunsch and Smith–Waterman). Further, a hierarchical clustering-based multiple alignment algorithm (ClustalW) is included for aligning a set of sequences. Besides, pattern identification and matching tasks such as finding composition, inverted repeats, DNA structure motifs, restriction site analysis and repeat analysis, are part of this module.

Algorithms for secondary structure prediction including transmembrane region detection, RNA structure prediction and analysis are also part of this module. The secondary structure prediction algorithms were trained (or retrained as appropriate) using a comprehensive dataset containing 731 high resolution protein structures (with resolutions ≤ 2 Å) that comprise a non-redundant dataset (redundancy has been removed through sequence comparisons, using a similarity cut-off of 25% with the Blosum62 substitution matrix). Use of a large dataset in training the prediction algorithms ensures high prediction accuracy. A comprehensive biophysical parameter computation ability has also been built into BioSuite, by extracting 36 different physico-chemical properties for protein molecules from the dataset and subsequently using them as training-sets in the prediction algorithms. Algorithms for predicting isoelectric point, peptide cleavage patterns, B-cell antigenicity from protein sequences are also included in this module. Yet another useful feature of this module is the domain building and analysing functionality. Programs are available for identifying domains, building consensus domain sequences, calibrating them and searching across a database. Hidden Markov models using sequence profiles are used for these purposes. In addition, the module has programs for studying molecular evolution, to cluster groups of sequences based on several criteria and to compute phylogenetic trees as well as to calculate evolutionary distances. Finally, algorithms for gene finding, gene assembly, probe and primer design, vector trimming and EST analysis are also part of this module. Two examples of using the programs of this module are illustrated in Figure 2 *a* and *b*.

3D Modelling and analysis

The 3D modelling and analysis module has capabilities to build, analyse and predict three-dimensional structures of macromolecules and macromolecular complexes. This module is further subdivided into the following sub-modules: (a) Homology modelling, (b) Threading, (c) Building proteins, (d) Building nucleic acids, (e) Building carbohydrates, (f) Generation of symmetry-related molecules, (g) Structural superposition, (h) Surfaces and volumes, (i) Binding site analysis, (j) Nucleic acid analysis, (k) Interactions, (l) Quality check, and (m) Fold detection. Example snapshots are shown in Figure 2 *c* and *d*.

Building the models of protein molecules by predicting their three-dimensional structures by comparative modelling techniques are enabled through the first two sub-modules, for which six algorithms are available that incorporate the latest concepts in these areas. Building nucleic acids and carbohydrates using geometric information is enabled through the building modules. A notable feature of the builder programs is the incorporation of 17 geometrical templates for nucleic acids and 12 templates for carbohydrates providing a handle to address the stereo-chemical variability in a large number of sugars. Several programs that can address visualization and analysis of crystallographically derived structures are also included in this module. For example, a lattice assembly of a protein molecule, as seen in its crystal structure can be generated easily. Structure validation tools for proteins and nucleic acids are enabled through the quality check programs. Extensive analysis is possible through the analysis and interactions functions, that can be used for analysing integral features of protein structure, protein–protein interactions as well as protein–ligand interactions. Finally, algorithms for classifying protein structures, in relation to the other protein structures known in the literature, are also included in this module through the fold detection routines. Here too, the unique integration of building,

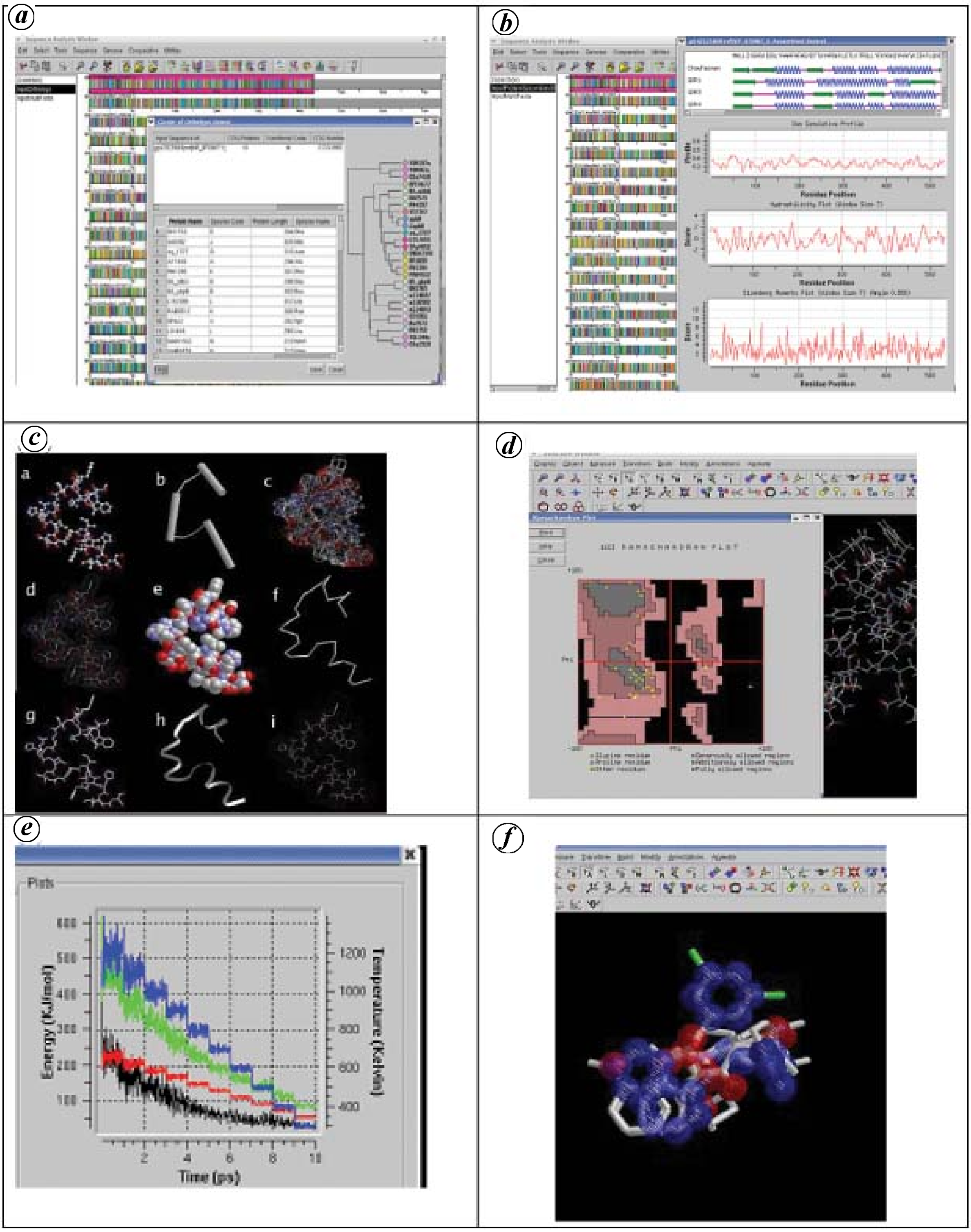


Figure 2. (*Contd...*)

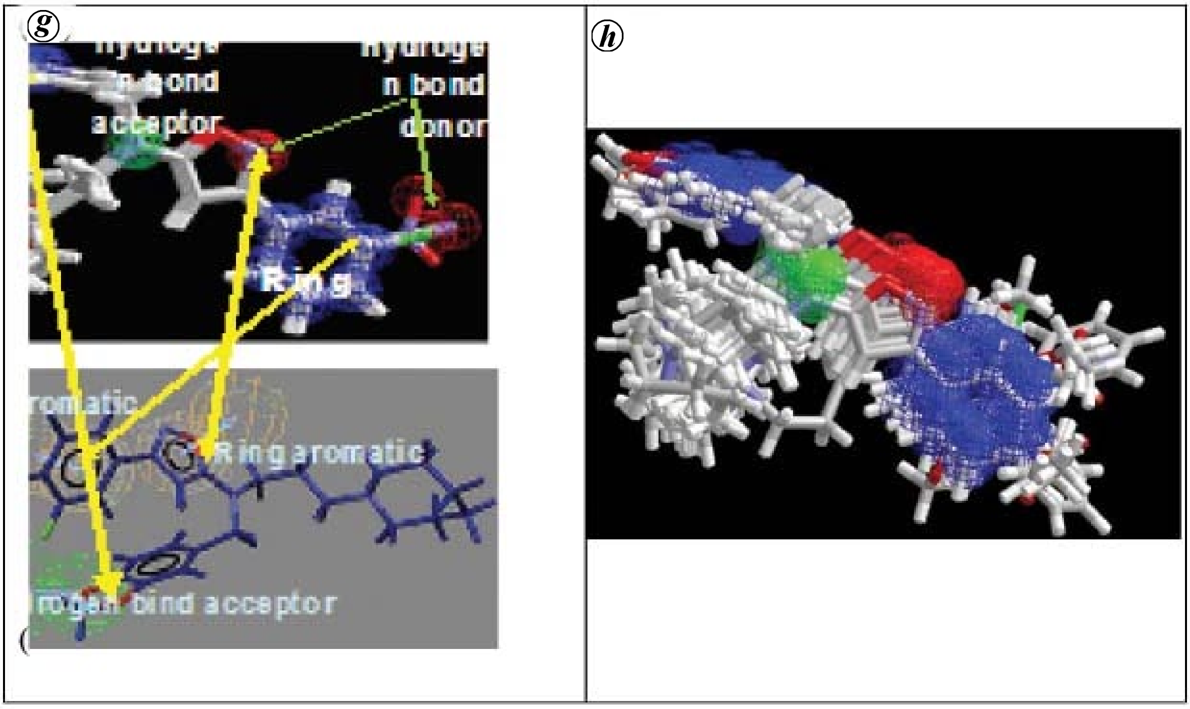


Figure 2 ***a–h.*** Example snapshots from various modules of BioSuite: ***a***, Genome comparison: Mapping Protein gi|42525869, from *Bacillus halorudians* to Clusters of Orthologous Groups (COG no. 1893), by using orthologues. A homologue of a lipase from Treponema denticola gi|42525869 was identified from *Bacillus halorudians*; ***b***, Protein secondary structure prediction using different methods and property profiles derived for the lipase protein sequence; ***c***, Different molecular representations in BioSuite – (a) ball-and-stick, (b) cartoons, (c) molecular surface, (d) van der Waals surface, (e) space fill, (f) C-alpha trace, (g) sticks, (h) ribbons, (i) solvent accessible surface; ***d***, Protein structure quality check using a Ramachandran plot; ***e***, An example of MD-analysis, variation in kinetic energy, potential energy, total energy, temperature during simulation; ***f***, An example of pharmacophore fitting, ***g***, Alignments produced by BioSuite derived pharmacophore model, and ***h***, An example of a field fit alignment: Molecular similarity between a pair of molecules is calculated by using the Gaussian function in BioSuite.

analysis and structural bioinformatics tools such as structure classification, all within one framework, significantly enhances the technical value of BioSuite.

Simulations

The ‘simulations’ module essentially simulates the behaviour of a molecule, in terms of its three-dimensional structure. The different submodules covered are, Forcefield, Energy minimization, Molecular dynamics, Monte Carlo simulations and Electrostatics. The molecular simulation of a system can conceptually be broken into three components: (a) generating a computational description of a biological/chemical system typically in terms of atoms, molecules and associated force field parameters, (b) the numerical solution of the equations which govern their evolution, and (c) the application of statistical mechanics to relate the behaviour of a few individual atoms/molecules to the collective behaviour of the very many. BioSuite is compatible both with the AMBER and the CHARMM force fields for macromolecules (proteins, nucleic acids and carbohydrates) and uses GAFF for small molecules (for e.g. natural substrates, drugs and drug-like substances). For each of the force fields, both treatments of the type of dielectric: either constant or distant dependent, are provided.

Several algorithms for first-order unconstrained energy minimization are contained in this module, providing a wide range of line search options. Thus, the coordinates of the molecular system can be adjusted so as to lower its energy, relative to the starting conformation, by using one of the following minimizers: Steepest descent algorithm, Conjugate gradient methods, Fletcher–Reeves algorithm, Polak–Ribiere algorithm, Polak–Ribiere plus algorithm and Shanno’s algorithm.

Further, to carry out molecular dynamics (MD) simulations, BioSuite provides NVE (micro-canonical), NVT (canonical), and NPT (isobaric–isothermal) ensembles for MD simulations with the choice of using velocity–verlet or leapfrog integrator. BioSuite also provides options for using SHAKE and RATTLE constraints.

MD being a deterministic approach, where the state of the system at any future time can be predicted from its current state, the tools provided in the suite can be used for solving Newton’s equations of motion for a given initial conformation, to study how the system evolves over time. Several intuitive and user-friendly tools are provided to analyse the resulting trajectories or time series of conformations. For example (Figure 2 *e*), plots at various energy levels along with the temperature, can be obtained. Plots generated with defined parameters show the structure and position at various energy levels, both of them present in two adjacent panels that can help to view the position of the molecule at a given temperature. The Monte Carlo method that generates configurations randomly and uses a special set of criteria to decide whether or not to accept each new configuration, is also part of this module.

In the electrostatics sub-module, BioSuite provides a solution for the linear Poisson–Boltzmann equation, to enable modelling of contributions of solvent, counterions and protein charges to electrostatic fields in molecules. Four choices for boundary conditions namely, zero, partial coulombic, full coulombic and focusing, are provided. For charge distribution, there are two options: trilinear and uniform. BioSuite has a very fast SOR solver, which utilizes spectral radius calculations to speed up convergence.

Drug design

This module provides the following functionalities: (a) Prediction of biological activities of unknown chemical entities using QSAR, (b) Identification of pharmacophores in biologically active molecules, (c) Superimposition of a set of molecules in 3D space by alignment, (d) Identification of the ligand poses in 3D space when it binds to a target using docking. Using the functionalities provided in the drug design module, one can identify lead-like molecules from a set of molecules, redesign them and predict their activities. Thus, lead optimization can be achieved iteratively. If the target structure is known, then the lead optimization can be done using the structure-based method, such as by docking.

The process of aligning a set of molecules in three-dimensional space, to find the superimposable regions of a group of molecules or to estimate molecular similarity can be performed by using either the ‘Field Fitting’ or the ‘RMS Fitting’ approach. The field fitting is done by aligning molecules using their electrostatic potentials and steric shapes, starting from their atomic coordinates and charges computed from Gaussian functions, while the ‘RMS fitting’ is done by minimizing the distances between specified atoms in the molecules. Flexible superposition can also be achieved by allowing rotations about single bonds.

For deriving and matching ‘3D-pharmacophores’, the following features are extracted/used: (a) Hydrogen bond donor, (b) Hydrogen bond acceptor, (c) Aliphatic hydrophobic group, (d) Aromatic ring, (e) Negatively charged group, and (f) Positively charged group. Pharmacophores are identified by using configurations of features common to a set of molecules. The pharmacophoric configurations are identified by a pruned exhaustive search, starting with small sets of features and extending them until no larger common configuration exists.

To carry out QSAR, where consistent relationships between the variations in the values of molecular properties and the biological activity for a series of compounds are sought, so that these ‘rules’ can be used to evaluate new chemical entities, a series of widely accepted feature extraction and statistical tools are provided within BioSuite. For example, a 2D-QSAR calculation uses either one or combinations of (a) Electronic, (b) Spatial, (c) Structural, (d) Thermodynamic and (e) Topological descriptors. BioSuite has the ability to compute 89 different descriptors. a few representative descriptors from different classes, e.g. Polarizability, HOMO and LUMO (electronic), Hf and Log P from (thermodynamic), log P, MR (thermodynamic), etc. and were compared with those computed from standard softwaers, using a dataset of 33 isooxazoles as potential thrombin receptor antagonists and in general, a high correlation (>0.9) was observed for the descriptor values.

Creating and refining a training set required for QSAR predictions are aided by (a) K-means, (b) K-nearest neighbours or (c) UPGMA hierarchical clustering algorithms. Tools are also provided for building user-defined data sets/training sets as well as for searching chemical databases. The QSAR model can be generated using regression techniques such as Multiple Linear Regression or Partial Least Squares. If the linearly independent descriptors for the molecules have to be eliminated while generating the model, then a dimensionality reduction can be performed by using either (a) Principal component analysis or (b) Discriminant analysis. Validation of the model to check the accuracy of the generated model can be performed by the K-fold cross validation technique

The structure-based drug design sub-module contains algorithms and utilities required for carrying out molecular docking. Using either simulated annealing or genetic algorithms (GA) based technique, the ligand conformations are searched and docked into the binding site of the macromolecule. In a simulated annealing-based method, the ligand’s current position, orientation and conformation are changed during each cycle, to reach the most energetically favourable conformation of the ligand bound to the target macromolecule. Thus these algorithms predict both the lowest energy conformation of the bound ligand as well as the best position and orientation for its binding to the target molecule, within the realm of the scientific capabilities of the approach.

A second popular algorithm is provided for this, the one based on genetic algorithms. The conformations of the ligand are encoded as a chromosome. The crossover and mutation operators are used to bring about random changes in the conformations of the ligand. A fitness function is defined for calculating the energy of the conformations generated. Through a number of runs of the GA cycle, a conformation having minimum energy is obtained.

Conformation search functionality generates the conformations for an input molecule, clusters the conformations and displays energy and torsion angle values of low energy conformations. This application generates conformations using two different methods, namely random conformation search and systematic conformation search. The random conformation search uses the simulated annealing algorithm. An option is provided to the user to select the rotatable bonds in the molecule. A few sample results from the drug-design modules are presented in Figure 2 *f–h*.

Performance evaluation

Evaluation has been an integral part of the entire development process. To start with, the choice of modules and the choice of algorithms themselves were evaluated, both at TCS and by the academic partners. The pseudo-codes and the SRS documents were then verified, followed by verification of the software codes by the TCS team. The scientific performance of the algorithms at various stages (versions 0.3, 0.7, 1.0a and 1.0) was evaluated independently by the academic partners at their institutions and any bugs reported or improvements suggested were subsequently considered and implemented into the suite, where appropriate. The outputs of each program were compared with those of other established academic codes/ commercial packages, to verify the scientific performance. They were also compared with the latest implementations of the chosen algorithms in the public domain, where available. The performance has been found to be comparable in all cases. While the utilities of many of the individual programs have been enhanced while implementing in BioSuite, the scientific capabilities and limitations of each of the programs are bounded by those of the corresponding original algorithms cited in Table 2.

An example of the manner in which the scientific performance was evaluated, is cited below. For testing the drug design module, 42 thymidine monophosphate kinase inhibitors were taken and minimization performed using both AMBER and CHARMM force fields with the conjugate gradient algorithm method. Conformational searches were tested with both systematic and randomized search methods. Alignments were satisfactory and we obtained low RMSD values for similar molecules, comparable to those obtained in Cerius. The time for computation was found to be good and comparable to other competitor software. The docking procedure is simple and user-friendly.

Prominent features of the package

For the most part, the existing software packages evolved out of academia, and were implementations of algorithms developed at different places and different times by different persons. As such, often there is no single 'superstructure' into which the algorithms fit seamlessly. To overcome these issues, BioSuite has been written in a modular fashion, which would permit the easy implementation of new algorithms as and when they are discovered. The unique partnership of the industry with academia harnesses the strengths of both communities, thus leading to a superior product both scientifically as well as according to software engineering standards. Some of the unique features of BioSuite are: (a) It is comprehensive, contains programs for carrying out sequence, whole genome and structure analysis, drug design, all under a common framework. (b) The software runs on simple personal computers on a Linux platform. (c) Domain identification and domain searching tools also available. (d) Transmembrane beta strand prediction, enhanced capability in building molecules in terms of the number of secondary structure templates available. (e) Enhanced capability in building larger carbohydrate structures, and (f) Code written fresh with CMMi-5 standards and consistency in coding methods to incorporate versatility in each program making up the entire suite, keeping in view of the genome-scale operations in bioinformatics.

Roadmap for the future

Going forward, several features are planned to be added to BioSuite to make it an even more useful platform for scientific research. Some developments in the pipeline are described below:

ADME

The Absorption, Distribution, Metabolism and Excretion profile (ADME) of a drug is an important determinant of its therapeutic efficacy. Accurately modelling the ADME properties of a candidate drug molecule is a necessary step to increase the chances that it will eventually become a successful drug. In the recent past, models have been developed for estimating various ADME-related properties such as blood-brain barrier penetration, human intestinal absorption, binding affinity to human serum albumin and $CaCO_2$ cell permeability. These will be integrated into the existing QSAR module of BioSuite.

Flexible docking

Docking, in BioSuite 1.0, explores the energetically optimal fit of a flexible small molecule with a rigid protein molecule. In subsequent releases, an improved version of the docking algorithm will be implemented that allows restricted flexibility in the protein molecule as well. This has been shown to be useful in improving the accuracy in prediction of the optimal binding conformation.

De novo drug design

An important requirement for drug design is the ability to generate novel molecules that bind to a known active site. Implementation of an algorithm is underway for the generation of novel binding candidates using a strategy of fragment docking followed by elaboration of selected fragments.

tRNA identification

A procedure for identifying tRNA genes in a genome will be included in the next version of BioSuite. The program identifies tRNAs based on the recognition of two intragenic control regions known as A and B boxes, a highly conserved part of B box, a transcription termination signal, and the evaluation of the spacing between these elements.

Improved whole genome comparison

MUMmer is an open source software package for the rapid alignment of very large DNA and amino acid sequences. A newer version of the MUMmer package has been integrated in BioSuite to find maximal unique matches between two genomes. The MUMmer output can also be viewed in the dot-plot format.

Improved graphics

Several techniques are being implemented to enhance the quality of the 3D graphics display in BioSuite while speeding up the display.

Scripting interface

While BioSuite provides a number of features and a vast array of functionality, users might want to implement their own procedures and programs. For this purpose, a scripting interface that exposes the functionality in BioSuite will be provided so that users can create their own workflows, develop and test new ideas and automate several tasks.

Sketcher

The next version of Bio-Suite will include a 2D sketcher for drawing molecules in a manner that chemists are familiar with and to automatically generate 3D structures for the molecules.

A high-performance version called Bio-Cluster for some of the memory intensive applications is also planned.

Hardware requirements and documentation

The minimum hardware requirements for BioSuite are as follows: Intel compatible ×86 Processor, 1.5 GHz, 256 MB RAM, 3 GB Free Hard Disk Space, Display capable of 1280 × 1024 pixel resolution, High end graphics card with 3D support for better viewing, Red-Hat Linux 8.0 or 9.0 or Fedora-Core 1/2 operating systems. BioSuite comes with its own set of documentation. The entire package is well documented and comes with easy to use tutorials, which reduce the learning curve and increase efficiency. Detailed documentation is available at the BioSuite website: http://www.atc.tcs.co.in/BioSuite/.

1. Huang, X., A contig assembly program based on sensitive detection of fragment overlaps. *Genomics*, 1992, **14**, 18–25.
2. Schuler, G. D., Sequence mapping by electronic PCR. *Genome Res*., 1997, **7**, 541–550.
3. Delcher, A. L., Harmon, D., Kasif, S., White, O. and Salzberg, S. L., Improved microbial gene identification with GLIMMER. *Nucleic Acids Res*., 1999, **27**, 4636–4641.
4. Kleffe, J., Hermann, K., Vahrson, W., Wittig, B. and Brendel, V., Logitlinear models for the prediction of splice sites in plant pre-mRNA sequences. *Nucleic Acids Res*., 1996, **24**, 4709–4717.
5. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J., Basic local alignment search tool. *J. Mol. Biol*., 1990, **215**, 403–410.
6. Needleman, S. B. and Wunsch, C. D., A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol*., 1970, **48**, 443–453.
7. Smith, T. F. and Waterman, M. S., Identification of common molecular subsequences. *J. Mol. Biol*., 1981, **147**, 195–197.
8. Thompson, J. D., Higgins, D. G. and Gibson, T. J., CLUSTALW improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res*., 1994, **22**, 4673–4680.
9. Arthur, L. D., Kasif, S., Fleschmann, R. D., Peterson, J., White, O. and Salzberg, S. L., Alignment of whole genomes. *Nucleic Acids Res*., 1999, **27**, 2369–2376.
10. Knuth, D. E., Morris, J. H. and Pratt, V. R., Fast pattern matching in strings. *SIAM J. Computing* 1977, **6**, 323–350.
11. Bailey, T. L. and Elkan, C., Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning J*., 1995, **21**, 51–83.
12. SantaLucia, J. Jr., Allawi, H. T. and Seneviratne, P. A., Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry*, 1996, **35**, 3555–3562.
13. Zuker, M., On finding all suboptimal foldings of an RNA molecule. *Science*, 1989, **244**, 48–52.
14. Jones, D. T., Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol*., 1999, **292**, 195–202.
15. Gromiha, M. M., Majumdar, R. and Ponnuswamy, P. K., Identification of membrane spanning beta strands in bacterial porins. *Protein Eng*., 1997, **10**, 497–500.
16. Durbin, R., Eddy, S., Krogh, A. and Mitchison, G., *Biological Sequence Analysis*: *Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, UK, 1998.
17. Mazumdar, A., Kolaskar, A. and Donald, S., GeneOrder: Comparing the order of genes in small genomes. *Bioinformatics*, 2001, **17**, 162–166.
18. Enright, A. J. and Ouzounis, C. A., GeneRAGE: a robust algorithm for sequence clustering and domain detection. *Bioinformatics*, 2000, **16**, 451–457.
19. Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. and Eisenberg, D., A combined algorithm for genome-wide prediction of protein function. *Nature*, 1999, **402**, 83–86.
20. Tamura, K. and Nei, M., Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evolut*., 1993, **10**, 512–526.

21. Fitch, W. M. and Margoliash, E., Construction of phylogenetic trees. *Science*, 1967, **155**, 279–284.
22. Bansal, M., Bhattacharyya, D. and Ravi, B., NUPARM and NUCGEN: software for analysis and generation of sequence dependent nucleic acid structures. *Comput. Appl. Biosci.*, 1995, **11**, 281–287.
23. Laskowski, R. A., MacArthur, M. W., Moss, D. S. and Thornton, J. M., PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.*, 1993, **26**, 283–291.
24. Sutcliffe, M. J., Haneef, I., Carney, D. and Blundell, T. L., Knowledge based modeling of homologous proteins, Part I: Three dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Eng.*, 1987, **1**, 377–384.
25. Baker, E. N. and Hubbard, R. E., Hydrogen bonding in globular proteins. *Progr. Biophys. Mol. Biol.*, 1984, **44**, 97–179.
26. Zhang, C. and Kim, S., Environment-dependent residue contact energies for proteins. *Proc. Nat. Acad. Sci.*, 2000, **97**, 2550–2555.
27. Orengo, C. A. and Taylor, W. R., SSAP: Sequential structure alignment program for protein structure comparison. *Methods Enzymol.*, 1996, **266**, 617–635.
28. Connolly, M. L., Computation of molecular volume. *J. Am. Chem. Soc.*, 1985, **107**, 1118–1124.
29. Brady, G. P. and Stouten, F. W. P., Fast prediction and visualization of protein binding pockets with PASS. *J. Computer-Aided Mol. Design*, 2000, **14**, 383–401.
30. Lichtarge, O., Bourne, H. R. and Cohen, F. E., An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, 1996, **257**, 342–358.
31. Gilbert, J. C. and Nocedal, J., Global convergence properties of conjugate gradient methods for optimization. *SIAM J. Optimization*, 1992, **2**, 21–42.
32. Watowich, S. J., Meyer, E. S., Hagstrom, R. and Josephs, R., A stable, rapidly converging conjugate gradient method for energy minimization. *J. Computat. Chem.*, 1988, **9**, 650–661.
33. Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S. Jr. and Weiner, P. K., A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.*, 1984, **106**, 765–784.
34. Jayaram, B., Sharp, K. A. and Honig, B., The electrostatic potential of B-DNA. *Biopolymers*, 1989, **28**, 975–993.
35. Nicholls, A. and Honig, B., A rapid finite difference algorithm, utilizing successive over-relaxation to solve the Poisson–Boltzmann equation. *J. Computat. Chem.*, 1991, **12**, 435–445.
36. Andersen, H. C., Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.*, 1980, **72**, 2384–2393.
37. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. and Haak, J. R., Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 1984, **81**, 3684–3690.
38. Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K. and Olson, A. J., Automated docking using a lamarckian genetic algorithm and empirical binding free energy function. *J. Computat. Chem.*, 1998, **19**, 1639–1662.
39. Goodman, J. M., *Chemical Applications of Molecular Modelling*, The Royal Society of Chemistry, London, 1998, pp. 61–69.
40. Good, A. C., Hodgkin, E. E. and Richards, W. G., Utilization of Gaussian functions for the rapid evaluation of molecular similarity. *J. Chem. Inf. Comput. Sci.*, 1992, **32**, 188.
41. Jones, G., Willet, P. and Glen, R. C., A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. *J. Comput.-Aided Mol. Des.*, 1995, **9**, 532.
42. Kurogi, Y. and Guner, O. F., Pharmacophore modeling and three-dimensional database searching for drug design using catalyst. *Curr. Med. Chem.*, 2001, **8**, 1035–1055.

Received 17 September 2005; revised accepted 26 October 2006

W584–W587 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl202

iCR: a web tool to identify conserved targets of a regulatory protein across the multiple related prokaryotic species

Sarita Ranjan, Jayshree Seshadri, Vaibhav Vindal, Sailu Yellaboina and Akash Ranjan*

Computational and Functional Genomics Group, Sun Centre of Excellence in Medical Bioinformatics,
Centre for DNA Fingerprinting and Diagnostics, EMBnet India Node, Hyderabad 500076, India

Received February 14, 2006; Revised and Accepted March 22, 2006

ABSTRACT

Gene regulatory circuits are often commonly shared between two closely related organisms. Our web tool iCR (identify Conserved target of a Regulon) makes use of this fact and identify conserved targets of a regulatory protein. iCR is a special refined extension of our previous tool PredictRegulon- that predicts genome wide, the potential binding sites and target operons of a regulatory protein in a single user selected genome. Like PredictRegulon, the iCR accepts known binding sites of a regulatory protein as ungapped multiple sequence alignment and provides the potential binding sites. However important differences are that the user can select more than one genome at a time and the output reports the genes that are common in two or more species. In order to achieve this, iCR makes use of Cluster of Orthologous Group (COG) indices for the genes. This tool analyses the upstream region of all user-selected prokaryote genome and gives the output based on conservation target orthologs. iCR also reports the Functional class codes based on COG classification for the encoded proteins of downstream genes which helps user understand the nature of the co-regulated genes at the result page itself. iCR is freely accessible at http://www.cdfd.org.in/icr/.

INTRODUCTION

Over last one and half decades, genomes of microorganisms have been sequenced at a highly accelerated pace. However, extracting useful information from such a large pool of genome data has become a major challenge of post genomics era. One approach to address this issue is to organize the large and complex genome into an ordered and manageable subsystem that can be tackled systematically. An important example of this approach is to study cellular processes and associated gene expression in terms of gene regulatory circuits. Each of these circuits contains a regulator and a list of its target sites (motifs) located upstream to a subset of genes that are being regulated (1–3). Such an approach will enable us to understand how the constituent genes of a genome come together to execute metabolic and physiological processes of a cell in response to a given regulator.

A large number of experimental and computational approaches are being attempted to understand how these genes come together to perform physiological function. The experimental approaches typically include microarray analysis of transcriptome (4,5). Subsequent to gathering the experimental data computational approaches are applied to search for common regulatory motifs and promoters present upstream to the up and down regulated genes and protein (6). Some of the computational tools like PHYLONET (7), BioProspector (8,9), Compare Prospector (9,10), MDscan (9,11), Motif Regressor (12), Bio Optimizer (13), PhyME (14) and so on are available for this purpose, but, most of these are either designed for eukaryotes or written to analyze the experimental data, such as micro array data, in terms of gene regulation.

An alternate approach could be to first select the regulator associated with a cellular process and then use computational approach to identify the potential target of regulatory protein which could then subsequently be followed up by experiments to validate the computationally identified targets. As a first step in this direction, we had previously proposed a tool called PredictRegulon, which finds targets of a regulatory protein in a genome based on limited set of known binding motif data (15). We have successfully used this tool to identify and validate the DtxR and IdeR targets in corynebacteria and mycobacteria, respectively (16,17). However an important limitation of Predictregulon was that it searches one genome at a time.

*To whom correspondence should be addressed. Tel: +91 40 27171442; Fax: +91 40 27171442; Email: akash@cdfd.org.in

Carrying out simultaneous search in multiple genomes offers many advantages, most important among these are ability of such approach to reveal the conserved regulatory targets across the multiple related genomes. This would increase the confidence of experimental biologist in taking up experimental validation. Further it was also felt that if we could group the targets based on class of genes that is being regulated then we could provide the overall impact of the regulator on the physiology of the organism.

We describe here iCR (identify Conserved target of a Regulon), a web server tool, for identification of conserved high priority targets of a regulatory protein from heterologous sequence data of prokaryotes (which includes regulatory sequences of genes and their orthologs in other species) where the user can easily distinguish biologically important motifs from background noise based on their cross species conservation.

PROGRAM DESCRIPTION

iCR is a CGI based web application written in Perl and C language. It uses a Shannon relative entropy based profile search method, similar to what was used in PredictRegulon tool. This application can utilize the available experimental data on binding sites of a transcription regulatory protein (18–20) to identify the regulons of a given regulator in genomes of various phylogenetically related bacterial species.

iCR is composed of three parts (Figure 1): (i) a front-end web interface for submitting the block aligned known binding motifs and for selection of species of choice; (ii) a search engine for scanning the upstream sequences; and (iii) a classification and reporting system for rendering the textual output produced by iCR into a meaningful grouping. Each of these components is discussed in detail in the help pages linked to the iCR home page. A brief description is being given here.

Input submission

iCR provides a web-based form for the input submission. The input form consists of two HTML pages. The first one accepts the sample motifs and the parameters defining the upstream region. On this page the known motifs can be copied either from sample input form or any authentic source and then be pasted in the web form in a block aligned fashion. The second page has a list of genomes organized in a taxonomically meaningful order for convenience in selection of multiple related species at a time and finally, the users need to specify the basis on which they want the predicted motifs to be grouped or classified on. The default or preferred option is Cluster of Orthologous Group (COG).

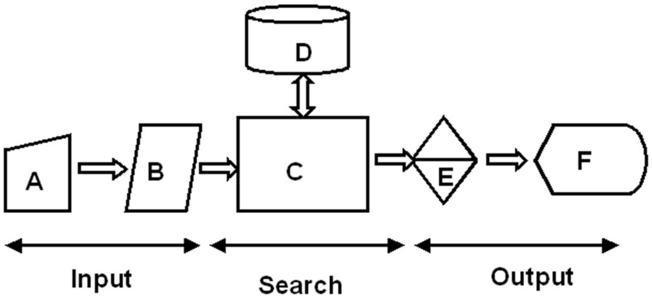


Figure 1. Architecture of iCR. iCR is a CGI application which collects input from user using html forms (A). B represents a Perl script that gathers the input from A launches the Search Engine (C) which looks up genome sequences and their annotations (D), and returns the potential targets as an output which is further classified based on COG/Class or Genome. The classified output is returned as HTML output (F).

Search engine

Parameters accepted from the input forms are passed to a search engine which uses the Shannon relative entropy based profile scan method to scan the upstream sequences for regulatory motifs. This method is described in our previous paper PredictRegulon (15). However this analysis is carried out on multiple user selected genome and the results are compiled together. Since the complete COG data were not available for many of the genes of various genomes, we updated these data by running COGNITOR (21,22). Each COG selected represents the best hits to proteins from at least three lineages.

The output of the search result is classified and grouped based on one of the three options—orthology, function class code or genome. Classification based on orthology (default option) lists all the orthologous targets of a regulator together emphasizing the fact that these are conserved targets of a given regulon.

Output

All the predicted and classified target motifs are presented as HTML table. This table has following columns: COG name, Functional class code, Genome, motif score, motif, Gene id mentioned in NCBI's ptt table, ORF number and gene product. The program predicts a number of motifs, the blue background color shows the high scoring motifs above the cut-off value. The motifs with yellow background color depicts exact match to the known binding sites.

Example usage

To demonstrate the typical application of iCR's regulon assignments, we chose to use known LexA-binding sites from *Bacillus subtilis* as a query set. These sites were collected from PRODORIC (19). We then selected different species belonging to Fermicutes (Bacillales, Lactobacillales, Clostridia and Mollicutes) simultaneously for search. We obtained the result classified on COG in which DNA motifs upstream to *lexA (COG1974), recA(COG0468), uvrB(COG0556), dinP(COG0389), rpsE(COG0098), rpsN(COG0098), rggD (COG0457)* and so on were picked up in many species together and therefore they qualify for conserved targets of LexA regulon (Table 1). Lex A is known to autoregulates itself (23). *recA* gene has been experimentally shown to be part of LexA regulon in *Escherichia coli* as well as *B.subtilis* (23,24). Homologs of *dinP* have also been shown to be regulated by LexA protein in *Bdellovibrio bacteriovorus* (25). LexA protein has been reported to interact with the regulatory region of *uvrB* in *B.subtilis* (19). All these observations confirm that the program is capable of identifying significant and high priority targets of a given regulator successfully. Additionally the result also highlights many motifs upstream to hypothetical genes/ ORFs. An experimental confirmation of interaction of these

Table 1. Output of iCR showing the conserved targets of LexA regulon in *Fermicutes*

COG	Class	Genome	Score	Position	Site	Gene	Synonym
COG1974	K	NC_004193	4.6875	−77	AGAACGAGTGTTTG	lexA	OB1669
COG1974	K	NC_003030	4.77125	−84	AGAACATAAGTTTG	lexA	CAC1832
COG1974	K	NC_002745	4.88271	−71	CGAACAAATGTTTG	lexA	SA1174
COG1974	K	NC_004557	4.82946	−80	AGAACATAAGTTTG	lexA	CTC01298
COG1974	K	NC_003366	4.83493	−70	AGAACATAAGTTTG	lexA	CPE1161
COG1974	K	NC_002570	4.72756	−77	AGAACTTATGTTTG	lexA	BH2356
COG1974	K	NC_000964	4.81601	−118	CGAACCTATGTTTG	lexA	BSU17850
COG1974	K	NC_003923	4.88303	−71	CGAACAAATGTTTG	lexA	MW1226
COG1974	K	NC_003212	4.82162	−79	CGAACCTTTGTTTG	—	LIN1340
COG1974	K	NC_002758	4.88182	−138	CGAACAAATGTTTG	lexA	SAV1339
COG1974	K	NC_003210	4.81541	−79	CGAACCTTTGTTTG	—	LMO1302
COG0468	L	NC_002570	4.64423	−121	CGAATAAATGTTCG	recA	BH2383
COG0468	L	NC_003212	4.67474	−138	CGAATAAATGTTCG	recA	LIN1435
COG0468	L	NC_003210	4.66915	−138	CGAATAAATGTTCG	recA	LMO1398
COG0468	L	NC_003923	4.40442	−143	AGCACGTTTGTTCG	recA	MW1168
COG0468	L	NC_002758	4.40302	−80	AGCACGTTTGTTCG	recA	SAV1285
COG0468	L	NC_003030	4.90549	−48	AGAACAAATGTTCG	recA	CAC1815
COG0468	L	NC_003366	5.01207	−34	AGAACTTATGTTCG	recA	CPE1673
COG0468	L	NC_004461	4.42484	−143	AGTACGTTTGTTCG	—	SE0963
COG0468	L	NC_000908	4.18494	−236	TGAACTGTTGTATG	recA	MG339
COG0468	L	NC_002745	4.40405	−143	AGCACGTTTGTTCG	recA	SA1128
COG0468	L	NC_004557	4.9426	−54	AGAACAGATGTTCG	recA	CTC01289
COG0556	L	NC_000964	4.7767	−122	CGAACTTTAGTTCG	uvrB	BSU35170
COG0556	L	NC_003923	4.8228	−105	CGAACAAACGTTTG	uvrB	MW0720
COG0556	L	NC_002745	4.82248	−105	CGAACAAACGTTTG	uvrB	SA0713
COG0556	L	NC_003030	4.93204	−29	CGAACAAATGTTTG	uvrB	CAC0502
COG0556	L	NC_002758	4.82157	−103	CGAACAAACGTTTG	uvrB	SAV0758
COG0556	L	NC_004193	4.65391	−69	CGAATACTTGTTCG	—	OB2488
COG0556	L	NC_003212	4.62091	−158	CGAAAATATGTTCG	uvrB	LIN2632
COG0556	L	NC_003210	4.61721	−160	CGAAAATATGTTCG	uvrB	LMO2489
COG0556	L	NC_004461	4.90087	−128	CGAACAAATGTTTG	—	SE0541
COG0389	L	NC_003366	4.82409	−26	TGAACATATGTTTG	dinP	CPE1566
COG0389	L	NC_003923	4.77999	−49	GGAACACGTGTTCG	—	MW1251
COG0389	L	NC_002758	4.33641	−6	AGAACATTTGTTCT	—	SAV1364
COG0389	L	NC_002745	4.81919	−49	AGAACACGTGTTCG	—	SA1196
COG0389	L	NC_003210	4.72978	−33	AGAACGCTTGTTCG	—	LMO1975
COG0389	L	NC_004461	4.32424	−75	AGAACAAATGTTCT	—	SE1046
COG0389	L	NC_003212	4.73647	−33	AGAACGCTTGTTCG	—	LIN2082
COG0389	L	NC_004557	4.82946	−40	AGAACATAAGTTTG	—	CTC00437
COG0389	L	NC_000964	4.37402	−68	CGAACATAAGTTCT	yqjW	BSU23710
COG0199	J	NC_004368	4.29015	−280	TGAACGTATGTACG	—	GBS0071
COG0199	J	NC_002662	4.9713	−280	CGAACGTATGTTCG	rpsN	L0391
COG0199	J	NC_003028	4.22998	−280	TGAACGTATGTACG	—	SP0222
COG0199	J	NC_003098	4.22977	−280	TGAACGTATGTACG	rpsN	SPR0202
COG0199	J	NC_002737	4.41534	−278	CGAACGTATGTACG	rpsN	SPY0064
COG0199	J	NC_003485	4.41477	−278	CGAACGTATGTACG	rpsN	SPYM18_0065
COG0199	J	NC_004432	4.29794	−140	CGAAATTGTGTATG	—	MYPE10040
COG0199	J	NC_004070	4.41397	−278	CGAACGTATGTACG	rpsN.1	SPYM3_0053
COG0199	J	NC_004116	4.2898	−280	TGAACGTATGTACG	—	SAG0071
COG1396	K	NC_003485	4.21446	−8	AGAAACCATGTTAG	—	SPYM18_0038
COG1396	K	NC_003923	4.32136	−263	GGAACAAGTGTACG	—	MW1228
COG1396	K	NC_004070	4.21434	−8	AGAAACCATGTTAG	—	SPYM3_0031
COG1396	K	NC_002570	4.4873	−118	GGAACGGGCGTTTG	—	BH0096
COG1396	K	NC_003028	4.47861	−127	TGAACAAATGTTGG	—	SP1115
COG1396	K	NC_002737	4.21453	−8	AGAAACCATGTTAG	—	SPY0037
COG1396	K	NC_004193	4.36271	−253	TGAACAGGAGTTAG	—	OB3501
COG1396	K	NC_003366	4.35319	−58	TGAACATTTGATTG	—	CPE2564
COG0098	J	NC_003028	4.38376	−109	AGAAGTGGTGTTCG	—	SP0227
COG0098	J	NC_004116	4.25066	−110	TGAAGTGGTGTTTG	rpsE	SAG0075
COG0098	J	NC_002737	4.23373	−110	TGAAGTGGTGTTTG	rpsE	SPY0069
COG0098	J	NC_004368	4.25082	−110	TGAAGTGGTGTTTG	rpsE	GBS0075
COG0098	J	NC_003098	4.38367	−109	AGAAGTGGTGTTCG	rpsE	SPR0206
COG0098	J	NC_004070	4.23345	−110	TGAAGTGGTGTTTG	rpsE	SPYM3_0057
COG0098	J	NC_004350	4.24208	−109	TGAAGTGGTGTTTG	rs5	SMU.2009
COG0098	J	NC_003485	4.23361	−110	TGAAGTGGTGTTTG	rpsE	SPYM18_0069
COG0457	R	NC_004557	4.34686	−240	GGAAGAAGAGTTTG	—	CTC02554
COG0457	R	NC_002570	4.38934	−268	CGAAGCAACGTTTG	—	BH3054
COG0457	R	NC_004557	4.39536	−233	AGAACAATTGTATG	—	CTC01089
COG0457	R	NC_002745	4.37946	−17	AGAAATGAGGTTCG	—	SA1448
COG0457	R	NC_003923	4.3797	−17	AGAAATGAGGTTCG	—	MW1570
COG0457	R	NC_003098	4.47855	−97	TGAACAAATGTTGG	rggD	SPR1022
COG0457	R	NC_002758	4.3788	−86	AGAAATGAGGTTCG	—	SAV1620

Note: Gene, Synonym column is as per NCBI ptt table. Class codes—K involved in transcription, L in DNA replication, recombination and repair, J represents orthologs involved in translation, ribosomal structure and biogenesis and so on.

motifs to LexA, followed by a functional assay based on known processes involved with a given regulator, could shed more lights on function of these hypothetical genes.

To test the sensitivity of the iCR predictions, we deleted two important and known binding motifs of LexA protein (present upstream to the *dinB* and *uvrB* in *B.subtilis*) from the input form and selected two species of Bacillales, *B.subtilis* and *Bacillus holodurans*. These two motifs were picked up on result page with blue background proving the reliability of predictions.

Certainly iCR results can serve as a useful starting point for molecular and cellular biologists for designing experiments to see the *in vitro* and *in vivo* effects of a regulatory protein in different systems.

CONCLUSION

To summarize, iCR is a web server that permits high throughput, detailed and fully automated prediction of potential binding targets of a regulatory protein in user selected prokaryotic species. iCR consists of 115 prokaryotic species arranged phylogenetically on the web interface. The first column on the result page, COG, is hyperlinked to NCBI and are fully navigable to allow users to have easy access to more related and descriptive information. The genome column shows the genome ID that is hyperlinked to a HTML page containing genome names corresponding to different IDs. For the user's convenience, functional class code column has also been linked to a page, which has a description of all the codes. iCR's strengths are in its free web accessibility, its comprehensiveness regarding choice of multiple species at a time, sorting of result based on COG and Class, and its interactive graphical interface.

ACKNOWLEDGEMENTS

Research in AR's laboratory is supported by grants from Council of Scientific and Industrial Research (CSIR) NMITLI, Department of Biotechnology, Department of Science and Technology, Govt. of India. S.R. and J.S. are supported by CSIR NMITLI Grant. V.V. is supported by UGC Research Fellowship and Y.S. is supported by CSIR Research Fellowship. Funding to pay the Open Access publication charges for this article was provided by Centre for DNA Fingerprinting and Diagnostics, Hyderabad.

Conflict of interest statement. None declared.

REFERENCES

1. Xing,B. and van der Laan,M.J. (2005) A causal inference approach for constructing transcriptional regulatory networks. *Bioinformatics*, **21**, 4007–4013.
2. Hershberg,R., Yeger-Lotem,E. and Margalit,H. (2005) Chromosomal organization is shaped by the transcription regulatory network. *Trends Genet*. **21**, 138–142.
3. Balazsi,G., Barabasi,A.L. and Oltvai,Z.N. (2005) Topological units of environmental signal processing in the transcriptional regulatory network of *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **102**, 7841–7846.
4. Ren,J. and Prescott,J.F. (2003) Analysis of virulence plasmid gene expression of intra-macrophage and *in vitro* grown Rhodococcus equi ATCC 33701. *Vet. Microbiol*., **94**, 167–182.
5. Rodriguez,G.M., Voskuil,M.I., Gold,B., Schoolnik,G.K. and Smith,I. (2002) ideR, an essential gene in Mycobacterium tuberculosis: role of IdeR in iron-dependent gene expression, iron metabolism, and oxidative stress response. *Infect. Immun*., **70**, 3371–3381.
6. Lin,L.H., Lee,H.C., Li,W.H. and Chen,B.S. (2005) Dynamic modeling of *cis*-regulatory circuits and gene expression prediction via cross-gene identification. *BMC Bioinformatics*, **6**, 258.
7. Wang,T. and Stormo,G.D. (2005) Identifying the conserved network of *cis*-regulatory sites of a eukaryotic genome. *Proc. Natl Acad. Sci. USA*, **102**, 17400–17405.
8. Liu,X., Brutlag,D.L. and Liu,J.S. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput*., 127–138.
9. Liu,Y., Wei,L., Batzoglou,S., Brutlag,D.L., Liu,J.S. and Liu,X.S. (2004) A suite of web-based programs to search for transcriptional regulatory motifs. *Nucleic Acids Res*., **32**, W204–W207.
10. Liu,Y., Liu,X.S., Wei,L., Altman,R.B. and Batzoglou,S. (2004) Eukaryotic regulatory element conservation analysis and identification using comparative genomics. *Genome Res*., **14**, 451–458.
11. Liu,X.S., Brutlag,D.L. and Liu,J.S. (2002) An algorithm for finding protein–DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat. Biotechnol*., **20**, 835–839.
12. Conlon,E.M., Liu,X.S., Lieb,J.D. and Liu,J.S. (2003) Integrating regulatory motif discovery and genome-wide expression analysis. *Proc. Natl Acad. Sci. USA*, **100**, 3339–3344.
13. Jensen,S.T. and Liu,J.S. (2004) BioOptimizer: a Bayesian scoring function approach to motif discovery. *Bioinformatics*, **20**, 1557–1564.
14. Sinha,S., Blanchette,M. and Tompa,M. (2004) PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics*, **5**, 170.
15. Yellaboina,S., Seshadri,J., Kumar,M.S. and Ranjan,A. (2004) PredictRegulon: a web server for the prediction of the regulatory protein binding sites and operons in prokaryote genomes. *Nucleic Acids Res*., **32**, W318–W320.
16. Yellaboina,S., Ranjan,S., Chakhaiyar,P., Hasnain,S.E. and Ranjan,A. (2004) Prediction of DtxR regulon: identification of binding sites and operons controlled by Diphtheria toxin repressor in Corynebacterium diphtheriae. *BMC Microbiol*., **4**, 38.
17. Prakash,P., Yellaboina,S., Ranjan,A. and Hasnain,S.E. (2005) Computational prediction and experimental verification of novel IdeR binding sites in the upstream sequences of Mycobacterium tuberculosis open reading frames. *Bioinformatics*, **21**, 2161–2166.
18. Salgado,H., Santos-Zavaleta,A., Gama-Castro,S., Millan-Zarate,D., Diaz-Peredo,E., Sanchez-Solano,F., Perez-Rueda,E., Bonavides-Martinez,C. and Collado-Vides,J. (2001) RegulonDB (Version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12. *Nucleic Acids Res*., **29**, 72–74.
19. Munch,R., Hiller,K., Barg,H., Heldt,D., Linz,S., Wingender,E. and Jahn,D. (2003) PRODORIC: prokaryotic database of gene regulation. *Nucleic Acids Res*., **31**, 266–269.
20. Ishii,T., Yoshida,K., Terai,G., Fujita,Y. and Nakai,K. (2001) DBTBS: a database of *Bacillus subtilis* promoters and transcription factors. *Nucleic Acids Res*., **29**, 278–280.
21. 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.
22. Tatusov,R.L., Natale,D.A., Garkavtsev, IV, Tatusova,T.A., Shankavaram,U.T., Rao,B.S., Kiryutin,B., Galperin,M.Y., Fedorova,N.D. and Koonin,E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res*., **29**, 22–28.
23. Little,J.W., Mount,D.W. and Yanisch-Perron,C.R. (1981) Purified lexA protein is a repressor of the recA and lexA genes. *Proc. Natl Acad. Sci. USA*, **78**, 4199–4203.
24. Groban,E.S., Johnson,M.B., Banky,P., Burnett,P.G., Calderon,G.L., Dwyer,E.C., Fuller,S.N., Gebre,B., King,L.M., Sheren,I.N. *et al*. (2005) Binding of the *Bacillus subtilis* LexA protein to the SOS operator. *Nucleic Acids Res*., **33**, 6287–6295.
25. Campoy,S., Salvador,N., Cortes,P., Erill,I. and Barbe,J. (2005) Expression of canonical SOS genes is not under LexA repression in *Bdellovibrio bacteriovorus*. *J. Bacteriol*., **187**, 5367–5375.

Comparative analysis of iron regulated genes in mycobacteria

Sailu Yellaboina, Sarita Ranjan, Vaibhav Vindal, Akash Ranjan*

Computational and Functional Genomics Group, Sun Centre of Excellence in Medical Bioinformatics, Centre for DNA Fingerprinting and Diagnostics, EMBnet India Node, Hyderabad 500076, India

Received 9 January 2006; revised 20 March 2006; accepted 28 March 2006

Available online 12 April 2006

Edited by Robert B. Russell

Abstract Iron dependent regulator, IdeR, regulates the expression of genes in response to intracellular iron levels in *M. tuberculosis*. Orthologs of IdeR are present in all the sequenced genomes of mycobacteria. We have used a computational approach to identify conserved IdeR regulated genes across the mycobacteria and the genes that are specific to each of the mycobacteria. Novel iron regulated genes that code for a predicted 4-hydroxy benzoyl coA hydrolase (Rv1847) and a protease dependent antibiotic regulatory system (Rv1846c, Rv0185c) are conserved across the mycobacteria. Although *Mycobacterium* natural-resistance-associated macrophage protein (Mramp) is present in all mycobacteria, it is, as predicted, an iron-regulated gene in only one species, *M. avium* subsp. *paratuberculosis*. We also observed an additional iron-regulated exochelin biosynthetic operon, which is present only in non-pathogenic *Mycobacterium*, *M. smegmatis*.

Keywords: Regulon; Modules; Prediction; Iron; Pathogen; Bacteria; *Mycobacterium*

1. Introduction

Iron is a cofactor for many enzymes and essential for growth of bacteria [1,2]. However, iron can also act as a potential catalyst of oxidative stress in bacteria. High amount of iron levels in a bacterium is countered by inducing synthesis of proteins involved in iron storage and oxidative stress defense to protect against iron-mediated oxidative damage [3,4].

Iron limitation leads to the growth restriction of many species of mycobacteria including *Mycobacterium tuberculosis*, which causes tuberculosis in humans [5]. Iron is an obligate cofactor for at least 40 different enzymes encoded in the *M. tuberculosis* genome.

In pathogenic bacteria, many virulence factors and iron acquisition systems are regulated by iron dependent transcription regulators [6]. There are two such regulators identified in *M. tuberculosis*, ferric uptake regulator (furA) and Iron dependent regulator (IdeR).

IdeR is a global regulator of iron response and belongs to the diphtheria toxin repressor (DtxR) family of transcription regulators [7]. Electrophoretic mobility shift assay (EMSA) and DNA footprinting analysis have lead to the identification of IdeR binding sites in upstream sequences of genes that code the proteins that are involved in biosynthesis of siderophores (MbtA, MbtB, MbtI), aromatic amino acids (PheA, HisE, HisG), lipopolysacaharide molecules (Rv3402c), lipids (AcpP), Peptidoglycan (MurB) and others annotated to be involved in iron storage (BfrA, BfrB) [8,9]. DNA microarray analysis of iron-dependent transcriptional profiles of wild-type and IdeR mutant of *M. tuberculosis* has lead to the identification of variety of other genes that code for the proteins like putative transporters (Rv0282, Rv0283, Rv0284), members of the glycine-rich PE/PPE family (Rv2123), membrane proteins involved in virulence (MmpL4, MmpS4), transcriptional regulators, enzymes involved in lipid metabolism (Rv1344, Rv1345, Rv1346, Rv1347) and amino acid metabolism (TrpE2, PheA) [10].

Orthologues of IdeR are present in all the sequenced genomes of mycobacteria. In this paper, we attempt to identify common and unique iron regulated genes in genomes of *M. leprae*, *M. avium* subsp. *paratuberculosis* and *M. smegamatis*. We applied a computational genomics tool – Predictregulon to identify the IdeR binding motifs and operon context of that motif [11]. Previously reported IdeR binding sites from *M. tuberculosis* were used to generate a recognition profile based on Shannon relative entropy, which was used to predict potential IdeR sites in sequenced genomes of mycobacteria. Further we have also predicted the other co-expressed genes that are potentially part of IdeR regulated operons. A sample of predicted motifs in *M. smegmatis* was experimentally verified by EMSA using recombinant *M. tuberculosis* IdeR.

2. Materials and methods

2.1. *In silico identification of IdeR binding sites*

Published and annotated genome sequences of *M. tuberculosis*, *M. leprae* and *M. avium* subsp. *paratuberculosis* were downloaded from NCBI ftp site (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). Unpublished genome sequence of *M. smegmatis* was downloaded from TIGR site (http://pathema.tigr.org/tigr-scripts/CMR/CmrHomePage.cgi). The known IdeR binding sites collected from the literature [8,9] were used to built IdeR binding site recognition profile and identify the IdeR binding sites as well as target genes in all the genomes of mycobacteria, using a method described previously [11,12].

2.2. *Cloning, expression and purification of* M. tuberculosis *IdeR*

pQE30 expression vector (Qiagen) with an N terminal 6× His tag was used to clone the ORF Rv2711 of *M. tuberculosis* that encodes IdeR. Briefly, Rv2711 was taken out from pRSET IdeR construct [13] with specific restriction enzyme sites (*Bam*H1 and *Hin*dIII) and

*Corresponding author. Fax: +91 40 27155610.
E-mail address: akash@cdfd.org.in (A. Ranjan).

Abbreviations: M. tuberculosis, *Mycobacterium tuberculosis*; *M. avium* subsp. *paratuberculosis*, *Mycobacterium* subsp. *paratuberculosis*; *M. smegmatis*, *Mycobacterium smegmatis*; DtxR, Diphtheria toxin repressor; IdeR, Iron-dependent regulator; RPS-BLAST, reversed position specific-basic local alignment search tool; EMSA, electrophoretic mobility shift assay

0014-5793/$32.00 © 2006 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.febslet.2006.03.090

the insert was cloned into the corresponding sites of pQE30 expression vector. *Escherichia coli* M15 cells transformed with the 6× His tagged chimeric construct were grown in 400 mL of LB medium supplemented with 100 μg/ml of ampicillin and 25 μg/ml of kanamycin. IPTG (0.2 mM) was added to a mid log phase culture. The cells were kept in an incubator shaker for another eight hours at 27 °C and 200 rpm to allow protein expression. Then, cells were harvested by centrifugation and resuspended in 10 ml of lysis buffer (50 mM NaH_2PO_4, 300 mM NaCl and 10 mM imidazole, pH 8) with 1 mM PMSF and disrupted using a sonicator. After a second round of centrifugation for 10 min at $10000 \times g$, the supernatant was applied to a Ni–NTA affinity column (Qiagen, USA). The supernatant was allowed to bind to Ni–NTA column. The recombinant protein was eluted with 200 mM imidazole and analyzed by SDS–PAGE after washing the column with 5 bed-volumes of wash buffer containing 20 mM imidazole.

2.3. Electrophoretic mobility shift assay

Double-stranded oligonucleotides containing the predicted binding motif (19 bp long) were end labeled with T4 polynucleotide kinase and [$^{32}P\gamma$]-ATP and were incubated with the purified recombinant IdeR protein in a binding reaction mixture. The binding reaction mixture (20-μl total volume) contains the DNA-binding buffer (20 mM Tris–HCl [pH 8.0], 2 mM DTT, 50 mM NaCl, 5 mM $MgCl_2$, 50% glycerol, and 5 μg of bovine serum albumin per ml), 10 μg of poly (dI–dC) per ml (for non-specific binding) and 200 μM $NiSO_4$. The reaction mixture was incubated at room temperature for 30 min and loaded onto 7% polyacrylamide gel containing 1× Tris–borate–EDTA buffer. No dye was added for loading. The gel was electrophoresed at 200 V for 2 h. Subsequently, the gel was dried and exposed to Storage Phosphor Image Plates for 4 h. The image plates were subsequently scanned in Storage Phosphor Imaging workstation.

3. Results and discussion

3.1. IdeR from various species of actinobacteria shows a similar DNA binding domain

In order to assess the rationale of using *M. tuberculosis* IdeR binding sites to identify the IdeR binding sites in other species, IdeR-orthologs from actinobacteria were aligned with each other using CLUSTALW. Alignment showed a very high sequence similarity at the N-terminal region which is involved in DNA binding (Fig. 1). This suggests that the target DNA motifs in various genomes can be recognized based on sequence recognition profile generated from experimentally defined IdeR target motifs from *M. tuberculosis*.

Table 1
Known IdeR binding sites from *M. tuberculosis*

Binding site	Gene
CAAGGTAAGGCTAGCCTTA	Rv1519
TTATGTTAGCCTTCCCTTA	Rv3403c
TTAACTTAGGCTTACCTAA	Rv3839
TTAGGCAAGGCTAGCCTTG	Rv1343c
CAAGGCTAGGCTTGCCTAA	Rv1344
TATGGCATGCCTAACCTAA	Rv1347c
TTCGGTAAGGCAACCCTTA	Rv1348
ATAGGTTAGGCTACCCTAG	Rv2122c
CTAGGGTACCCTAACCTAT	Rv2123
AGAGGTAAGGCTAACCTCA	Rv3402c
TTAGTGGAGTCTAACCTAA	Rv1876
GTAGGTTAGGCTACATTTA	Rv2386c
CTAGGAAAGCCTTTCCTGA	Rv3841
TTAGCTTATGCAATGCTAA	Rv0282
TTAGGCTAGGCTTAGTTGC	Rv0451c
TTAGCACAGGCTGCCCTAA	Rv2383c
TTAGGGCAGCCTGTGCTAA	Rv2384

3.2. In silico prediction of IdeR binding sites and target operons

A recognition profile of experimentally defined IdeR binding sites (Table 1) from *M. tuberculosis* was used to identify the potential IdeR binding sites and downstream operons/genes in genomes of *M. avium* subsp. *paratuberculosis* (Tables 2 and 3), *M. smegamatis* (Tables 4 and 5), *M. leprae* (Tables 6 and 7). Function for the proteins encoded by these genes were predicted by RPS-BLAST (reversed position specific-basic local alignment search tool) search against conserved domain database [14]. Table 8 lists the distribution of orthologous genes of IdeR regulated genes belonging to different functional category across the mycobacteria.

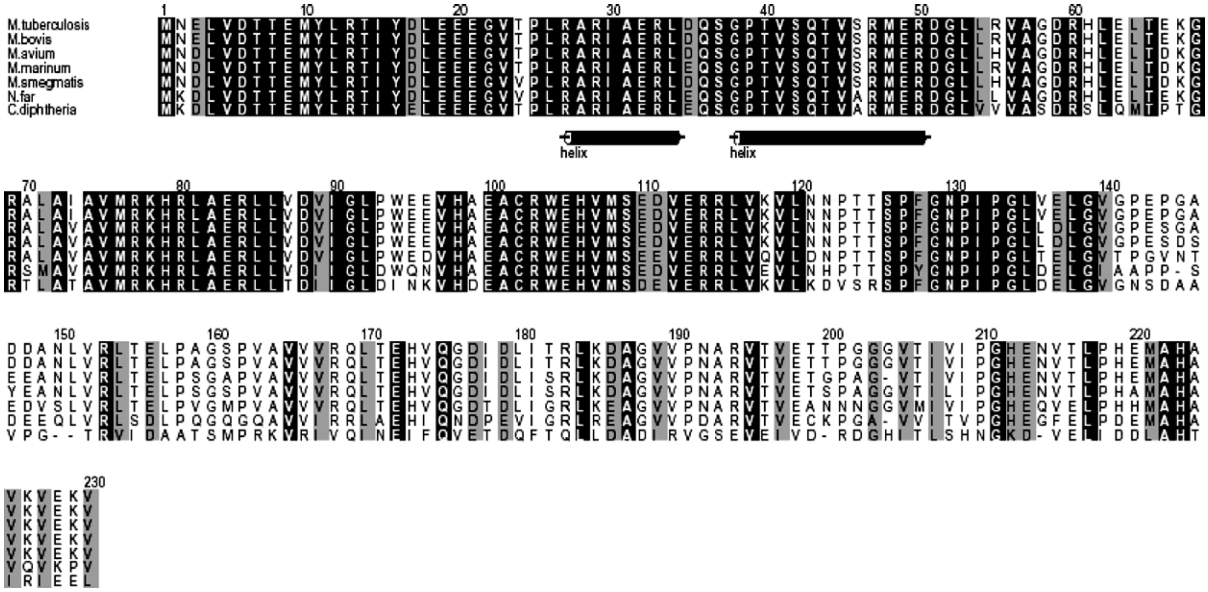


Fig. 1. Alignment of IdeR orthologues from different species of actinobacteria reveals a highly conserved DNA binding domain The arial black shadow show identity and the gray show similarity. Two helices (labeled as helix) are part of helix turn helix that assists in IdeR box recognition.

Table 2
Predicted IdeR binding sites in *M avium* subsp. *paratuberculosis*

Score	Position	Binding site	Gene	Synonym	Product
6.41034	−184	TTAGGTTAGACTCACCTAA	–	MAP1594c	Hypothetical protein
6.35589	−243	ATAGGCAAGGCTGCCCTAA	–	MAP1559c	Hypothetical protein
6.33364	−209	TTAGTGGAGTCTAACCTAA	*bfrA*	MAP1595	BfrA
6.22698	−78	TTAGGTAAGCCTAAGTTAA	*pheA*	MAP0193	PheA
6.20315	−32	TTAACTTAGGCTTACCTAA	–	MAP0192c	Hypothetical protein
6.18548	−94	TTAGCACAGGCTGCCCTTA	*mbtA*	MAP2178	MbtA
6.08146	−202	TTAGGGCAGCCTTGCCTAT	–	MAP1560	Hypothetical protein
6.07653	−25	ATAGGTTAGGCTACATTTA	*trpE2*	MAP2205c	TrpE2
5.89751	−46	ATAGTGCACACTATCCTAA	–	MAP2052c	Hypothetical protein
5.85458	−32	TAAGGGCAGCCTGTGCTAA	*mbtB*	MAP2177c	MbtB
5.81294	−55	TTAGGTAAGCCTAGCATCC	–	MAP0794	Hypothetical protein
5.80159	−27	TTAGGTACGGCTAGCCTCA	–	MAP0024c	Hypothetical protein
5.75148	−12	TTAGGTAAACCTTGGCTAT	–	MAP4065	Hypothetical protein
5.74424	−285	ATAGCCAAGGTTTACCTAA	–	MAP4064c	Hypothetical protein
5.7243	−38	GGATGCTAGGCTTACCTAA	–	MAP0793c	Hypothetical protein
5.71252	−56	TTTAGCTAGGCTACGCTAA	–	MAP1762c	Hypothetical protein
5.65231	−341	TAAGGCTAGCGTTGCCTAA	*fadD33_2*	MAP1554c	Fadd33_2
5.65231	−79	TAAGGCTAGCGTTGCCTAA	–	MAP1555c	Hypothetical protein
5.63035	−65	TTATGCAATGCTAACTTCA	–	MAP3778	Hypothetical protein
5.61853	−90	ATAGAGAATACTATTCTCA	–	MAP0680	Hypothetical protein
5.61329	−26	GCAGGTCAGGCTACCGTTA	*murB*	MAP3975	MurB
5.50085	−182	TTTGGTAAGGCAACCCTTA	–	MAP2414c	Hypothetical protein
5.47614	−189	CTACGCCAACCTCACCTTA	–	MAP2111c	Hypothetical protein
5.47185	−49	TTCGGTGACGCTAGACTGA	–	MAP2908c	Hypothetical protein
5.45568	−43	TGAGGCTAGCCGTACCTAA	–	MAP0025	Hypothetical protein
5.39833	−56	TTAGGGAAAGCTTAGGTAT	–	MAP2018c	Hypothetical protein
5.38891	−31	TTACGTCAAGCTGGCCTTC	*viuB*	MAP2960c	ViuB

Table 3
Predicted IdeR regulated operons in *M. avium* subsp. *paratuberculosis*

Synonym	Gene	COG no.	Product
MAP1594c	–	–	Bacterioferritin-associated ferredoxin
MAP1595	*bfrA*	COG2193	BfrA
MAP1558c	–	COG0501	Zn-dependent protease
MAP1559c	–	COG3682	Transcription regulator
MAP1560	–	COG2050	Possibly involved in aromatic compounds catabolism
MAP0191c	–	COG1316	Hypothetical protein
MAP0192c	–	COG4175	Hypothetical protein
MAP0193	*pheA*	COG0077	PheA
MAP0194	–	COG0406	Fructose-2,6-bisphosphatase
MAP2169c	*mbtH_3*	COG3251	MbtH_3
MAP2170c	*mbtG*	COG3486	MbtG
MAP2171c	*mbtF*	COG1020	MbtF
MAP2172c	–	COG1020	Putative non-ribosomal peptide synthetase
MAP2173c	*mbtE*	COG1020	MbtE
MAP2174c	*mbtD*	COG3321	MbtD
MAP2175c	*mbtC*	COG3321	MbtC
MAP2176c	–	COG3208	Thio esterase (similar to mbtB)
MAP2177c	*mbtB*	COG1020	MbtB
MAP2178	*mbtA*	COG1021	MbtA
MAP2179	–	–	Hypothetical protein
MAP2205c	*trpE2*	COG0147	TrpE2
MAP2206	–	COG3329	Predicted permease
MAP2051c	–	COG2124	Cytochrome P450 monooxygenase
MAP2052c	–	–	Bacterial regulatory proteins, tetR family
MAP2053	–	–	Hypothetical protein

(*continued on next page*)

Table 3 (*continued*)

Synonym	Gene	COG no.	Product
MAP0791c	–	COG2226	Hypothetical protein
MAP0792c	–	COG2141	F420-dependent N5,N10-methylene tetrahydromethanopterin reductase
MAP0793c	–	COG0654	Monooxygenase, FAD-binding
MAP0794	–	COG1309	Bacterial regulatory proteins, tetR family
MAP0795	–	COG2141	Luciferase-like monooxygenase
MAP0024c	–	COG5651	PPE-repeat proteins
MAP0025	–	COG0236	Acyl carrier protein
MAP0026	*fadD33_1*	COG0318	FadD33_1
MAP4064c	–	COG3315	*O*-Methyltransferase involved in polyketide biosynthesis
MAP4065	–	COG1914	Mramp
MAP1760c	–	COG2837	Predicted iron-dependent peroxidase
MAP1761c	–	COG2822	Predicted periplasmic lipoprotein involved in iron transport
MAP1762c	–	COG0672	FTR1, high-affinity Fe^{2+}/Pb^{2+} permease
MAP1553c	*fadE14*	COG1960	FadE14
MAP1554c	*fadD33_2*	COG0318	FadD33_2
MAP1555c	–	COG0236	Acyl carrier protein
MAP3777	–	COG3315	*O*-Methyltransferase involved in polyketide biosynthesis
MAP3778	–	COG0464	Hypothetical protein
MAP3779	–		
MAP3780	–		
MAP3781	–		
MAP0677c	–	COG2159	Hypothetical protein
MAP0678c	–	COG2329	Enzyme involved in biosynthesis of extracellular polysaccharides
MAP0679c	*fdxB*	COG0633	FdxB
MAP0680	–	COG0318	Acyl-CoA synthetases (AMP-forming)/AMP-acid ligases II
MAP0681	–	COG1960	Acyl-CoA dehydrogenase
MAP0682	–	COG1960	Putative acyl-CoA dehydrogenase
MAP0683	–	COG1024	Enoyl-CoA hydratase/isomerase family
MAP3973c	–	COG0388	Predicted amidohydrolase
MAP3974c	–	COG3832	Predicted lactoylglutathione lyase
MAP3975	*murB*	COG0812	MurB
MAP3976	–	COG1376	Putative lipoprotein
MAP2412c	–	COG3173	Predicted aminoglycoside phosphotransferase
MAP2413c	–	COG1132	ABC-type multidrug/protein/lipid transport system
MAP2414c	–	COG1132	ABC-type multidrug/protein/lipid transport system
MAP2109c	–	COG2516	Predicted Fe–S oxidoreductases
MAP2110c	–	COG1575	1,4-Dihydroxy-2-naphthoate octaprenyltransferase
MAP2111c	–	COG1463	ABC-type transport system, resistance to organic solvents, periplasmic
MAP2958c	*xerC*	COG4974	XerC
MAP2959c	–	COG1304	L-Lactate dehydrogenase
MAP2960c	*viuB*	COG2375	ViuB

Note: Genes that are part of an operon are together.

3.3. Experimental validation of predicted binding sites

A sample of predicted regulator binding motifs in (Table 6) upstream sequences of the *M. smegmatis* genes that code for a hypothetical protein (MSMEG6382), periplasmic component of ABC-type Fe^{3+}-hydroxamate transport system (MSMEG0020), iron utilization protein (MSMEG5028 and MSMEG0011) and a predicted 4-Hydroxy benzoyl co-A hydrolase (MSMEG3634) were experimentally verified by EMSA using recombinant IdeR from *M. tuberculosis*. Double stranded 19-mer synthetic oligonucleotides corresponding to the predicted DNA-binding sites were labeled with [$^{32}P\gamma$]-ATP and mixed with purified IdeR in presence of Nickel ions and was assayed for the formation of DNA–protein complex using EMSA. Nickel was used as the divalent metal in the binding reactions on account of its redox stability compared with ferrous ion. IdeR is able to retard the electrophoretic mobility of all the double stranded oligonucleotides though the level of affinity may vary (Fig. 2). A synthetic motif-ds (5′-TTTTCATGAAGTCTTCTAA-3′), which was used as a negative control, did not show any complex formation. These results indicate that the predicted IdeR-binding sites can indeed bind to IdeR.

Table 4
Predicted IdeR binding sites in *M. leprae*

Score	Position	Binding site	Gene	Synonym	Product
4.91319	−213	ATAGGCAAGGCTGCCCTAA	–	ML2063	Possible regulator
4.8888	−269	TTAGTGGAGTCTAACCTAA	*bfrA*	ML2038	Bacterioferritin
4.57039	−208	CGAGGTTAGACTAAGCTAA	*hisE*	ML130	Phosphoribosyl-ATP pyrophosphatase
4.49503	−6	GTAGGCCAGTCTATCGTTA	*murB*	ML2447	UDP-*N*-acetylenolpyruvoylglucosamine reductase
4.292	−243	GTATCCTAGGCTAGCCTGG	*fdxA*	ML1489	Ferredoxin (Fe–S co-factor)
4.25015	−69	CCAGACCAGGCTACCCTAG	–	ML0453	Conserved hypothetical protein
4.22436	−69	GGATGACAGGCTGACCTGA	*glpK*	ML2314	Glycerol kinase
4.19852	−78	TTACGCTAGTCTCAAGTAA	–	ML1689	Possible hydrolase
4.14559	−361	TTATACAAGTCTTTGCTTT	*ilvG*	ML2083	Acetolactate synthase II
4.13935	−130	CTAGGGAAGGGTACCCTCG	–	ML0591	Putative membrane protein
4.12623	−158	CTCGCGGAGCCTTCGCTGA	–	ML2158	Hypothetical protein
4.12616	7	TTAGCTTACGCAATGCTAA	–	ML2537	Conserved hypothetical protein

Table 5
Predicted IdeR regulated operons in *M. leprae*

Gene	Synonym	COG no.	Product
–	ML2063	COG3682	Possible regulator
–	ML2064	COG0501	Integral membrane protein
–	ML2035	–	Amycolatopsis Mediterranei U32 Nacd Nitrite Reductase
bfrA	ML2038	COG2193	Bacterioferritin
hisE	ML1309	COG0140	Phosphoribosyl-ATP pyrophosphatase
hisG	ML1310	COG0040	ATP phosphoribosyltransferase
–	ML2446	COG1376	Possible lipoprotein
murB	ML2447	COG0812	UDP-*N*-acetylenolpyruvoylglucosamine reductase
	ML1488	COG0436	Putative aspartate aminotransferase [EC:2.6.1.1]
fdxA	ML1489	COG1146	Ferredoxin
–	ML0450	COG0214	Putative pyridoxine biosynthesis protein
–	ML0451	COG0494	NTP pyrophosphohydrolases including
–	ML0452	COG0438	Putative glycosyltransferase
–	ML0453	COG1560	Phosphatidylinositol synthase Pgsa
glpK	ML2314	COG0554	Glycerol kinase
gltS	ML1688	COG0008	Glutamyl-Trna synthase
–	ML1689	COG0179	Possible hydrolase
ilvG	ML2083	COG0028	Acetolactate synthase II
–	ML0589	COG0842	ABC-type multidrug transport system
–	ML0590	COG1131	ABC-type multidrug transport system
–	ML0591	–	Putative membrane protein
–	ML2534	–	PE-family protein
–	ML2535	COG1674	DNA segregation Atpase Ftsk/Spoiiie
–	ML2536	–	Conserved membrane protein
–	ML2537	COG0464	Atpase, AAA family

Note: Genes that are part of an operon are together.

3.4. *Conserved iron dependent modules in Mycobacterium species*

A comparative analysis of IdeR target genes in various mycobacteria enabled us to identify the common iron dependent genes. Conservation of these genes in the predicted IdeR regulons suggests an important role of their cognate gene products in iron metabolism.

Orthlogues of the *pheA* (Rv3838c) and Rv3837c in *Mycobacterium* species are predicted to be regulated by IdeR (Table 8). The gene *pheA* codes for a predicted prephenate dehydratase and other gene Rv3837 codes for 2,3-PDG dependent phosphoglycerate mutase. They both belong to the same operon and are likely to be involved in similar function.

The gene *trpE2* (Rv2386c) is conserved across the predicted IdeR regulons (Table 8). The gene *trpE2* has been predicted to code for an isochorismate synthase that can catalyze the conversion of chorismate to isochorismate, the precursor for salicylate [15]. Later its orthologue *ybtS* in *Yersinia enterocolitica* has been suggested to catalyze formation of salicylate from chorismate [16].

The genes that code for an iron storage protein (BfrA), siderophore biosynthesis and siderophore transport system are also conserved across the IdeR regulon of mycobacteria (Table

Table 6
Predicted IdeR binding sites in *M. smegmatis*

Score	Position	Binding site	Synonym	Product
6.1538	−204	TTAGCGGAGTCTAACCTTA	MSMEG3564	Bacterioferritin
6.13839	−95	TTAGCACAGGCTGTCCTAA	MSMEG4510	MbtA
6.13198	−148	TTAGGCAACGCTAAGCTAA	MSMEG5992	Hypothetical protein
6.07216	−22	TTAGGACAGCCTGTGCTAA	MSMEG4509	Dihydroaeruginoic acid synthetase
6.06436	−62	**TTAAGTTAGGCTTACCTCA**	MSMEG6382	Conserved hypothetical protein
6.04681	−142	**TTAGGGAAGCCTTGCCTAT**	MSMEG3634	Predicted 4-hydroxy benzoyl co-A hydrolase
5.97454	−10	CTAGGTTAGGCTACATTTA	MSMEG4518	TrpE2
5.9718	−67	TTAGGTAACGCTGACCTCA	MSMEG6385	BfrB
5.95729	−185	**ATAGCGAAGGCTAACCTAT**	MSMEG0020	FxuD protein
5.84773	−85	ATAGGTTAGCCTTCGCTAT	MSMEG0021	Aspartate 1-decarboxylase
5.82649	−62	TGAGGTAAGCCTAACTTAA	MSMEG6381	PheA
5.802	−190	TAAGGTTAGACTCCGCTAA	MSMEG3566	Fadd16
5.77512	−31	**TAAGGGTACGCTTACCTTA**	MSMEG5028	Iron utilization protein
5.74281	−37	TAAGCCTAGCCTACCTTAA	MSMEG2135	Acyl carrier protein Acp
5.71354	−146	ATAGGTAAGCCTAACTTTG	MSMEG1669	SdhC
5.69514	−51	CAAAGTTAGGCTTTCCTTA	MSMEG2263	Transcriptional activator Benr
5.66394	−130	GAAGGTAAAGCTACCCTCA	MSMEG2132	Siderophore biosynthesis protein
5.55146	−36	TGAGGCTAGCCTTCGTTAA	MSMEG0093	Hypothetical protein
5.54236	−375	GTCGGCAAGCCTTTCCTGA	MSMEG6536	β-Lactamase
5.49362	−53	GTAGAGCAGTCTCACCTAG	MSMEG5650	Citrate synthase I
5.47364	−278	CCAGGAAAGGCTCAACTGA	MSMEG6837	Enoyl-Coa hydratase
5.44417	−91	CAAAGTTAGGCTTACCTAT	MSMEG1670	Cytidine deaminase
5.43696	−57	TTAGCTTAGGCATACATAA	MSMEG0605	ATPase
5.42989	−20	**TTAGGTTACCCTCAGCTGT**	MSMEG0011	Iron utilization protein
5.41895	−325	GTAGGTCAATCTCAGCTCA	MSMEG1984	Conserved hypothetical protein
5.40656	−34	TATAGTAAGGCTAACCTAA	MSMEG3635	Hypothetical protein
5.39683	−12	CAAGGCTAGCCAAGGCTAA	MSMEG5398	Membrane protein
5.37462	−170	TTAGCCTTGGCTAGCCTTG	MSMEG5396	Conserved hypothetical protein
5.32741	−115	ATTGGTAAGCCTTACCTTT	MSMEG0016	MbtH protein–related protein
5.29507	−71	GTAGCTAATTCTGTCCTTC	MSMEG3327	Monooxygenase
5.29163	−78	CCAAGCGAGGCTGGCCTCA	MSMEG2989	Hypothetical protein
5.27999	−258	CGAGGGCAGTCTGTCCTTC	MSMEG0436	Peptidyl-arginine deiminase superfamily
5.27347	−47	ATAAGCAAAGCTATCGTCA	MSMEG3668	Antifungal protein precursor
5.2614	−337	TGAGGTAAACCTAATCTTG	MSMEG3643	CBS domain protein
5.26078	−50	TTTGGCAAGGCTATCGTTG	MSMEG1947	Uncharacterized membrane proteins
5.19307	−88	CGAGCACACGCTGGCCTCA	MSMEG2501	ABC transporter
5.19306	−72	CAAGATTAGGTTTACCTCA	MSMEG3642	Iron ABC transporter

8). The mycobactin biosynthesis operon is conserved across the IdeR regulons of mycobacteria, whereas the exochelin biosynthesis operon is present only in IdeR regulon of *M. smegmatis*.

The operon containing the genes Rv0282, Rv0283, and Rv0284 is also conserved across the predicted IdeR regulon of mycobacteria (Table 8). The gene Rv0282 predicted to code FtsK, a protein implicated to have role in cell division and peptidoglycan synthesis or modification [17,18]. The gene Rv0283 codes for a hypothetical protein. The gene Rv0284 code for the protein belonging to the AAA-superfamily of ATPases associated with a wide variety of cellular activities, including membrane fusion, proteolysis, and DNA replication [19].

Genes Rv1847 and Rv1846c were reported to have strong predicted IdeR binding site [13] are divergently transcribed in *M. tuberculosis*. Their cognate orthologues in other mycobacteria also show strong predicted IdeR binding site. Conservation of these genes across the predicted IdeR regulons of mycobacteria suggests that the genes could be potential targets of IdeR and could play impartant role in iron metabolism (Table 8).

The gene Rv1847 code for a predicted 4-hydroxy benzoyl coA thioesterase (*paaI*) and its downstream genes code for subunits of urease. Homologs of Rv1847 are widely distributed in sequenced genomes of bacteria, but none of them are characterized. Homologue of Rv1847 in *E. coli* is known as *ybdB*, which is associated with the genes *entA*, *entB*, *entE* and *entC* that code for the enzymes involved in siderophore (enterochelin) biosynthesis. Hence, it is likely that product of Rv1847 and its orthologs might be involved in the siderophore biosynthesis pathway.

The genes Rv1846c–Rv1845c and their cognate orthologs in other mycobacteria belong to the same operon. The gene Rv1846 codes for a BlaI family of transcription regulator and the other gene Rv1845c code for BlaR1 family of protein. The two families of proteins together confer resistance to variety of β-lactum antibiotics and widely distributed in pathogenic bacteria. In *Staphylococcus aureus*, BlaR1 family of protein MecR1, present in the cytoplasmic membrane, detects the presence of the β-lactum by means of an extracellular penicillin binding-domain and transmits the signal via a second intracellular zinc metalloprotease signaling domain. Binding of a β-lactum to MecR1 stimulates the autocatalytic conversion of intracellular zinc metalloprotease signaling domain of MecR1 from an inactive proenzyme to an active protease. The activated form of MecR1 cleaves BlaI family of transcription regulator, MecI and de-represses the transcription of *β*-lactamase [20,21]. It would be interesting to study what role these gene play in mycobacteria.

3.5. *Iron regulated genes that are present in other Mycobacterium species but not in* M. tuberculosis

Analysis of genes that are under the control of IdeR in *M. avium* and *M. smegmatis* reveals novel iron regulated genes,

Table 7
Predicted IdeR regulated operons in *M. smegmatis*

Gene	Product
MSMEG3564	Bacterioferritin
MSMEG4502	MbtH protein–related protein
MSMEG4503	MbtG
MSMEG4504	Peptide synthetase
MSMEG4505	Peptide synthetase
MSMEG4506	Polyketide synthase
MSMEG4507	Polyketide synthase
MSMEG4508	Dihydroaeruginoic acid synthetase
MSMEG4509	Dihydroaeruginoic acid synthetase
MSMEG4510	MbtA
MSMEG5992	Hypothetical protein
MSMEG5993*	Hypothetical protein
MSMEG6380	Phosphoglycerate mutase family
MSMEG6381	PheA
MSMEG6382	Conserved hypothetical protein
MSMEG6383	Hypothetical membrane protein
MSMEG3630	Urease accessory protein UreF
MSMEG3631	Urease
MSMEG3632	Urease
MSMEG3633	Urease
MSMEG3634	Predicted 4-hydroxy benzoyl co-A hydrolase
MSMEG3635	Hypothetical protein
MSMEG3636	Transcription regulator
MSMEG3637	Peptidase family M48 family
MSMEG4517	Ismsm2
MSMEG4518	TrpE2
MSMEG6385	BfrB
MSMEG0011	Iron utilization protein
MSMEG0012	Ferric exochelin uptake (FxuC)
MSMEG0013	Ferric exochelin uptake (FxuA)
MSMEG0014	Ferric exochelin uptake (FxuB)
MSMEG0015	Ferric exochelin biosynthesis (FxbA)
MSMEG0016	MbtH protein–related protein
MSMEG0017	Exit protein
MSMEG0018	Exit protein
MSMEG0019	Peptide synthetase homolog
MSMEG0020	FxuD protein
MSMEG0021	Aspartate 1-decarboxylase
MSMEG0022	L-Ornithine N5-Oxygenase
MSMEG3565	Hypothetical protein
MSMEG3566	Fadd16
MSMEG3567	BFD-like [2Fe–2S] binding domain family
MSMEG5028	Iron utilization protein
MSMEG2133	FadE13
MSMEG2134	AMP-binding family protein
MSMEG2135	Acyl carrier protein Acp
MSMEG1666	SdhB
MSMEG1667	Succinate dehydrogenase
MSMEG1668	Succinate dehydrogenase
MSMEG1669	SdhC
MSMEG1670	Cytidine deaminase
MSMEG2263	BenABC operon transcriptional activator BenR
MSMEG2132	Siderophore biosynthesis protein
MSMEG0093	Hypothetical protein
MSMEG6536	β-Lactamase
MSMEG5650	Citrate synthase I
MSMEG6837	Enoyl-Coa hydratase
MSMEG0605	ATPase
MSMEG0606	Protein of unknown function (DUF690) superfamily
MSMEG0607	Ftsk/SpoIIIE family protein
MSMEG0608	PE
MSMEG0609	PPE family domain protein
MSMEG0610	PE family protein
MSMEG0611	Conserved hypothetical protein
MSMEG1983	Monooxygenase
MSMEG1984	Conserved hypothetical protein
MSMEG1985	AMP-binding domain protein
MSMEG5398	Membrane protein
MSMEG5399	Putative lipoprotein
MSMEG5394	Protein of unknown function (DUF501) family
MSMEG5395	Enolase
MSMEG5396	Conserved hypothetical protein
MSMEG5398	FTR1, high-affinity Fe^{2+}/Pb^{2+} permease
MSMEG5399	Predicted periplasmic lipoprotein involved in iron transport
MSMEG5310	Predicted iron-dependent peroxidase
MSMEG3326	Drug transporter
MSMEG3327	Monooxygenase
MSMEG3328	Repressor protein
MSMEG2989	Hypothetical protein
MSMEG0436	Porphyromonas-type peptidyl-arginine deiminase superfamily
MSMEG0437	Conserved hypothetical protein
MSMEG0438	Amino acid permease
MSMEG3668	Antifungal protein precursor
MSMEG3640	Iron(III) ABC transporter
MSMEG3641	Iron ABC transporter
MSMEG3642	Iron ABC transporter
MSMEG3643	CBS domain protein
MSMEG3644	CBS domain protein
MSMEG3645	Conserved hypothetical protein
MSMEG3646	Malate synthase G
MSMEG3647	Conserved hypothetical protein
MSMEG1947	Uncharacterized membrane proteins

Note: Genes that are part of an operon are together.

predicted to be involved in iron transport. These include the genes that code for a predicted iron permease (MAP1761c, MSMEG5398), iron transporter (MAP1761c, MSMEG5399)

Table 8
Distribution of orthologues of IdeR regulated genes across mycobacteria

Gene	Mtub	Mavi	Mlep	Msme
Aromatic amino acid metabolism				
pheA	Rv3838c	MAP0193	*ML0078	MSMEG6381
fbp	Rv3837c	MAP0194	*ML0079	MSMEG6380
hisE	Rv2122c	*MAP1847c	ML1309	*MSMEG4186
hisG	Rv2121c	*MAP1846c	ML1310	*MSMEG4185
trpE2	Rv2386c	MAP2205c	-	MSMEG4518
paaI	Rv1847	MAP1560	-	MSMEG3634
Urease				
ureA	Rv1848	-	-	MSMEG3633
ureB	Rv1849	-	-	MSMEG3632
ureC	Rv1850	-	-	MSMEG3631
ureF	Rv1851	-	-	MSMEG3630
uerG	Rv1852	-	-	*MSMEG3628
ureD	Rv1853	-	-	*MSMEG1084
Fatty acid metabolism				
fadD	Rv1344	MAP1555c	-	MSMEG2135
fadE	Rv1345	MAP1554c	-	MSMEG2134
fadB	Rv1346	MAP1553c	-	MSMEG2133
Cell wall biosynthesis				
-	Rv1347	*MAP3149c	-	MSMEG2132
murB	Rv0482	MAP3975	ML2447	MSMEG0920
Siderophore biosynthesis				
mbtJ	Rv2385	MAP2197	-	-
mbtA	Rv2384	MAP2178	-	MSMEG4510
mbtB	Rv2383c	MAP2177c	-	MSMEG4509
mbtC	Rv2382c	MAP2175c	-	MSMEG4507
mbtD	Rv2381c	MAP2174c	-	MSMEG4506
mbtE	Rv2380c	MAP2173c	-	MSMEG4505
mbtF	Rv2379c	MAP2171c	-	MSMEG4504
mbtG	Rv2378c	MAP2170c	-	MSMEG4503
mbtH	Rv2377c	MAP1872c	-	MSMEG4502
Siderophore transport				
-	Rv1348	MAP2414c	-	MSMEG6516
-	Rv1349	MAP2413c	-	MSMEG6515
fecB	Rv3044	MAP3092	ML1729	*MSMEG2319
Iron storage				
bfrA	Rv1876	MAP1595	ML2038	MSMEG3564
bfrB	Rv3841	-	-	MSMEG6385
Transmembrane transport				
mmpL4	Rv0450c	*MAP3751	*ML2378	-
mmpS4	Rv0451c	*MAP1241c	*ML2377	*MSMEG0373
Membrane proteins				
	Rv0282	MAP3778	ML2537	MSMEG0605
	Rv0283	MAP3779	ML2536	MSMEG0606
	Rv0284	MAP3780	ML2535	MSMEG0607
Transcription regulators				
-	Rv1846c	MAP1559c	ML2063	MSMEG3636
-	Rv1845c	MAP1558c	ML2064	MSMEG3637
-	Rv1404	MAP1131	ML0550	MSMEG5546
acrR	Rv0452	*MAP3945	-	-
PPE family				
-	Rv2123	-	-	-
Hypothetical proteins				
-	Rv3402c	-	-	*MSMEG2303
-	Rv3403c	-	-	-
-	Rv2663	-	-	-

Notes: 1. Genes that are part of an operon are together. 2. '-' Represents, corresponding orthologs are not present in the genome. 3. '*' Represents, upstream sequence of corresponding genes do not have predicted IdeR bindings site. 4. Mtub – *Mycobacterium tuberculosis*; Mavi – *Mycobacterium avium* subsp. *paratuberculosis*; Mlep – *Mycobacterium leprae*; Msme – *Mycobacterium smegmatis*.

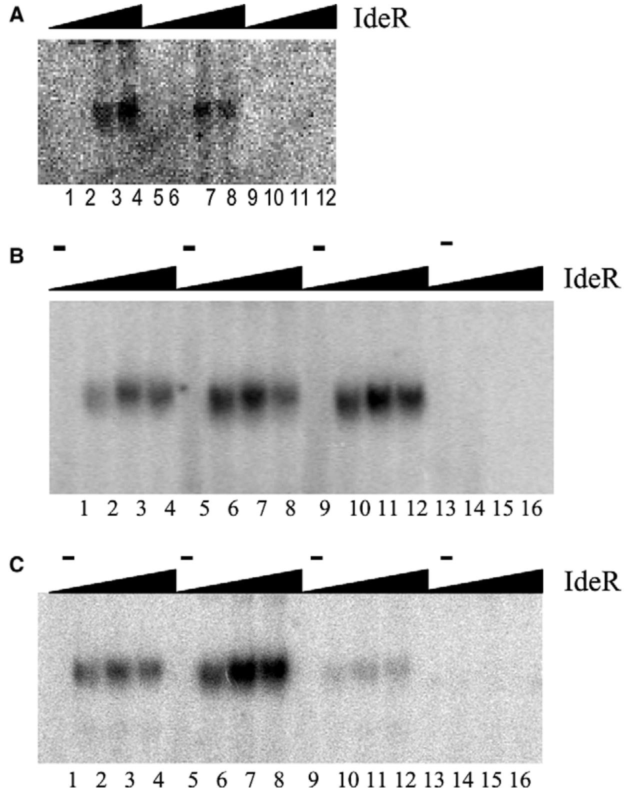


Fig. 2. IdeR binds to the predicted IdeR binding regulatory motifs in *M. smegmatis*. The lanes indicated by (-) have the probe alone without IdeR. Increasing concentration of IdeR was added to ^{32}P-labelled DNA probes in the presence of 200 μM Ni^{2+} and complexes were resolved on a 7% Tris–borate polyacrylamide gel. Binding conditions and gel electrophoresis are described in Section 2. (A) Radiolabeled IdeR binding motif (ds-5′TTAGGATAGCTTTACCTAA-3′) as positive control (lanes 1–4), radiolabeled MSMEG0020 motif (lanes 5–8), radiolabeled motif without IdeR binding site (5′-TTTTCAT-GAAGTCTTCTAA-3′) (lanes 9–12), IdeR was added in increasing concentration from 0 to 10 pmol. (B) Radiolabeled MSMEG5028 motif (lanes 1–4), radiolabeled IdeR binding motif (lanes 5–8), radiolabeled MSMEG3634 motif (lanes 9–12), radiolabeled motif without IdeR binding site (lanes 13–16). (C) Radiolabeled MSMEG0011 motif (lanes 1–4), radiolabeled IdeR binding motif (lanes 5–8), radiolabeled MSMEG6382 motif (lanes 9–12), radiolabeled motif without IdeR binding site (lanes 13–16). IdeR was added in increasing concentration from 5 to 20 pmol (B and C).

and iron dependent peroxidase (MAP1760c, MSMEG5310). These genes are part of same operon and are well represented in sequenced bacterial genomes. They could play a similar role in iron transport as reported in case of *Candida albicans* and *Saccaromyces cerviciae* [22,23]. Further, the peroxidase dependent iron transport system could also have role in peroxide stress defense.

In addition to the peroxidase dependent transport system, IdeR can also regulate the genes that code for predicted siderophore interacting protein (nfa7590, MSMEG5028, MSMEG0011) and *Mycobacterium* natural-resistance-associated macrophage protein, Mramp (MAP4065), in *M. avium* subsp. *paratuberculosis*. The genes that code for the siderophore interacting protein are similar to the *viuB* of *Vibrio cholerae* [24]. The protein, Mramp, is an ortholog of natural-resistance-associated macrophage protein (Nramp) and is known to compete with later for the same divalent-cations, for intracellular survival of mycobacteria [25].

3.6. *Iron regulated genes that are specific to* M. tuberculosis

The genes Rv0450c (*mmpL4*), Rv0451c (*mmpS4*), Rv3402c, Rv3403c, Rv2123 (PPE), and Rv2663 are specific to IdeR regulon of pathogenic *Mycobacterium*, *M. tuberculosis.* The protein, MmpL4, belongs to a family of MmpL proteins that are known to have a role in virulence and drug resistance [26].

4. Summary

Computational analysis of IdeR regulated genes across the mycobacteria has lead to identification of many conserved iron regulated genes. These include genes involved in aromatic amino acid metabolism, fatty acid metabolism, siderophore biosynthesis, siderophore transport as well as iron storage. The genes, which code for predicted 4-Hydroxy benzoyl coA hydrolase (Rv1847), and a protease dependent antibiotic regulatory system (Rv1846c, Rv0185c) were also observed as one of the most conserved IdeR regulated genes. In addition, IdeR can also regulate the genes that code for predicted citrate dependent iron transport system (FecB), siderophore interacting protein (ViuB). Analysis of IdeR regulated genes in *M. avium* subsp. *paratuberculosis* and *M. smegmatis* identifies novel iron regulated genes, which code for predicted iron permease, iron transporter and iron dependent peroxidase, which could be involved in iron transport, peroxide stress defense and control of intracellular iron levels. We also predicted the gene encoding natural-resistance-associated macrophage protein (Mramp) in *M. avium* subsp. *Paratuberculosis* and an operon for the biosynthesis of exochelin in *M. smegmatis* to be iron regulated.

Acknowledgement: We thank TIGR for making unfinished M.smegmatics MC2 sequences available through their web portal. Research in A.R.'s laboratory is supported by grants from Council of Scientific & Industrial Research (CSIR) NMITLI, Department of Biotechnology, Department of Science and Technology, Govt. of India. S.R. is supported by CSIR NMITLI Grant. Y.S. is supported by CSIR Research Fellowship and V.V. is supported by UGC fellowship.

References

[1] Castagnetto, J.M., Hennessy, S.W., Roberts, V.A., Getzoff, E.D., Tainer, J.A. and Pique, M.E. (2002) MDB: the Metalloprotein Database and Browser at The Scripps Research Institute. Nucleic Acids Res. 30, 379–382.

[2] Ho, W.L., Yu, R.C. and Chou, C.C. (2004) Effect of iron limitation on the growth and cytotoxin production of *Salmonella choleraesuis* SC-5. Int. J. Food Microbiol. 3, 295–302.

[3] Martinez, A. and Kolter, R. (1997) Protection of DNA during oxidative stress by the non specific DNA-binding protein Dps. J. Bacteriol. 179, 5188–5194.

[4] Smith, J.L. (2004) The physiological role of ferritin-like compounds in bacteria. Crit. Rev. Microbiol. 30 (3), 173–185.

[5] Ratledge, C. (2004) Iron, mycobacteria and tuberculosis. Tuberculosis 84, 110–130.

[6] Litwin, C.M. and Calderwood, S.B. (1993) Role of iron in regulation of virulence genes. Clin. Microbiol. Rev. 6, 137–149, (review).

[7] Schmitt, M.P., Predich, M., Doukhan, L., Smith, I. and Holmes, R.K. (1995) Characterization of an iron-dependent regulatory protein (IdeR) of *Mycobacterium tuberculosis* as a functional homolog of the diphtheria toxin repressor (DtxR) from *Corynebacterium diphtheriae*. Infect Immun. 11, 4284–4289.

[8] 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. 5, 287–298, Erratum in: Tuber Lung Dis. 6, 382.

[9] Gold, B., Rodriguez, G.M., Marras, S.A., Pentecost, M. and Smith, I. (2001) The *Mycobacterium tuberculosis* IdeR is a dual functional regulator that controls transcription of genes involved in iron acquisition, iron storage and survival in macrophages. Mol. Microbiol. 3, 851–865.

[10] Rodriguez, G.M., Voskuil, M.I., Gold, B., Schoolnik, G.K. and Smith, I. (2002) IdeR, An essential gene in mycobacterium tuberculosis: role of IdeR in iron-dependent gene expression, iron metabolism, and oxidative stress response. Infect Immun. 7, 3371–3381.

[11] Yellaboina, S., Seshadri, J., Kumar, M.S. and Ranjan, A. (2004) PredictRegulon: a webserver for the prediction of the regulatory protein binding sites and operons in prokaryote genomes. Nucleic Acids Res. 32, W318–W320.

[12] Yellaboina, S., Ranjan, S., Chakhaiyar, P., Hasnain, S.E. and Ranjan, A. (2004) Prediction of DtxR regulon: identification of binding sites and operons controlled by Diphtheria toxin repressor in *Corynebacterium diphtheriae*. BMC Microbiol. 4, 1471–2180.

[13] Prakash, P., Yellaboina, S., Ranjan, A. and Hasnain, S.E. (2005) Computational prediction and experimental verification of novel IdeR binding sites in the upstream sequences of *Mycobacterium tuberculosis* open reading frames. Bioinformatics 21, 2161–2166.

[14] Marchler-Bauer, A., Anderson, J.B., DeWeese-Scott, C., Fedorova, N.D., Geer, L.Y., He, S., Hurwitz, D.I., Jackson, J.D., Jacobs, A.R., Lanczycki, C.J., Liebert, C.A., Liu, C., Madej, T., Marchler, G.H., Mazumder, R., Nikolskaya, A.N., Panchenko, A.R., Rao, B.S., Shoemaker, B.A., Simonyan, V., Song, J.S., Thiessen, P.A., Vasudevan, S., Wang, Y., Yamashita, R.A., Yin, J.J. and Bryant, S.H. (2003) CDD a curated Entrez database of conserved domain alignments. Nucleic Acids Res. 31, 383–387.

[15] Quadri, L.E., Sello, J., Keating, T.A., Weinreb, P.H. and Walsh, C.T. (1998) Identification of a *Mycobacterium tuberculosis* gene cluster encoding the biosynthetic enzymes for assembly of the virulence-conferring siderophore mycobactin. Chem. Biol. 5, 631–645.

[16] Pelludat, C., Brem, D. and Heesemann, J. (2003) Irp9, encoded by the high-pathogenicity island of *Yersinia enterocolitica*, is able to convert chorismate into salicylate, the precursor of the siderophore yersiniabactin. J. Bacteriol. 18, 5648–5653.

[17] Begg, K.J., Dewar, S.J. and Donachie, W.D. (1995) A new *Escherichia coli* cell division gene, ftsK. J. Bacteriol. 21, 6211–6222.

[18] Daniel, R.A., Williams, A.M. and Errington, J. (1996) A complex four-gene operon containing essential cell division gene pbpB in *Bacillus subtilis*. J. Bacteriol. 8, 2343–2450.

[19] Frickey, T. and Lupas, A.N. (2004) Phylogenetic analysis of AAA proteins. J. Struct. Biol. 1–2, 2–10.

[20] Hanique, S., Colombo, M.L., Goormaghtigh, E., Soumillion, P., Frere, J.M. and Joris, B. (2004) Evidence of an intramolecular interaction between the two domains of the BlaR1 penicillin receptor during the signal transduction. J. Biol. Chem. 14, 14264–14272.

[21] Wilke, M.S., Hills, T.L., Zhang, H.Z., Chambers, H.F. and Strynadka, N.C. (2004) Crystal structures of the Apo and penicillin-acylated forms of the BlaR1 β lactam sensor of *Staphylococcus aureus*. J. Biol. Chem. 45, 47278–47287.

[22] Stearman, R., Yuan, D.S., Yamaguchi-Iwai, Y., Klausner, R.D. and Dancis, A. (1996) A permease-oxidase complex involved in high-affinity iron uptake in yeast. Science 271, 1552–1557.

[23] Ramanan, N. and Wang, Y. (2000) A high-affinity iron permease essential for *Candida albicans* virulence. Science 288, 1062–1064.

[24] Butterton, J.R. and Calderwood, S.B. (1994) Identification, cloning, and sequencing of a gene required for ferric vibriobactin utilization by *Vibrio cholerae*. J. Bacteriol. 18, 5631–5638.

[25] Agranoff, D., Monahan, I.M., Mangan, J.A., Butcher, P.D. and Krishna, S. (1999) *Mycobacterium tuberculosis* expresses a novel pH-dependent divalent cation transporter belonging to the Nramp family. J. Exp. Med. 5, 717–724.

[26] Domenech, P., Reed, M.B. and Barry, C.E. (2005) Contribution of the *Mycobacterium tuberculosis* MmpL protein family to virulence and drug resistance. Infect Immun. 6, 3492–34501.