Studies on cAMP Receptor Proteins from mycobacteria

Thesis submitted to

Department of Biochemistry
School of Life Sciences
University of Hyderabad
HYDERABAD
INDIA

For the degree of
Doctor of Philosophy

By

Yusuf Akhter
Bachelor of Science, University of Lucknow, 2002
Master of Science, Hamdard University, 2004

Laboratory of Molecular and Cellular Biology
Centre for DNA Fingerprinting and Diagnostics
HYDERABAD
INDIA

Registration Number: 05LBPH04
2009

CERTIFICATE

This is to certify that the thesis entitled, **"Studies on cAMP Receptor Proteins from mycobacteria"** submitted by **Mr. Yusuf Akhter** for the Degree of **Doctor of Philosophy** to University of Hyderabad 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.

Prof. Seyed E. Hasnain
Thesis Supervisor
Vice Chancellor
University of Hyderabad

Dean, School of Life Sciences
University of Hyderabad

Head, Department of Biochemistry
University of Hyderabad

DECLARATION

The research work presented in this thesis entitled **"Studies on cAMP Receptor Proteins from mycobacteria"**, has been carried out by me at the Centre for DNA Fingerprinting and Diagnostics, Hyderabad, under the guidance of **Prof. Seyed E. Hasnain.** 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.

Yusuf Akhter

PhD Candidate

'We dance around in a ring and suppose, the secret sits in the middle and knows'

– Robert Frost

ACKNOWLEDGEMENTS

The work presented in this thesis was accomplished with the help of many colleagues and friends. It is a pleasant opportunity to express my gratitude to thank all the people who have helped me directly or indirectly in their various capacities during the tenure of my Ph.D.

My heartfelt thanks to Prof. Seyed E. Hasnain whom I have worked with since June, 2004. He provided a motivating, enthusiastic and critical atmosphere during discussions. I thank him for his immense support, outstanding guidance and encouragement during my stay. It was a great pleasure for me to conduct this thesis under his supervision. I owe him lots of gratitude for giving the freedom of thought, amiable environment and motivation which provided in me a confidence in analyzing my research problems. In spite of his busy schedule as the Vice Chancellor of the University of Hyderabad, he was always available for discussing the progress of my research work. He, as my supervisor, has provided constructive comments during my thesis as well as on the preliminary formalities towards the completion of this thesis. I feel privileged to be associated with him and words fail to express my deepest regards towards him.

I also thanks to Dr. Matthias Wilmanns, who was kindly agreed to be my host supervisor at European Molecular Biology Laboratory, Hamburg, Germany for my DAAD fellowship stay. He was always excited about my projects and was interested in discussions despite of his busy schedule as the head of EMBL-outstation. I would like to extend my thanks to Dr. Vivian Pogenberg, Dr. Santosh Panjikar and Dr. Simon Holton for helping me with methods in crystallography. I also want to acknowledge Christian Poulsen, who is always a friend indeed and stand by me whenever I needed any kind of support.

A special thanks to Dr. Krishnaveni Mohareer for being a commendable senior colleague. Her considerate intellectual and moral supports throughout the course of study have been incredible. It would be unfair if I would not thank Sandeep, he is all-time good friend and colleague.

I would like to express my thanks to Director, CDFD, Dr. J Gowrishankar for all his support. I am thankful to the Registrar and the Head, Dept. of Biochemistry at University of Hyderabad for permitting me to register as a Ph.D. student.

I am deeply indebted and would like to express my sincere thanks to Dr. Niyaz Ahmed, Dr. Shekhar C. Mande, Dr. Sangita Mukhopadhaya and Dr. Murali Bashyam for their constant encouragement and help. I am immensely grateful to Dr. Nasreen Z. Ehtesham for her kind help and the joyful discussions. I also take this opportunity to thank Nitin Pathak for excellent support during the tenure.

I am thankful to CSIR for the financial support given to me as PhD student. I am also thankful to DAAD (German Academic Exchange Service) for a long term doctoral fellowship. Project support from Department of Biotechnology and the Centre for DNA Fingerprinting and Diagnostics is also acknowledged. I would like to thank “The Bill and Melinda Gate’s foundation” for Global Health travel award which was utilized to attend “Keystone Symposia” on tuberculosis held at Keystone resort, Colorado, USA, where I have presented the work. A travel grant from the organizers is acknowledged to attend “International School of Biological Crystallization” held at Granada, Spain where I have presented a poster. I am thankful to Department of Science and technology (Govt. of India) for nomination and travel award to attend “Lindau Meeting of Nobel laureates and Young researchers 2007”, Germany.

I also owe sincere thanks to the people in the CDFD and EMBL Administration. The work would have been not possible without their help.

I am highly thankful to Khalid, Kaiser, Khursheed and Sreejit for their support as my friends who together with my other lab-mates Zameer, N. Sudhir, Wasim and Krishnamurthy made a wonderful working atmosphere in the lab. We had great time both in and out of the lab in the form of picnic, watching movies, and delicious parties.

Many cheerful thanks to JRF 2004 batch: Sandeep, Tabrez, Ratheesh, Devi, Debashree, Shiny, Kaiser, Aisha, Gita and Jisha for being good friends.

My parents have been constant source of unflinching support and encouragement and this has been a great source of strength and motivation for me to accomplish my objectives. My sisters (Hina, Mina and Uzma) and brothers (Tabraiz and Tabish) have extended all their affection and moral support without which it would not have been possible to pursue my work. I dedicate this thesis to my family.

Yusuf

Table of Contents

Chapter 2

cAMP Receptor Protein Regulons in mycobacteria **34-64**

Chapter 3

Biophysical and Biochemical features of *Mtb*-CRP **65-85**

Chapter 4

List of figures

List of tables

List of Abbreviations

˚C	:	Degree centigrade
aa	:	amino acid
AU	:	Arbitrary Units
ATP	:	Adenosine-5'-triphosphate
BCG	:	Bacillus of Calmette Guerin
BSA	:	bovine serum albumin
bp	:	base pair
CRP	:	cAMP receptor protein
cAMP	:	3'-5'-cyclic adenosine monophosphate
DNA	:	Deoxyribonucleic acid
dNTP	:	Deoxynucleotide triphosphate
DNAse	:	deoxyribonuclease
E. coli	:	*Escherichia coli*
EMSA	:	electrophoretic mobility shift assay
EtBr	:	Ethidium Bromide
FNR	:	Fumarate Nitrate reductase reguator
Frd	:	Fumarate reductase
HEPES	:	N-(2-hydroxyethyl)piperazine-N'-(2-ethanesulfonic acid)
IPTG	:	Isopropyl-b-D-thiogalactopyranoside
INH	:	Isoniazid
kb	:	Kilo base pair
kDa	:	Kilo Dalton(s)
KOH	:	potassium hydroxide
Mtb	:	*Mycobacterium tuberculosis*
mg	:	Milli gram (10^{-3} gram)
min	:	Minute(s)
ml	:	Millilitres (10^{-3} litres)
mM	:	Millimolar
nM	:	Nanomoles (10^{-9} moles)

mmol	:	Millimoles (10-3 moles)
ng	:	Nano gram (10-9 gram)
NRP	:	Non replicating persistent stage
OD	:	Optical density
ORF	:	Open reading frame
PAGE	:	polyacrylamide gel electrophoresis
PBS	:	Phosphate buffer saline
PCR	:	Polymerase chain reaction
Pfam	:	Protein family search database
PMSF	:	Phenylmethanesulphonylfluoride
rpm	:	Rotations per minute
SDS	:	Sodium dodecyl sulfate
TB	:	Tuberculosis
V	:	Volts
WHO	:	World Health Organization
µg	:	Micro gram (10^{-6} grams)
µl	:	Micro litre (10^{-6} litre)
µM	:	Micro molar (10^{-6} Molar)

Chapter 1: cAMP receptor proteins from mycobacteria - An Introduction

1.1 Tuberculosis: Infection to cure

Tuberculosis (TB) is a serious infectious disease caused by *Mycobacterium tuberculosis* (*Mtb*) or closely related mycobacterial strains. The infection is acquired in similar way as common cold by inhaling microscopic droplets from the atmosphere, in this case containing the bacteria. The droplets are transmitted from the respiratory tract of a contagious person when talking, coughing or spitting and can remain airborne for hours. Because only few bacteria are needed to start an infection the transmission is very efficient (Schluger and Rom, 1998; WHO factsheet, 2006). Once the bacteria enter the lungs it will on site be recognized by alveolar macrophages or dendritic cells which will phagocytose the bacteria. From this point there will be four possible outcomes: 1) the bacteria will enter the phagolysosomes and be killed leaving the host with no risk of developing TB at any time; 2) the bacteria will multiply and grow immediately resulting in a primary infection; 3) the bacteria will become dormant and never cause disease and leave the host non-contagious and with no symptoms and 4) the dormant bacteria will be reactivated at a later stage and cause a disease (Figure 1.1; Schluger and Rom, 1998).

The molecular mechanism behind these outcomes remains unclear. However, TB is always associated with the presence of macroscopic structures in the lung known as tubercles, also giving name to the disease.

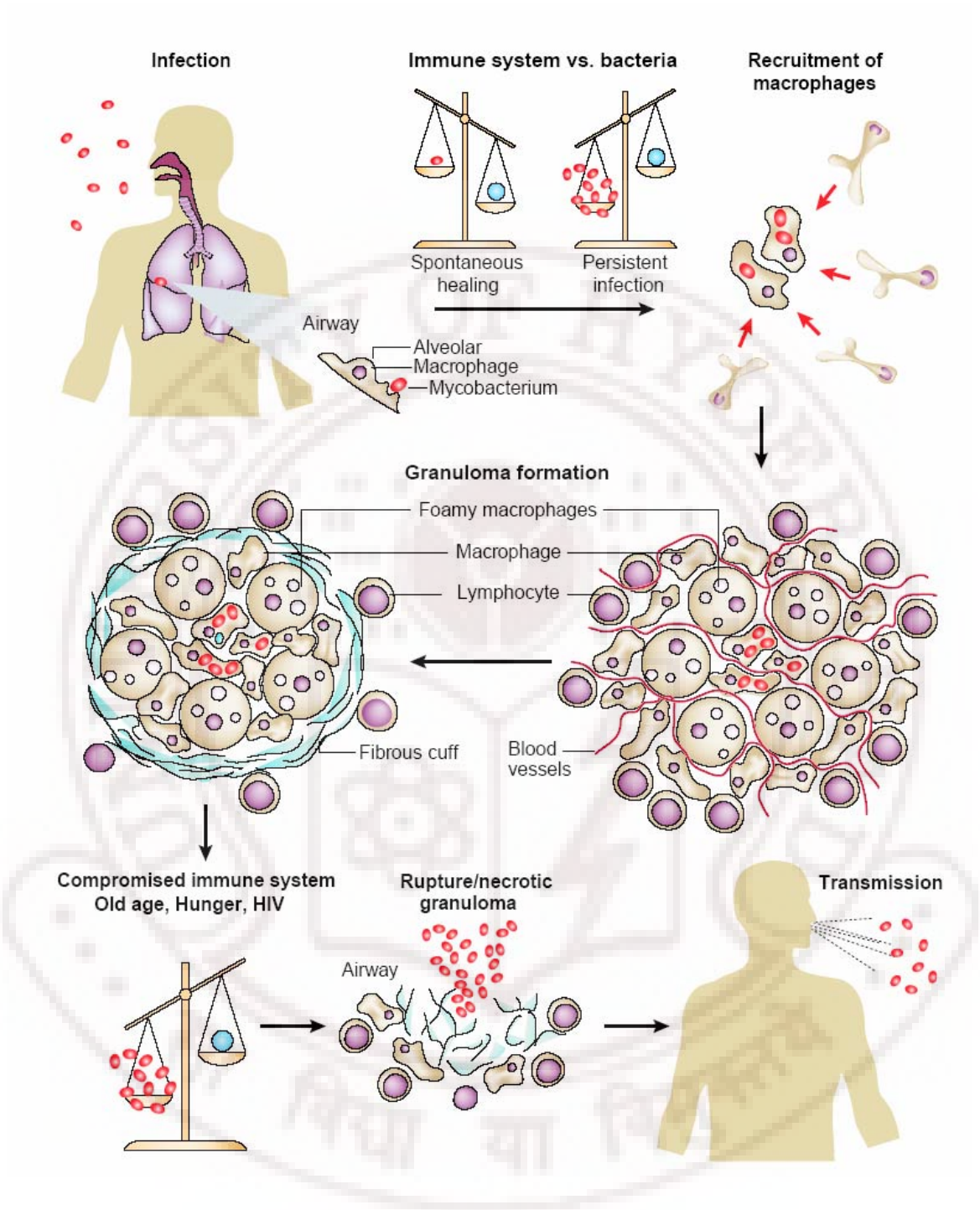


Figure 1.1: The infection cycle of *Mtb*. The bacteria are inhaled and phagocytosed by alveolar macrophages. The recognition and uptake of the mycobacteria will lead to a located pro-inflammatory response that attracts white blood cells and fibroblasts. The recruited cells will build the granuloma around the infected macrophages, and start to seal the area with collagen and other extracellular matrix components. In the containment/latent phase the fibrous cuff is advanced and the number of blood vessels in the granuloma is markedly reduced. During the containment phase the infection has no symptoms and is not contagious.

Under conditions where the immune system is compromised, the bacteria can escape from the degraded/ruptured granulomas. Now the bacteria can again easily be transmitted to the atmosphere and start a new infection cycle. The figure has been modified from [Schluger and Rom, 1998].

Tubercles are a persistent group of mononuclear phagocytes and associated cells that are clumping together as a consequence of the infection (Adams, 1976). This structure is often referred to as "granuloma" (Adams, 1976). The granuloma are built after the infected cells migrate deeper into the tissues where they induce a localized pro-inflammatory response that recruits additional mononuclear phagocytes to the area. The phagocytes will start to organize in the complex granuloma structure, encapsulating the infected macrophages, by specialized phagocytes and surrounding cells. The area gets finally sealed off with a fibrous cuff of collagen and other extracellular matrix components (Russell, 2007). Under these conditions the bacteria becomes dormant and the host has no symptoms. The bacteria can remain inside the host for decades and reactivate from the dormant state in case of failure of immune surveillance (Lillebaek *et al.*, 2002; Stewart *et al.*, 2003). When the bacteria escape the granuloma it effectively reinfect the lung tissues of the host and thereby allow it to spread to a new host. Infection with *Mtb* for this reason could be of serious consequence in conditions of severe malnutrition, HIV infection or any situation that impairs the immune system (Schaible *et al.*, 2007; Corbett *et al.*, 2003). The World Health Organization, has

estimated that about two billion people are infected by *Mtb* or similar mycobacteria (WHO factsheet). Of these 5-10% will develop a progressive disease (WHO factsheet, 2006; Schluger and Rom, 1998). Left untreated TB is fatal, but low cost medicine have existed for half a century with cure rates up to 95% (Frieden *et al.*, 2003). A downside of the current therapy is however the long treatment period of 6 months, which results in many incomplete treatments. This have, over many years, lead to growing number of drug-resistant, multidrug-resistant and extensively drug-resistent strains (Aziz *et al.*, 2006). Currently the situation is calling for new effective drugs or vaccines and more knowledge about these infectious bacteria (Sacchettini *et al.*, 2008).

1.1.1 Models to study tuberculosis

TB is not restricted to humans, but also commonly seen in mammals (Murphy *et al.,* 2008). One way to better understand the infection cycle is therefore to use animal models. When analyzing data from such models it is important to note that under natural conditions, *Mtb* exclusively infects human (Brosch et at, 2002). Infection with *Mtb* is however not restricted to human, and under experimental conditions *Mtb* is often administrated to immuno-defficient animals, injected intravenous or inhaled in high dose (Stewart *et al.,* 2003).

For historical reasons, mice have been the first choice for *in vivo* experiments, due to the ready availability of mutants and their relative ease to genetically manipulate. Two drawbacks of the mice model are the difficulty in controlling the dormant stage of the infection cycle, and the formation and the structure of the granuloma which can not be compared with humans (Flynn, 2006). The guinea pig and rabbits are thought to be better models, as their granulomas are well organized and become necrotic as in humans. On the other hand guinea pigs are very susceptible for infection at even very low dose of *Mtb*, whereas rabbits are resistant to infection and need to be infected with a related mycobacteria (Flynn, 2006). As in mice, it is therefore difficult to investigate the dormant stage of infection in these models. Although debatable, monkeys have recently been considered as a model to investigate TB. These non-human primates develop TB as in humans, and the data obtained are of more value than any other animal model (Lin *et al.,* 2006). The infection cycle in monkeys includes all the four possible outcomes of infection mentioned, including the latent stage and a more complex granuloma formation similar to human (Lin *et al.,* 2006). Although different animal models and homozygotic strains make it difficult to extrapolate the data (Converse *et al.,* 2009), these *in vivo* models have been used successfully and widely in investigating different virulent mycobacterial strains including genetically manipulated strains. It is indeed very important to understand how bacteria survive inside the phagosome and how bacteria

translocate within and between cells. Therefore, cultured cell lines and primary derived macrophages have also taken a central position in TB research (Vergne *et al.,* 2004).

More unexpected approaches using amoeba and zebrafish have also been used (Tobin and Ramakrishnan, 2008; Hagedorn *et al.*, 2009). The zebrafish is normally infected with *M. marinum* which is a genetically close relative of *Mtb* (Tobin and Ramakrishnan, 2008). Interestingly, the granuloma structure found in infected zebrafish resembles that found in human, increasing the value of the model (Tobin and Ramakrishnan, 2008). In contrast to other models, the zebrafish has the obvious advantage of being transparent. This makes it possible to investigate the infection by microscopy in real time at the cellular level *in vivo*, bringing the two disciplines together.

1.1.2 Inside the macrophage

Besides single experiments using samples from patients (Mwandumba *et al.*, 2004), most data on phagocytosed mycobacteria have been carried out on primary macrophages or cell lines (Vergne *et al.*, 2004). Under normal conditions macrophages effectively kill microorganisms by directing them to compartments known as phagolysosomes, having toxic oxygen and nitrogen metabolites, low pH and hydrolases (Nathan and Shiloh, 2000; Anes *et al.*, 2006). Some

pathological bacteria including *Mtb* can however block the phagosome maturation and escape these hostile conditions (Russell, 2007; Ray *et al.*, 2009). The secreted or released bacterial proteins and lipids could provide clues to the ability of *Mtb* to arrest phagosome maturation (Vergne *et al.*, 2004). Inside the bacteria-friendly vacuole the bacteria can replicate and enter into dormancy (Russell, 2007). Recent data reveal that *Mtb* enters matured phagolysosomes and escapes into the cytoplasm after more than two days (Wel *et al.*, 2007). A similar strategy has been found for a number of gram positive and negative bacteria (Ray *et al.*, 2009) as well as for related mycobacteria *M. marinum* (Tobin and Ramakrishnan, 2008). However these bacteria escape the unwanted conditions of the phagolysosome within 30 minutes (Ray *et al.*, 2009). It is not known how mycobacteria escape the phagolysosomes, however in other related bacteria specific proteins are involved that can disrupt the phagosomal membrane (Ray *et al.*, 2009). It remains to be found if *Mtb* use similar strategies and how they avoid other stresses like acidification, hypoxia and starvation within the phagocytes.

1.1.3 Mycobacterial apparatus against antibacterial armory:

Normally macrophages swiftly kill microorganism by translocating them to the phagolysosomes, cellular compartments containing low pH, hydrolases, toxic oxygen and nitrogen metabolites. However, *Mtb* seems

to arrest phagosome maturation (Kinchen & Ravichandran, 2008). The thick mycobacterial cell wall provides a degree of resistance against the microbicidal mechanisms of macrophages, which is complemented by the expression of a range of enzymes that can detoxify oxidative radicals. In addition, as with other slow-growing mycobacteria, live *Mtb* interfere with intracellular trafficking events after uptake by phagocytes, allowing the bacteria to occupy an immature phagosomal compartment that is screened from the most potent antimicrobial armoury. These properties allow *Mtb* to replicate inside macrophages during the early stages of infection. Dormant *Mtb* resides within harsh macrophage environment for decades. It is important to understand the host pathogen interactions to learn how *Mtb* circumvents the host defenses and causes disease or stays latent without being detected in host for decades.

1.2 Transcription regulators: New targets in combating old threat

The increasing emergence of drug-resistant TB, especially multidrug-resistant TB (MDR-TB, resistant to at least two frontline drugs such as isoniazid and rifampin), is particularly alarming. MDR-TB has already caused several fatal outbreaks and poses a significant threat to the treatment and control of the disease in some parts of the world, where the incidence of MDR-TB can be as high as 14% (WHO, 2003). The standard TB therapy is ineffective in controlling MDR-TB in high MDR-

TB incidence areas (Kimerling *et al.*, 1999). Of the estimated 0.5 million cases of MDR-TB in 2007, 27 countries (of which 15 are in the European region) together account for 85% of all such cases. According to WHO, XDR-TB is a future threat defined as TB that has developed resistance to at least rifampicin and isoniazid (resistance to these first line anti-TB drugs, MDR-TB), as well as to any member of the quinolone family and at least one of the following second-line anti-TB injectable drugs: kanamycin, capreomycin, or amikacin. By the end of 2008, 55 countries and territories had reported at least one case of extensively drug resistant TB (XDR-TB) (http://www.who.int/en/; WHO-factsheet, 2009). There is much concern that the TB situation may become even worse with the spread of HIV worldwide, a virus that weakens the host immune system and allows latent TB to reactivate and makes the person more susceptible to re-infection with either drug-susceptible or drug-resistant strains. The combination of drug resistant TB and HIV infection is a growing problem that presents serious challenges for effective TB control. In view of this situation, the World Health Organization (WHO), declared TB a global emergency (WHO, 1993). There is an urgent need to develop new TB drugs. However, no new TB drugs have been developed after the last one was developed more than 40 years. Although TB can be cured with the current therapy, the six months needed to treat the disease is too long, and the treatment often has significant toxicity. These factors make patient compliance to therapy very difficult, and this

noncompliance frequently selects for drug-resistant TB bacteria. The current TB problem clearly demonstrates the need for a re-evaluation of our knowledge of physiology of *Mtb* and the need for characterizing new and better drugs targets. Therapeutic interventions that are not only active against drug-resistant TB, but also, more importantly, can shorten the requirement for six months of therapy.

To rationally develop new antitubercular agents, it is essential to study the comparative genetics and physiology of *Mtb* and other non-pathogenic mycobacteria. The post-genomic TB era has now enabled researchers to undertake a global analysis of genetic differences between *Mtb*, *M. bovis* and various BCG substrains encompassing almost every open reading frame. Among the genes present in virulent strains but deleted from BCG strains, there seems to be no classical virulence elements, but there does seem to be an over-representation of transcriptional regulators, both repressors and activators. The adaptability of *Mtb* to its environment is underpinned by a complex array of these over 200 annotated transcriptional regulators. It is speculated that these transcriptional regulators may be involved in adapting to environmental changes like the intracellular hostile environment.

1.2.1 cAMP receptor proteins/Fumarate and nitrate reductase regulators: physiological switch to global stress response

Bacteria live in extremely diverse environments. These environments are not stable and expose bacteria to rapid changes and often a paucity of resources (including hostile intracellular situations). Bacteria must have the capacity to respond quickly to transitory conditions and imbalanced resources by activating alternative gene programs in order to make use of short-lived opportunities or to shut of unneeded metabolic routes. The CRP/FNR family of regulators is well adapted to potentiate bacterial metabolic versatility. Functions span the control of virulence factors, enzymes of aromatic ring degradation, nitrogen fixation, photosynthesis, and various types of respiration, making CRP/FNR regulators a most versatile group. The first member of the family was the cyclic adenosine monophosphate (cAMP)-binding protein, CRP (cAMP receptor protein), also called CAP for catabolite activator protein. This protein represents the paradigm of a genetic regulator, its properties having attained textbook status. It was found that oxygen or redox sensitive gene regulatory circuits are controlled by another global regulator called FNR. FNR stands for fumarate and nitrate reductase regulator and involved in regulating of anaerobic metabolic pathways of *Escherichia coli* [Shaw *et al.*, 1983]. A remarkable versatility of sensory mechanisms integrated into the CRP/FNR scaffold has been defined. The different signal molecules recognized by CRP/FNR

regulators as well as their sensory modules have been also discussed. These aspects include among others the signals NO, CO, 2-oxoglutarate and temperature and a binding capability of CRP/FNR regulators beyond that of cAMP and a [4Fe-4S] cluster [Figure 1.2].

Using common structural features and functional considerations, Fischer proposed a tripartite classification of CRP/FNR regulators [Fischer, 1994] into the groups CRP, NtcA, and FNR, of what were then only 22 proteins. The FNR cluster was divided into FNR-type regulators and the FixK subgroup. By about mid 1999 the number of regulators had risen to 56, for which the original classification held up in a phylogenetic analysis, except for introduction of a new group of Dnr regulators [Vollack *et al.*, 1999]. A better analysis comprising 64 regulators maintained a virtually constant picture [Green *et al.*, 2001]. Due to the rapid advances in genome analysis, the number of candidate members of the CRP/FNR family has now increased enormously.

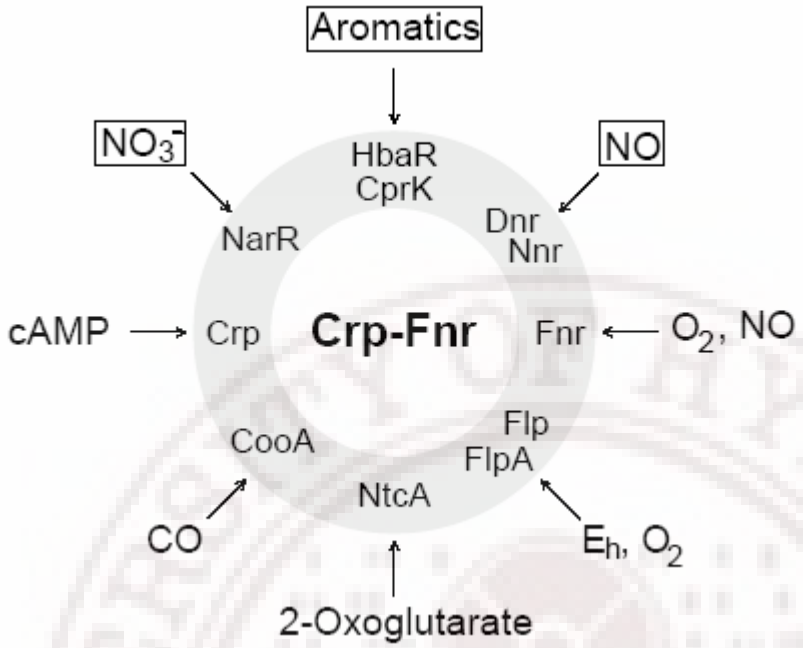


Figure 1.2: Signals processed by regulators of the CRP/FNR family. Boxed signals have not been proven to interact directly with the cognate regulator. Eh, midpoint potential. Figure has been modified from Köner *et al.*, 2003.

The CRP/FNR proteins are rather constant in size with approximately 230-250 amino acid residues in most members. The proteins are characterized by a C-terminally located helix-turn-helix (HTH) structural motif consisting of two α-helices joined by a turn, which fits into the major groove of DNA. Regulators bind *via* this motif in the promoter region of target genes and thus exert functions of activators or repressors. In addition to the HTH motif, CRP/FNR regulators have a large nucleotide binding domain [Kolb *et al.*, 1993] that extends from the N-terminus over roughly 170 residues. By itself this domain, defined by similarity to cAMP-binding domains, does not confer specificity onto the regulator family. Rather it places it in a pool of over 1200 nucleotide-

binding proteins of various functions, which encompass all life forms. Thus, by definition, members of the CRP/FNR superfamily are characterized by the presence of both domains.

1.2.2 Structural overview of cAMP receptor proteins: The allosteric activation

The activation process of CRP upon binding to cAMP involves unique features of allostery distinguished from those of other enzymes and membrane receptors. CRP is a homodimeric protein and each monomer consists of two domains (Figure 1.3,1G6N) [Lin *et al.*, 2002; Won *et al.*, 2008]. The N-terminal cAMP-binding domain (residues 1–137) and the C-terminal DNA-binding domain (residues 138–209). The region between C- and D-helices that connects the two domains is called a hinge, typically defined as the L134–D138 region. The larger, N-terminal domain (NTD), which is predominantly β-stranded, is basically responsible for CRP dimerization (see Figure 1.3), and cAMP binding to this domain results in the functional activation of CRP.

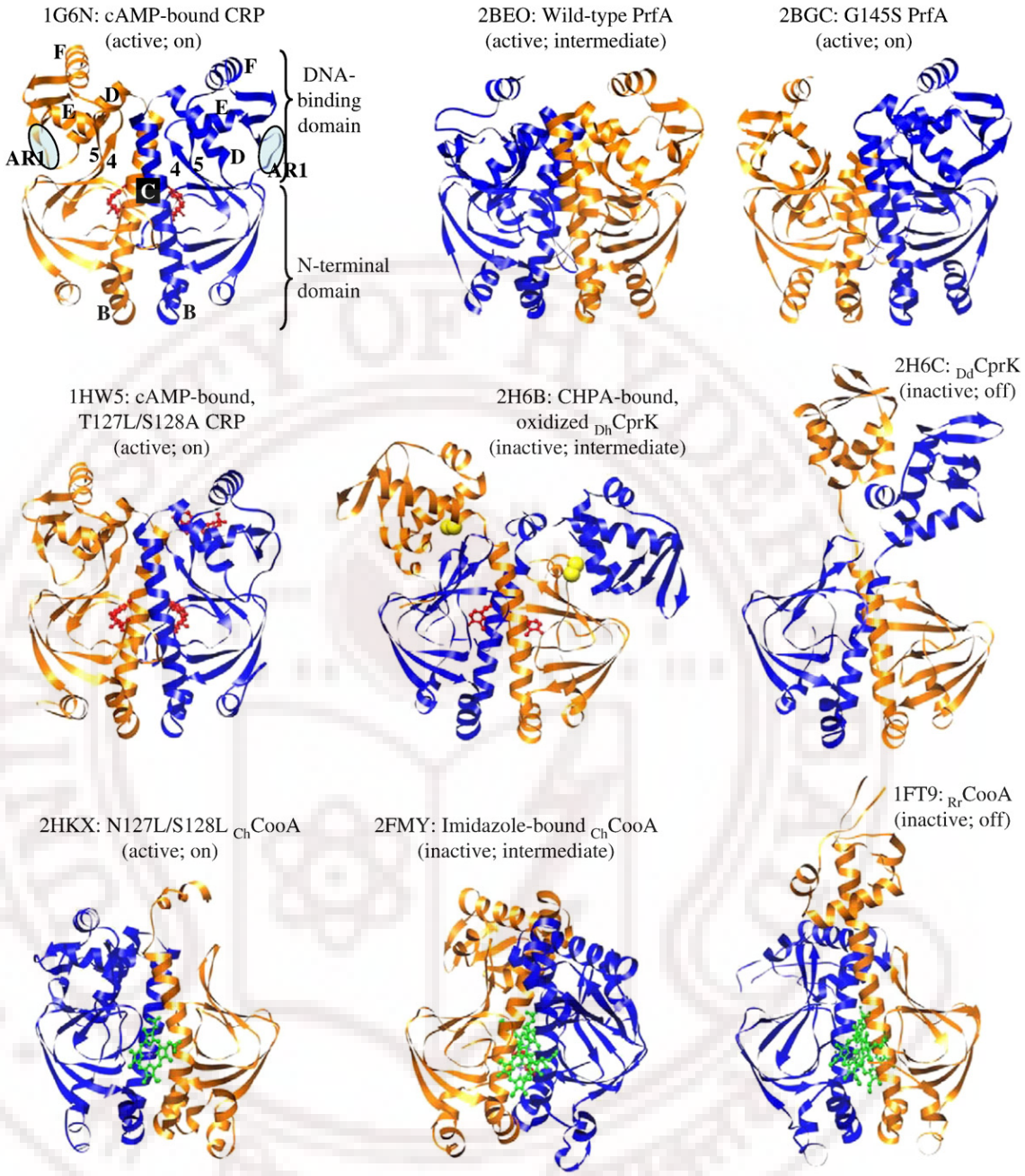


Figure1.3: Ribbon drawings of CRP-family protein structures. PDB code and molecular description is provided on top of each structure. Activity and conformational state of each molecule is indicated in parenthesis. Based on the PDB, A chains and B chains are colored blue and orange, respectively. The bound ligands (cAMP, imidazole, and CHPA) and heme groups are shown as ball-and-stick presentations, colored red and green, respectively. Disulfide bond in CprK is represented in atom-colored spheres enlarged (yellow). Each domain and the secondary structure elements (letters for helices and numbers for strands) addressed in text are denoted on the 1G6N structure. All structures were globally matched to the 1G6N structure, to appear in similar orientations. Figure has been modified from Won *et al.*, 2009.

When activated by cAMP, CRP recognizes specific DNA sites *via* the C-terminal F-helix (αF) which forms a typical helix–turn–helix (HTH) motif, together with the neighboring E-helix (αE) (Figure 1.3, 1G6N). Binding of CRP to DNA bends the DNA by about 87° [Parkinson *et al.*, 1996] and recruits RNA polymerase (RNAP) to initiate transcription of many catabolite genes. This smaller, DNA-binding domain (DBD), which is mainly populated by α-helices, also contains the so-called AR1 (activating region 1; residues 154–164; Figure 1.3, 1G6N) that serves as a main ligand for the interaction with RNAP, at both the CRP-dependent class I and class II promoters [Lawson *et al.*, 2004]. Even in the absence of promoter DNA, CRP can interact with RNAP *via* the AR1 [Heyduk *et al.*, 1993]. The interaction with RNAP at class II promoters is complemented by NTD of CRP, through the AR2 (residues 19, 21, 96, and 101) and the AR3 (residues 52–55, 58). Finally, the cAMP–CRP machinery is now presumed to regulate transcription at nearly 200 different promoters [Hollands *et al.*, 2007]. Since its isolation in the early 1970s, the allosteric conformational change of CRP has been studied in considerable detail, by many biochemical and biophysical probes. In particular, comprehensive understanding of the protein structure by X-ray crystallography has critically contributed to an insight into the CRP allostery. Up to now, more than ten 3-dimensional structure coordinates of CRP (Table 1.1) have been deposited in the Protein Data Bank (PDB), including the three different functional states of complexes (CRP–cAMP

[Weber and Steitz, 1987], CRP–cAMP–DNA [Schultz *et al.*, 1991], and CRP–cAMP–DNA–RNAP [Benoff *et al.*, 2002]) and some mutants (namely CRP*) that are constitutively active even in the absence of cAMP [Vaney *et al.*, 1989]. Unfortunately, the crystal structure of the ligand-free form (apo-CRP) from *E.coli* was not available till recently, NMR data at the level of secondary structure has been published [Won *et al.*, 2009]. Thus, the paucity of 3D-structural information of apo-CRP has highlighted the need for a definite elucidation of the CRP allostery. However, the ligand-free, inactive structures are now alternatively available from the recent crystal structures of other CRP-homologous proteins [Lanzilotta *et al.*, 2000, Gallagher *et al.*, 2009; Kumar *et al.*, 2009] (Table 1.1 and Figure 1.3). Active or intermediate conformations are also additionally available from the CRP-family proteins [Eiting *et al.*, 2005] (Table 1.1 and Figure 1.3). Although each member of CRP family proteins may possess unique process of structural activation, it would be reasonable to anticipate that the homologous proteins certainly share many common features of allostery. Comparative insight into the recent structural achievements on the CRP homologous proteins would provide useful information to fill in the details of the CRP allosterism.

Table 1.1: CRP-related crystal structures deposited in the Protein Data Bank. Modified from Won *et al.*, 2009.

PDB code	Origin	Variation (mutation)	Ligands[a] bound to the protein dimer	Activity	Deposition/release year	Resolution (Å)
CRP (known effector: cAMP)						
1G6N[b]	*E. coli*	–	2 cAMP	Active	2000/2000	2.1
1I5Z	*E. coli*	–	3 cAMP	Active	2001/2003	1.9
2GZW	*E. coli*	–	4 cAMP	Active	2006/2007	2.21
n.a.[c] [Ref. 22]	*E. coli*	A144T	2 cAMP	Active	1987	2.4
n.a. [Ref. 23]	*E. coli*	A144T	1 cAMP, 1 adenosine	Active	1987	2.4
1HW5	*E. coli*	T127L/S128A	3 cAMP	Active	2001/2001	1.82
1I6X	*E. coli*	D53H	3 cAMP	Active	2001/2003	2.2
CRP–DNA complex						
1CGP	*E. coli*	–	2 cAMP	Active	1991/1994	3.0
2CGP	*E. coli*	–	4 cAMP	Active	1997/1998	2.2
1RUN	*E. coli*	–	2 cAMP	Active	1996/1997	2.7
1J59[d]	*E. coli*	–	2 cAMP	Active	2002/2002	2.5
1O3T/1O3Q, 1O3R[e]	*E. coli*	–	2 cAMP	Active	2003/2003	2.8/3.0
1ZRF/1ZRD, 1ZRC	*E. coli*	–	2 cAMP	Active	2005/2006	2.1/2.8
1RUO	*E. coli*	E181F	2 cAMP	Active	1996/1997	2.7
1O3S[e]	*E. coli*	E181D	2 cAMP	Active	2003/2003	3.0
CRP–DNA–RNAPαCTD complex						
1LB2	*E. coli*	–	2 cAMP	Active	2002/2002	3.1
CooA (known effector: CO)						
1FT9	*R. rubrum*	–	–	Inactive	2000/2000	2.6
2FMY	*C. hydrogenoformans*	–	2 Imidazole	Inactive	2006/2007	2.2
2HKX	*C. hydrogenoformans*	N127L/S128L	–	Active	2006/2007	2.3
PrfA (effector unknown yet)						
2BEO	*L. monocytogenes*	–	–	Active	2004/2005	2.7
2BGC	*L. monocytogenes*	G145S	–	Active	2004/2005	2.3
CprK (known effector: CHPA[e])						
2H6C	*D. dehalogenans*	Reduced	–	Inactive	2006/2006	2.9
2H6B	*D. hafniense*	Oxidized	2 CHPA[f]	Inactive	2006/2006	2.2

[a] Nonspecific ligands such as buffer constituent are excluded.
[b] Updated from 1GAP and 3GAP.
[c] Not available (not deposited in PDB).
[d] Updated from 1BER.
[e] Updated from 1DBC, 1DB7, 1DB8, and 1DB9, respectively.
[f] 3-chloro-4hydroxyphenylacetate.

1.2.3 History and current understanding of CRP

The first crystal structure of CRP was solved with two cAMPs bound to the NTDs (1G6N in Figure 1.3) [Weber and Steitz, 1987]. One of controversial points of this structure was the observed asymmetry in relative orientation of domains between subunits. In the 'open' subunit (chain B, orange, in Figure 1.3), there is a large cleft between the DBD and NTD, which is not present in the 'closed' subunit (chain A in Figure 1.3). Similar asymmetry has been also seen in other crystal structures of CRP mutants [Vaney *et al.*, 1989] and CRP-family proteins [Lanzilotta *et al.*, 2000] (Figure 1.3). Although the asymmetry observed in the cAMP–

CRP crystal has been interpreted as an important nature for its functional mechanism, NMR studies in solution could never find any evidence of structural asymmetry in both the apo-CRP and cAMP–CRP [Won *et al.*, 2009]. Thus, the asymmetry in the cAMP–CRP crystal can probably be an artifact due to crystal packing forces, as suggested from the molecular dynamics (MD) simulation results that implied the symmetric structure in solution [Harman, 2001]. Nevertheless, it has been still ambiguous which subunit is more relevant to real conformation in solution. The earlier approach of MD simulation by García and Harman indicated the closed conformation of both subunits [García and Harman, 1996], consistent with that observed in the crystal structures of cAMP–CRP–DNA and cAMP–CRP–DNA–RNAP complexes [Won *et al.*, 2001]. In contrast, recent MD simulation showed open conformation of both subunits in cAMP–CRP [Harman, 2001]. Then, based on electrostatics calculations, the authors have proposed that the open conformation moves to closed conformation upon binding to DNA, resulting in the DNA bending observed in the cAMP–CRP–DNA complex structure. However, the results by Krueger *et al.,* using small angle neutron scattering (SANS) and energy minimization of the crystal structure commonly indicated exclusively closed form for both subunits [Krueger *et al.*, 1998]. In addition, the energy minimization using the starting structure with cAMP removed showed open conformation. Also based on the SANS results, Krueger *et al.,* have finally suggested that the

cAMP binding induces conformational change in CRP from the open form to more compacted closed form. Thus, it is regarded that the closed geometry, rather than the open form, is a more responsible on-state conformation. Based on the cAMP–CRP and cAMP–CRP–DNA structures, it has long been assumed that a CRP dimer can bind two molecules of anti-conformation cAMP, one at each NTD. However, another crystal structure of the cAMP–CRP–DNA complex latterly solved showed two additional cAMPs with a syn-conformation, each of which bound to the DBD and contacted with a part of NTD and DNA [Passner and Steitz, 1997]. The syn-cAMP binding was also evident from the crystal structure of a mutant (T127L/S128A) CRP, where two anti- and one syn-cAMP bound [Tutar, 2009]. Finally, it has been confirmed in solution without DNA that CRP binds four cAMP molecules as its maximum [Won *et al.*, 2009]. These results have created ambiguity about interpreting the binding cooperativity between the four cAMPs and the number of CRP conformers depending on the number of cAMP bound, however the observation of syn-cAMP binding from earlier spectroscopic experiments [Toyama *et al.*, 1991] could be explained. However, only the anti-cAMP binding is assumed to be responsible for the global allosteric transition of CRP, since the overall conformations of the DNA-bound CRP structures were similar, regardless of syn-cAMP binding [Schultz *et al.*, 1991]. Kinetic studies using fluorescence probes have suggested that conformational change precedes the formation of CRP–cAMP4 complex

[Małecki *et al.*, 2000]. It has been finally supported by biochemical and modeling experiments on certain mutants that bind a maximum of two cAMPs [Scott & Jarjous, 2005]. The anti-cAMP bound to NTD is directly coordinated by five intra-subunit residues (G71, E72, R82, S83 and T127) and another one residue (S128) from the opposite subunit [Passner and Steitz, 2000]. Additional indirect interaction with the anti-cAMP is made by an intra-subunit residue R123, which forms a nearby salt bridge with the directly contacting residue E72. Recently it has been revealed that the R123 is important for proper cAMP affinity, but not critical for the conformational transition of CRP [Youn *et al.*, 2007]. Without apo-CRP structure available, the allosteric transition by the anti-cAMP binding has been indirectly probed by biochemical and biophysical investigations. Additionally, isolation of the CRP* phenotype mutants has contributed considerably to deducing the mechanism of conformational change [Passner and Steitz, 2000]. CRP* designates a cAMP-independent mutant of CRP; even in the absence of cAMP it activates transcription. Currently known sites for the CRP* mutation involve positions 53, 62, 127/128 (double mutation), 138, 140, 141, 142, 144, 148 and 195 [Youn *et al.*, 2006]. Up to now, two kinds of constitutively active CRP* A144TCRP [Vaney *et al.*, 1989] and T127L/S128A-CRP [Chu *et al.*, 2001]) structures have been solved by X-ray crystallography (Table 1.1) and their overall conformations were highly comparable to that of cAMP–CRP (Figure 1.3). Unfortunately,

these structures were also crystallized in ligand-bound states; the ligand-free forms of CRP* mutants have not been solved yet. However, it would be reasonable to assume that the CRP* mutants without cAMP already resemble the active conformation of cAMP–CRP, since they are constitutively active regardless of cAMP binding. Based on the SANS measurements, Krueger *et al.,* have revealed that the T127L/S128A-CRP, already in the active conformation, undergoes little, or possibly no, structural change upon binding of cAMP [Krueger *et al.*, 1998]. Current understanding of the allosteric conformational change of CRP can be summarized into rigid-body movements that involve subunit realignment and domain rearrangement, as reviewed by Harman [Harman, 2001]. The main consequence of that global transition is protrusion of the DNA-binding F-helices, formerly buried in apo-CRP.

1.2.4 Structural overview and global transition of other CRP-family proteins

Among the CRP-family members, few species of proteins have been solved in structure (Figure1.3): CRP and CooA from Gram negative bacteria and PrfA and CprK from Gram positive bacteria. The structures can be categorized into three groups. The first group, designated as “off” state, represents the ligand-free, inactive conformations: CooA from *Rhodospirillum rubrum* (RrCooA) [Lanzilotta *et al.*, 2000] and CprK from

Desulfitobacterium dehalogenans (DdCprK) [Joyce *et al.*, 2006]. The last group depicts a fully active, namely "on" state including the cAMP-bound CRP (cAMP–CRP) and constitutively active mutants: T127L/S128A-CRP (CRP*LA) [Chu *et al.*, 2001], G145SPrfA (PrfA*145) [Eiting *et al.*, 2005], and N127L/S128L-ChCooA (ChCooA*LL) [Borjigin *et al.*, 2007]. The other structures are designated as "intermediate" state: the ligand-bound, inactive CooA from *Carboxydothermus hydrogenoformans* (Im-ChCooA) [Komori *et al.*, 2007], the ligand-bound, inactive CprK from *Desulfitobacterium hafniense* (CHPA-DhCprKox) [Joyce *et al.*, 2006], and the ligand-free, active PrfA from *Listeria monocytogenes* [Eiting *et al.*, 2005].

1.2.4.1 PrfA

PrfA is distinguished from CRP in that it has no known effector for activation. Nonetheless, it can positively regulate the expression of many virulence genes in *L. monocytogenes*. Accordingly, the structure already resembles the on-state form of CRP and, what is more, possesses additional three α-helices at the C-terminus that further stabilize DBD (Figure1.3). However, the relatively low DNA-binding affinity of PrfA was attributable to the partial disorder or flexible property in the inter-helix loop region (αE–αF loop) of the HTH motif [Eiting *et al.*, 2005]. In addition, overall geometry of the HTH α-helices significantly deviates from that in the active CRP form. Thus, the authors postulated that PrfA would require an as yet unidentified co-factor for mature activation and

suggested the tunnel structure between domains as a putative cofactor binding site. In these respects, we regarded the wild type PrfA as an intermediate (ligand-free, partially active) state, rather than the on state. The possibility of cofactor requirement is also supported by the fact that the virulence gene expression under the regulation by PrfA is also dependent on environmental factors. In contrast, the G145S mutation of PrfA dramatically enhances the DNA binding affinity, thereby enabling the protein to induce the over expression of virulence genes, regardless of environmental change. It is consistent with CRP, of which A144T mutation also leads to the CRP* phenotype. Thus, the PrfA*145 in Figure 1.3 is regarded as a maturely active form comparable to the ligand-bound, active state of CRP. The main event induced by the mutation in PrfA was the rearrangement of the HTH motif. In particular, the structural ordering of the αE–αF loop (chain A in Figure 1.3) could be obtained by shortening the region, which resulted from the shift of αE to the C-terminus and elongation of αF toward N-terminus.

1.2.4.2 CprK

CprK originates from anaerobic bacteria capable of halo-respiration *via* a reductive dehalogenation of halo-organic compounds. The genes responsible for halo-respiration are transcriptionally regulated by CprK and the chemical compound CHPA (3-chloro-4-hydroxyphenylacetate) has been identified as an effector molecule of the protein. However, CprK is distinguished from CRP in that it is deactivated under aerobic

conditions, by an intra-molecular disulfide bond formation. Thus, the oxidized, ligand-bound structure of CprK from *D. hafniense* (CHPA-DhCprKox), which is inactive, was grouped into the intermediate state, while the reduced, ligand-free form from *D. dehalogenans* (DdCprK) was assigned to the off state. Apparently, CHPA-DhCprKox structure looks similar to the cAMP–CRP structure (Figure 1.3). Asymmetry between subunits was also appreciable, as in cAMP–CRP, with relatively more open and more closed conformations. However, since it was oxidized, the structure exhibits unique features as follows. Compared to CRP, CprK possesses two extra regions: a polypeptide extension at the N-terminus and an additional helix (αG) at the C-terminus (Figure1.3). The C11 residue in the N-terminal tail forms a disulfide bond with the C200 in the DNA-binding αF of the opposite subunit, while the C-terminal αG contacts with NTD of the opposite subunit. These interactions result in the domain-swapped conformation and in consequence the orientation of the HTH motif remains incompatible with DNA binding. In addition, this conformation was suggested to facilitate a putative tetramer formation of the protein, by facing both DBDs in a dimer onto the same parts of the other dimer [Joyce *et al.*, 2006]. In summary, the domain swapped conformation, induced by oxidation, would be responsible for the inertness of the ligand-bound DhCprK at oxidized state. Then, in comparison with this structure, the reduced, ligand-free DdCprK structure could provide important information about the ligand-induced

allosteric transition. The most distinct feature of the ligand-free form was a dimerization of the C-terminal DBDs, whereby the DBDs are positioned apart from NTDs and the HTH motifs are constrained incompatible with DNA binding (Figure 1.3). Consequently, it could be postulated that the ligand binding to CprK moves the DBDs toward the NTDs, which disrupts the DBD dimer interface and leads to repositioning of the HTH motif at a required position fit to DNA binding. However, oxidation of CprK would inhibit the activation by driving the domains to a false direction; thereby anchoring the DBDs onto the NTDs of opposite subunits.

1.2.4.3 CooA

The functional ligand of CooA is CO and the protein activates transcription of genes involved in CO oxidation to get energy under anaerobic environment. CooA is distinguished from CRP in that its NTD contains two additional parts: a heme molecule as a prosthetic group to sense CO and an N-terminal extension of which a part provides an internal ligand to the heme. Then, CO binds to the heme by replacing the N-terminal ligand, thereby triggering the allosteric activation of the protein. The first structure of CooA from *R. rubrum* (RrCooA) has been solved in the absence of CO, thus as an inactive form (off state), where the intra-subunit residue H77 acts as one heme ligand to bind the heme while the residue P2 from the opposite subunit provides the other ligand to be displaced by CO [Lanzilotta *et al.*, 2000].

Displacing of the N-terminal heme ligand could be observed when imidazole, instead of CO, bound to the CooA from *C. hydrogenoformans* (ChCooA) [Komori *et al.*, 2007]. However in Figure 1.3, the Im-ChCooA structure is regarded as an intermediate state, as it was still inactive. In contrast, the N127L/ S128L mutant of ChCooA (ChCooA*LL) is constitutively active even without CO, representing an on-state conformation [Borjigin *et al.*, 2007]. The ligand-free RrCooA structure showed an asymmetry, with one subunit closed and the other open, as observed in the cAMP–CRP. Although the observed asymmetry in RrCooA was far prominent, neither subunit adopted a proper domain orientation for DNA binding. Among the two subunits, the open subunit (chain B in Figure 1.3), where DBD is stretched straightforward from NTD, has been interpreted as a real off state [Lanzilotta *et al.*, 2000]. However, the heterogeneity between subunits, although it occurred probably owing to crystal packing forces [Chan, 2000], has been suggested alternatively to reflect flexibility in the hinge and DBD regions [Joyce *et al.*, 2006]. In this respect, Komori *et al.,* (2007) have argued that the closed subunit in the RrCooA structure, where DBD is positioned closer in proximity with NTD, could represent the physiologically relevant conformation of the off state. However, the off-state structure of the DdCprK is more consistent with the open subunit of RrCooA (Figure 1.3). No matter which is correct, DBD becomes further shifted toward NTD in the Im-ChCooA and ChCooA*LL. This observation supports the hypothesis [Krueger *et al.*,

1998] that the conformational change of CRP would be driven from a relatively open to closed conformation. The on-state ChCooA*LL assumes the domain orientation well matched to that of the active CRP (Figure 1.3). In this structure, despite the absence of bound CO, the N-terminal tail was fully expelled from the heme pocket and situated between the DNA- and heme-binding domains, thus stabilizing the favored orientation of domains [Borjigin *et al.*, 2007]. Actually, one subunit (chain B in Figure 1.3) in the ChCooA*LL structure did not exist at on state, since it was unexpectedly devoid of the heme. In this heme-free subunit, the N-terminal tail is oriented away from the main body, not interacting with DBD. Accordingly, the DBD in the heme-free subunit was not visible in electron-density maps, probably due to disorder or flexibility. Thus, the N-terminal tail of CooA seems to be important for both the off-state and on-state conformations. Functional importance of the N-terminus has also been evidenced by mutational experiments [Borjigin *et al.*, 2007]. The releasing of The N-terminal heme ligand was also observed in the Im-ChCooA structure. However, in this structure, the N-terminal region is still restricted in the vicinity of the heme, by hydrogen bonding to the imidazole. Although DBD was close to NTD, it remained to be rotated around the hinge region, for proper DNA-binding orientation [Komori *et al.*, 2007]. Consequently, the Im-ChCooA could not be active.

1.2.5 cAMP receptor protein from *Mycobacterium tuberculosis* : A potential drug target

Mtb contains a single CRP/FNR homologue coded by ORF Rv3676 (Cole *et al.*, 1998). Orthologue of Rv3676 (*Mtb*-CRP) is present in all sequenced and unfinished mycobacterial genomes. The DNA binding and cAMP binding properties of recombinant Rv3676 were described based on computational predictions (Bai *et al.*, 2005). This study was further extended to provide insights into unusual and novel biochemical properties of Rv3676 protein including crystallization and preliminary X-ray diffraction data. Previously, Mattow *et al.,* (2001), while comparing the proteome profiles of *Mtb* and *M. bovis* BCG, observed differences in electrophoretic mobility of CRP proteins (Mattow *et al.*, 2001). Subsequently, such a mobility shift was attributed to point mutations in both the DNA and cAMP binding domains in CRP of *M. bovis* BCG. These mutations were ascribed to the impaired DNA binding activity of *M. bovis* BCG-CRP in comparison to *Mtb*-CRP (Spreadbury *et al.*, 2005). This might suggest these mutations to be one of the contributing factors in attenuation of the virulence of *M. bovis* BCG. In other studies, *Mtb* deletion mutants corresponding to Rv3676 (*Mtb*-CRP) revealed growth defects in laboratory cultures of bone marrow derived macrophages and in mouse models of tuberculosis (Rickman *et al.*, 2005).

1.2.6 Structural studies on *Mtb*-CRP

Recently, Gallagher and colleagues reported the structure of a cAMP free form of CRP (Gallagher *et al.*, 2009). They found that the dimer was asymmetric in both domains, but the N-domains differ only by isolated rotamers and smooth global deformations, generally preserving the local environment of each residue. The differences in the C-domains were more profound, involving many residues that have completely different local environments, including different H-bonds. The RMSD between the two C-domains (70 Ca positions) was 3.1 Å. This report however was contradicted in the later published structure of cAMP free form of *Mtb*-CRP (Kumar *et al.*, 2009). Unlike the structure of unliganded apo *Mtb*-CRP reported by Gallagher and colleagues, the two DNA-binding domains in this structure match well with each other with an RMSD of 1.1 Å (over 54 Cα positions). The only difference in the conformation of the DNA-binding domains is the relative orientation of the D-helix with respect to the rest of the domain. It is therefore likely that the conformational difference in the DNA-binding domains of the two subunits, as observed in the earlier reported *Mtb*-CRP structure (Gallagher *et al.*, 2009), is due to crystal packing effects rather than inherent conformational heterogeneity (Kumar *et al.*, 2009). Very recently, *Mtb*-CRP-cAMP structure was also reported (Reddy *et al.*, 2009). They presented two crystallographic structures: that of *Mtb*-CRP bound to cAMP and that of *Mtb*-CRP bound to an N6-cAMP ligand. Determination of both the CRP-

cAMP and CRP -N6-cAMP structures has enabled them to explore various allosteric models, giving more detailed insight into the allosteric behavior of this transcriptional regulator. Their structures of the cAMP and N6-cAMP bound CRP allow for the first time a comparison of apo CRP to cAMP and inhibitor-bound CRP within the same species. Crystallographic analysis of these structures has lead to a model in which the *Mtb* CRP exists as a symmetric dimer, in which both subunits exist in an open form. Furthermore, their structures implicate a model for the allosteric switch from the inactive apo form to the active cAMP-bound form. This model implies that binding of cAMP triggers alterations in the cAMP contacting residues and shifts the N-terminal domain, consequently allowing the DNA binding domain to accept DNA. The structural basis of the CRP inhibitor is based upon abrogation of dimer interactions and a concomitant helix unwinding at the interface of one of the subunits.

1.3 Objectives of current study

Mtb infects human macrophages and dendritic cells where it has to survive within unfavorable antibacterial environment presented by host apparatus. Eubacteria have evolved in a way to respond and cope with extremely diverse harsh environments. In case of intracellular pathogens it has to deal with hypoxia, starvation and acidic pH etc. In order to

respond quickly to these environmental changes bacteria have to switch on certain alternative metabolic pathways as well as to shut down some of them. It always involves some of the transcription regulators which regulate the expression of related proteins. CRP/FNR family of proteins is closely related to regulation of alternative gene programs during stress in bacteria.

Given the fact that the CRP/FNR family transcription regulators are important for survival of intracellular micro organisms, the present work deals with the structure-function studies on *Mtb*-CRP. The overall objectives of this thesis are:

1. cAMP Receptor Protein Regulons in mycobacterial species including *Mtb*, *M. bovis*, *M. leprae*, *M. smegmatis*, and *M. smegmatis*;
2. Biophysical and biochemical characterization of *Mtb*-CRP;
3. Structural studies on *Mtb*-CRP-cAMP-DNA ternary complex using X-ray crystallography.

Chapter 2: cAMP Receptor Protein Regulons in mycobacteria

A part of the work presented in this chapter has been published as:

ELSEVIER

Available online at www.sciencedirect.com
ScienceDirect
Gene 407 (2008) 148–158

GENE
www.elsevier.com/locate/gene

Genome scale portrait of cAMP-receptor protein (CRP) regulons in mycobacteria points to their role in pathogenesis

Yusuf Akhter [a,e,1], Sailu Yellaboina [b], Aisha Farhana [a], Akash Ranjan [b], Niyaz Ahmed [c], Seyed E. Hasnain [d,e,*]

[a] *Laboratory of Molecular and Cellular Biology, CDFD, Hyderabad, 500076, India*
[b] *Laboratory of Computational and Functional Genomics and CDFD-Sun Microsystems Centre of Excellence in Medical Bioinformatics, CDFD, Hyderabad, 500076, India*
[c] *Pathogen Evolution Laboratory, CDFD, Hyderabad, 500076, India*
[d] *Jawaharlal Nehru Centre for Advanced Scientific Research, Jakkur, Bangalore, 560012, India*
[e] *University of Hyderabad, Hyderabad, 500046, India*

Received 23 March 2007; received in revised form 2 October 2007; accepted 5 October 2007
Available online 22 October 2007
Received by G. Pesole

2.1. Introduction

Cyclic AMP receptor proteins (CRP)–FNR superfamily of transcription factors regulate a diversity of physiological processes in bacteria. These proteins predominantly regulate intracellular reactions related to carbon, sulfur and nitrogen metabolism, denitrification, nitrogen fixation, aerobic and anaerobic respiration and expression of virulence genes in response to a variety of environmental and metabolic signals (Körner *et al.*, 2003; Green *et al.*, 2001). With more than 370 family members, these DNA-binding proteins primarily function as positive regulators except that a few members of this family also act as negative regulator of transcription. FNR protein from *E. coli* acts both as activator as well as repressor of transcription depending upon the distance between the binding site and the transcription start-points (Barnard *et al.*, 2003). Some of the distinguishing features of CRP/FNR members include the presence of a nucleotide binding domain and a helix turn helix DNA binding motif at the N-and C- terminal, respectively. The classical cAMP-binding domain (Schultz *et al.*, 1991) is a versatile structure that has evolved to accommodate different functional specificities in response to a wide range of signals (Körner *et al.*, 2003; Green *et al.*, 2001). The best-studied prototype proteins from this

superfamily are CRP and FNR of *E. coli* that regulate expression of numerous genes in response to starvation and hypoxia, respectively.

In case of *E. coli* CRP, cAMP acts as a secondary signal and binds to the CRP protein to form cAMP-CRP complex. This complex then binds to the promoters carrying specific DNA elements related to the consensus motif TGTGANNNNNNTCACA (Berg and von Happel, 1988) thereby regulating expression of downstream genes. However, in case of FNR protein, redox states of bound metals act as the signal and activate the FNR that further binds to the promoter containing specific DNA elements (consensus TTGATNNNNATCAA) and regulate gene expression (Spiro, 1994).

Mycobacterium tuberculosis (*Mtb*) H37Rv contains a single CRP/FNR homologue coded by ORF *Rv3676* (Cole *et al.*, 1998). Orthologue of *Rv3676* (*Mtb*-CRP) is present in all sequenced and unfinished mycobacterial genomes. Recently, the DNA binding and cAMP binding properties of recombinant Rv3676 were described based on computational predictions (Bai *et al.*, 2005). Previously, Mattow *et al.* (2001), while comparing the proteome profiles of *Mtb* and *M. bovis* BCG, observed differences in electrophoretic mobility of CRP proteins (Mattow *et al.*, 2001). Subsequently, such a mobility shift was attributed to point mutations in both the DNA and cAMP binding domains in CRP of *M. bovis* BCG. These mutations were ascribed to the impaired DNA binding

activity of *M. bovis* BCG-CRP in comparison to *Mtb*-CRP (Spreadbury *et al.*, 2005) thereby providing to their importance as contributing factors in attenuation of virulence of *M. bovis* BCG. In other studies, *Mtb* deletion mutants corresponding to *Rv3676* (*Mtb*-CRP) revealed growth defects in laboratory cultures of bone marrow derived macrophages and in mouse models of tuberculosis (Rickman *et al.*, 2005).

Apart from some initial studies, not much is known about the population-wide repertoire of CRP-regulated genes *per-se* in different clinical settings and their cellular functions. This chapter describes CRP regulated novel genes of *Mtb* and prediction of their abundance and operon context in the genomes of *M. leprae, M. avium* subsp. *paratuberculosis* and *M. smegmatis.* This part of work also attempts to identify common genes across the genus mycobacteria, which could be regulated by CRP.

2.2 Materials and Methods

2.2.1. Source of genome sequence

Published and annotated genome sequences of *Mtb, 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/tigrscripts/CMR/CmrHomePage.cgi). Sequences corresponding to *Mtb*-CRP-binding sites identified previously (Bai *et al.*, 2005; Rickman *et al.*, 2005) were also obtained. A local archive of genome sequences was assembled, curated and stacked. This resource was queried for known and putative homologies and to build consensus and alignments at a genome wide interface.

2.2.2 Prediction of CRP-binding sites

CRP-binding site recognition profile was calculated by positional Shannon relative entropy method as described earlier (Yellaboina *et al.*, 2004; Prakash *et al.*, 2005, Yellaboina *et al.*, 2006). Sequences corresponding to *Mtb*-CRP-binding sites identified previously (Bai *et al.*, 2005; Rickman *et al.*, 2005) were used to generate input profile. The binding site profile thus generated was used to scan upstream sequences of all the genes of each of the mycobacterial genomes. The score of each site was calculated as the sum of the respective positional Shannon relative entropy of each of the four possible bases. A maximally scoring site was selected from the upstream sequence of each of the gene we looked at. The lowest score among the input binding sites was considered as cut-off score. The sites scoring higher than the cut-off value were predicted as potential binding sites that conform to the consensus sequence. For each of predicted regulon (*Mtb, M. avium, M. leprae* and *M.*

smegmatis) sequence logo was generated by using Weblogo server (Crooks *et al.*, 2004). The height of each stack of letters represented the degree of sequence conservation measured in bits. The height of each letter within a stack is proportional to its frequency at that position in the binding site. The letters were sorted with the most frequent on top (Figure 2.2).

2.2.3 Prediction of operons and function annotation

Genes transcribed co-directionally and downstream to the predicted binding sites in *M. leprae, M. avium* subsp. *paratuberculosis* and *M. smegmatis* were chosen as potentially co-regulated genes (operons). These were used according to one or more of the following criteria (Yellaboina *et al.*, 2004; Tundup *et al.*, 2006; Yellaboina *et al.*, 2006): [1] co-directionally transcribed orthologous gene pairs, conserved in at least 4 genomes; [2] genes belonging to the same cluster of orthologous gene function category whose intergenic distance is less than 200 base pairs; [3] if the first three letters in the gene names are identical (gene names for putative genes were assigned from COG database); and [4] if the intergenic distance is less than 90 base pairs.

The functions of regulated genes were predicted if not done so previously. RPS-BLAST search against conserved domain database (CD search) at NCBI server (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi)

was used to infer the annotation. In principle, RPS-BLAST is a tool opposite of PSI-BLAST which searches the profiles against a database of sequences, hence the 'Reverse'. RPS-BLAST uses a BLAST-like algorithm, finding single- or double-word hits and then performing an ungapped extension on these candidate matches. If a sufficiently high-scoring ungapped alignment is produced, a gapped extension is performed and those (gapped) alignments with sufficiently low expect value have been reported (Marchler-Bauer *et al.*, 2002).

2.3 Results and Discussion

2.3.1 CRP regulators from mycobacteria have conserved DNA binding domains

It was aimed at exploiting the use of CRP binding sites in *Mtb* for the prediction of regulons in mycobacteria. DNA binding helix-turn-helix (HTH) motif was predicted at EMBOSS server (http://bioweb.pasteur.fr/seqanal/interfaces/helixturnhelix-imple.html).

The sequence (176-LTQEEIAQLVGASRETVNKALA-196) from *Mtb*-CRP was identified as the likely HTH motif involved in DNA binding as it elicited maximum score.

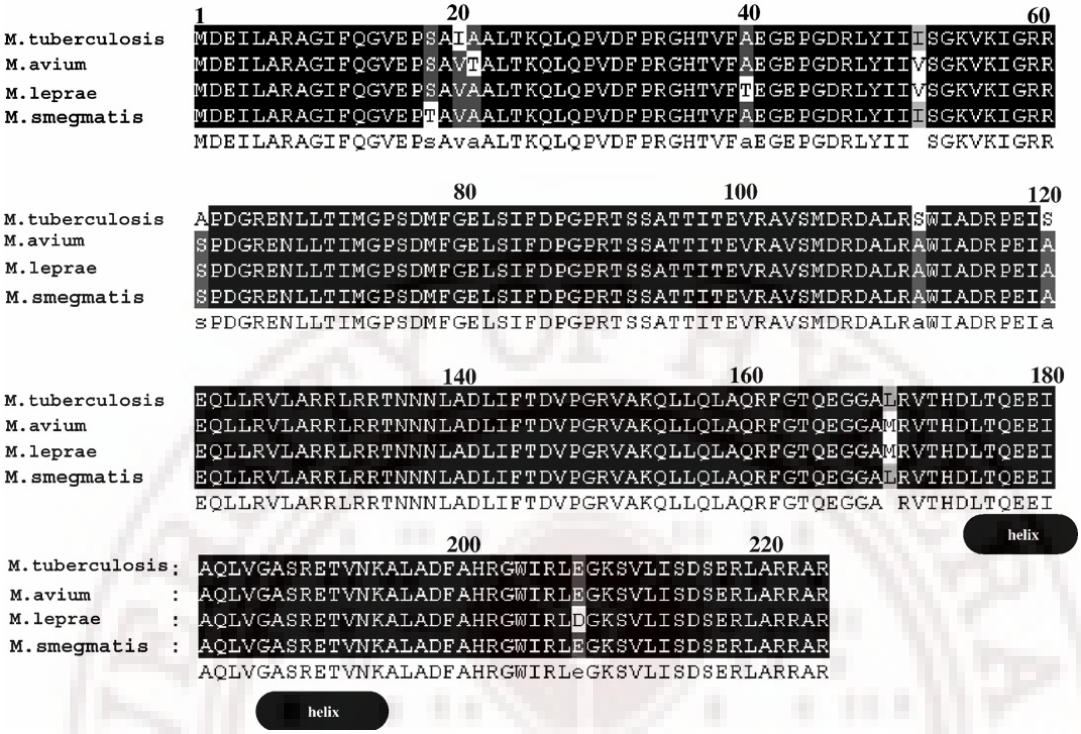


Figure 2.1: Alignment of CRP orthologues from different species of mycobacteria reveals a highly conserved DNA binding domain. *Mtb*-CRP sequence was used as reference sequence and compared with other orthologues. Arial black shadow shows identity and the gray show similarity. Two helices (labeled as helix) are part of helix turn helix that assists in CRP box recognition.

This HTH motif from *Mtb*-CRP was selected and compared with the orthologs. The bioinformatics analysis of CRP regulators from different mycobacterial species was validated through CLUSTAL W based alignments, which revealed high-level identity in their DNA binding domains (Figure 2.1). In the comparative amino acid sequence analyses with *Mtb*-CRP, it was found that CRP orthologues from *M. avium, M. leprae* and *M. smegmatis* were 96%, 96% and 97% identical respectively.

2.3.2 Novel CRP binding sites in *Mtb* genome

In earlier computational predictions of CRP-regulon, 73 binding elements in *Mtb* genome were observed (Bai *et al.*, 2005). Of the top 44 binding sites identified in the present work, 25 were reported previously (Bai *et al.*, 2005). Thus, 19 new *Mtb*-CRP binding sites upstream of various operons in *Mtb* genome showed up (Table 2.1). The CRP-binding element from fumarate reductase (*Rv1552*) received the highest score (Table 2.1). While this DNA element was also reported as a potential CRP-binding site in earlier reports (Bai *et al.*, 2005; Rickman *et al.*, 2005, Spreadbury *et al.*, 2005), experimental evidence for specific DNA: protein interaction between this DNA element and recombinant purified *Mtb*-CRP has been provided in chapter 3 (Akhter *et al.*, 2007). Also, Spreadbury *et al.* (2005) proposed some potential genes as members of the CRP-regulon. There are obvious overlaps between the results from previous studies and the present one, but the present study could identify some novel members of the CRP-regulon in *Mtb* genome. While previous studies (Bai *et al.*, 2005, Spreadbury *et al.*, 2005) utilized information from *E. coli* CRP-regulon, in this study only the available information from *Mtb*-CRP-regulon has been used.

Table 2.1: Predicted CRP binding sites in *M. tuberculosis* genome

Score	Position	Binding site	Gene	Synonym	Product
3.96721	−287	AATGTGATCTAGGTCACGTG	frdA	Rv1552	Fumarate reductase
3.92739	−176	ATTGTGAGTTGGATCACGTT	sucC	Rv0951	Succinyl-CoA synthetase subunit beta
3.92564	−156	GCCGTGAGATTCGTCACGTC	–	Rv1810	Hypothetical protein
3.92181	−11	CGTGTGAACGATGTCACGCC	galU	Rv0993	UTP-glucose-1-phosphateU-transferase
3.90978	−4	ACTGTGACGCCCGTCACAAC	–	Rv0104	cAMP binding protein
3.90294	−182	GCTGTGAGCCGAATCACGAC	–	Rv3645	Membrane linked adenylate cyclase
3.90045	−378	ATCGTGAAGCCGTTCACGCT	glnD	Rv2918c	PII uridylyl-transferase
3.89854	−160	GTTGTGACGGGCGTCACAGT	ctpB	Rv0103c	Cation-transporter p-type ATPaseB
3.89686	−41	CGCGTGACATGTGTCACATG	PE_PGRS44	Rv2591	PE-PGRS
3.89517	−141	AGTGTGATTTACATCACATA	–	Rv2700	Secreted alanine rich protein
3.89328	−44	GTCGTGATACGACTCACGCG	echA6	Rv0905	Enoyl-CoA hydratase
3.89246	−169	ACCGTGACGCCCCTCACGGC	fadD21	Rv1185c	Acyl-CoA synthase
3.89159	−136	AGTGTGAACAAGCTCACATG	–	Rv0885	Hypothetical protein
3.88671	−92	CATGTGAGCTTGTTCACACT	serC	Rv0884c	Phosphoserine aminotransferase
3.88342	−110	GGCGTGACATCGTTCACACG	–	Rv0992c	5-formyltetrahydrofolatecyclo-ligase
3.88095	−120	AACGTCACATGCGTCACGGC	–	Rv1581c	Probable phiRv1 phage protein
3.87653	−154	AACGTGATCCAACTCACAAT	–	Rv0950c	Hypothetical protein
3.87318	−117	TATGTGATGTAAATCACACT	–	Rv2699c	Hypothetical protein
3.86965	−3	CGCGTGAGTCGTATCACGAC	accD3	Rv0904c	Acetyl-CoAcarboxylaseCo-transferase
3.86178	−90	GCTGTCAAATCCGTCACGAA	–	Rv2336	Hypothetical protein
3.86153	−80	GATGTGACTCAAGTGACACG	–	Rv1159	Conseved transmembrane protein
3.85583	−306	TGCGTGAGGAGCCTCACGGC	–	Rv2650c	PhiRv2 prophage protein
3.85125	−70	CGTGTCACTTGAGTCACATC	–	Rv1158c	Hypothetical ALA,PRO-rich protein
3.84869	−70	ATCGTGACTTTGCTGACGTG	–	Rv0019c	Hypothetical protein
3.84804	—216	GCGGTGATCGGCGTCACGCC	PE24	Rv2408	PE
3.84665	−177	GACGTCAACCAGTTCACGCT	mmaA3	Rv0643c	Methoxy mycolic acid synthase
3.83995	−80	ATGGTGATCTAGTTCACGAA	–	Rv1230c	Membrane protein
3.83989	−236	CGCGTGACTGAAATCACAAC	–	Rv1566c	Possible inv protein
3.83976	−229	CGTGTGACAGCTGTGACGGT	wag22	Rv1759c	PE-PGRS
3.82802	−367	ACCGTCACAGCTGTCACACG	–	Rv1760	Hypothetical protein
3.82654	−86	GATGTGATGCACTTGACATC	PE33	Rv3650	PE
3.82347	−245	GTGGTGAGCTGGTTCACACC	–	Rv2407	Ribonuclease Z
3.81909	−35	GGTGTGAACCAGCTCACCAC	–	Rv2406c	Hypothetical protein
3.81879	−341	TTCGTGAGGCGTGTGACGAA	–	Rv3113	Phosphatase
3.81542	−7	GACGTCATGGATTTCACGAC	fadE27	Rv3505	Acyl-CoA dehydrogenase
3.81481	−50	ACCGTGACATCGATGACAGC	–	Rv3031	Glycoside hydrolase
3.81167	−168	ACGGTGACAGCGCTCACGGT	moaX	Rv3323c	MOAD-MOAE fusion protein
3.81104	−272	TAGGTGACCAAACTCACGCT	PPE11	Rv0453	PPE
3.80557	−318	CCTGTGACCGGTGTCACTGC	ephA	Rv3617	Probable epoxide hydrolase
3.80304	−136	CGTGTGACCAAACTCACCGC	PE15	Rv1386	PE
3.79969	−175	ATCGTGACACCGGTAACGGC	–	Rv0520	Methyltransferase/methylase
3.79829	−371	GATGTGACCGTGGTAACGTA	pdhC	Rv2495c	Dihydrolipoamide acetyltransferase
3.79647	−69	AGTGTAACGCATATCACGTG	–	Rv0452	Transcriptional regulatory protein
3.7954	−288	ATAGTGACGGCCGTCACAGC	–	Rv3690	Conserved membrane protein

***New identified sites**

Table 2.2: Predicted CRP binding sites in *M. avium subsp. paratuberculosis* genome

Score	Position	Binding site	Gene	Synonym	Product
4.04838	−353	AGTGTGACCTCCGTCACATC	–	MAP3737	Hypothetical protein
4.02131	−137	AGTGTGATCTAGGTGACGTG	–	MAP3267c	Hypothetical protein
4.00972	−82	GGTGTGATTTACCTCACACC	–	MAP2817	Hypothetical protein
4.00494	−116	GGTGTGAGGTAAATCACACC	–	MAP2816c	Hypothetical protein
3.99554	−383	GACGTGACGAGGTTCACGGC	fadD29	MAP3284c	FadD29
3.99383	−56	ACTGTGAGATTGCTCACAGT	–	MAP3671	Hypothetical protein

3.98389	−45	GCCGTGATACGACTCACGAG	echA6	MAP0840	Enoyl-CoA hydratase
3.98388	−62	GTTGTGAGCCCGCTCACAGG	–	MAP0174	Hypothetical protein
3.98224	−37	GCCGTGATTCAGTTCACACC	–	MAP0227c	Hypothetical protein
3.9796	−25	CGTGTGACCAACCTCACATT	–	MAP2149c	Hypothetical protein
3.97915	−111	TTTGTGAGCCGCTTCACACC	–	MAP2220	Ribonuclease Z
3.97805	−101	AACGTCACAAACGTCACGGT	lpqR	MAP0670	LpqR
3.97759	−373	GGCGTCACCGAGGTCACGCT	–	MAP0727	Hypothetical protein
3.97627	−3	CTCGTGAGTCGTATCACGGC	accD3	MAP0839c	AccD3
3.96913	−37	GGTGTGAAGCGGCTCACAAA	–	MAP2219c	Hypothetical protein
3.96489	−41	AACGTGAACGCCGTGACGGC	pepC	MAP0632	Putative aminopeptidase 2
3.96428	−69	GACGTCAACCAGTTCACGCT	recC	MAP4094c	RecC
3.96198	−293	TTTGTGATTGATCTCACGGA	–	MAP1418c	Hypothetical protein
3.96178	−399	AACGTCACCCAACTCACGAG	–	MAP1601	Hypothetical protein
3.95747	−40	AACGTGACGGGTGTGACGGA	pknD	MAP3387c	PknD
3.95283	−187	TCCGTCAACCGCGTCACGTC	echA21	MAP0249c	Enoyl-CoA hydratase
3.94986	−4	ATCGTGATAGCGGTGACGAT	–	MAP2875	Hypothetical protein
3.9487	−87	CGGGTGACCCCGGTCACGCT	ephD	MAP1955c	Short chain dehydrogenase
3.94584	−157	AGCGTGACCGGGGTCACCCG	dlaT	MAP1956	Dihydrolipoamide acyltransferase
3.94403	−127	GGTGTCAGCATGGTCACATA	–	MAP2018c	Hypothetical protein
3.94196	−101	GACGTGACCTCGATGACACG	–	MAP0097c	Hypothetical protein
3.93609	−119	ATTGTGATTTGGATCACCTA	sucC	MAP0896	Succinyl-CoA synthetase subunit beta
3.93595	−35	CACGTCACCTAGATCACACT	hsp18_3	MAP3268	hsp18_3
3.93244	−62	ATCGTCACCGCTATCACGAT	fadD13	MAP2874c	FadD13
3.93233	−1	ATGGTGAACAAATTCACGAC	lpqL_1	MAP3906	LpqL_1
3.92976	−111	AACGTGATTGGCATGACGAG	–	MAP1597	Hypothetical protein
3.92919	−96	CGGGTGAAGCGGGTCACGAC	–	MAP0683	Hypothetical protein
3.92773	−32	ACCGTCATGGAGATCACGGG	fadA3	MAP2407c	Acetyl-CoA acetyltransferase
3.92509	−163	GACGTCACGGCTTTCACGGC	–	MAP1333	Hypothetical protein
3.92501	−285	GCGGTGATCTACGTGACGCC	–	MAP1644	Hypothetical protein
3.92295	−78	CGCGTGAGTAGGGTGACATC	–	MAP2860	Hypothetical protein
3.92295	−392	CGCGTGAGTAGGGTGACATC	–	MAP2861	Short chain dehydrogenase
3.92038	−3	GCCGTGATGGACCTGACGCA	–	MAP2928c	Short chain dehydrogenase
3.91804	−248	TCGGTGACCCGTTTCACGCC	–	MAP1475	Hypothetical protein
3.9169	−65	CGGGTGACCCGGCTCACGGT	valS	MAP2271c	valyl-tRNA synthetase
3.90643	−244	ATTGTGACCGGCCTCACTGC	–	MAP0948	Hypothetical protein
3.90494	−269	ACTGTTAGTTACATCACACC	–	MAP1693c	Hypothetical protein
3.90214	−89	TGGGTGACCCTTGTCACAGC	–	MAP0725	Hypothetical protein
3.90027	−318	GGCGTGAACAACCTCACCGG	–	MAP2531	Hypothetical protein
3.89953	−189	ACCGTTATCGTTGTCACGCA	–	MAP2500	Hypothetical protein
3.89821	−366	ATCGTAAAGCCGTTCACGCT	glnD	MAP2986c	PII uridylyl-transferase
3.89757	−325	AGCGTGACCTCGCTTACACC	–	MAP1558c	Hypothetical protein
3.89655	−93	ACTGTGAATTAGTTAACAAG	fadE24	MAP3188	FadE24
3.89528	−104	AAGGTCAAGACCGTCACGTC	fadE9	MAP4214c	FadE9
3.89286	−56	TACGTCACCGGGGTGACGTT	–	MAP2780	Hypothetical protein
3.89245	−294	TGCGTGATGCCCTTGACGAA	fadD2	MAP3714	acyl-CoA synthase
3.89228	−43	AGTGTGAGGTGTATTACACA	–	MAP3952	Hypothetical protein
3.89158	−96	GGTGTGACGAGTTTCACTAC	fbpC1	MAP0217	FbpC1
3.8888	−365	TTTGTGACTCACCTCACTTG	sodA	MAP0187c	SodA
3.8888	−25	TTTGTGACTCACCTCACTTG	–	MAP0188c	Hypothetical protein
3.88826	−369	GCGGTGATCTGGCTGACGTG	embR_1	MAP0230	EmbR_1
3.888	−3	GAGGTGACGCAATTGACGCC	atsG	MAP3791c	AtsG
3.88752	−117	ATTGTTAGCGCGGTCACAGA	–	MAP2060c	Hypothetical protein
3.88613	−208	GGCGTCACCCTGCTGACGGT	–	MAP0053c	Hypothetical protein
3.8858	−8	TATGTGATTTGTATAACGCA	–	MAP3944c	Hypothetical protein
3.88484	−119	GATGTCAGGGTGGTGACATG	parB	MAP4344c	ParB
3.88483	−19	GCAGTGAGGCCGGTCACAAT	–	MAP0947c	Hypothetical protein
3.884	−397	CAAGTGAGGTGAGTCACAAA	–	MAP0189	Hypothetical protein

Table 2.3: Predicted CRP binding sites in *M. leprae* genome

Score	Position	Binding site	Gene	Synonym	Product
3.71657	−390	ACTGTGAACCAAGTCACTAC	–	ML0201	Hypothetical protein
3.70052	−139	CGTGTGACTGATGTGACACG	–	ML0185	Hypothetical protein
3.69878	−11	CGTGTGAACGATGTCACTCC	galU	ML0182	UTP-glucose-1-phosphate uridylyltransferase
3.68332	−168	ATTGTGATTTGTATCACTGT	sucC	ML0155	Succinyl-CoA synthetase subunit beta
3.68277	−362	GTTGTGACCCATCTCACTGT	sodA	ML0072	Superoxide dismutase
3.67869	−240	TGCGTGATCTGCTTGACGAT	–	ML0298	Sulfur carrier protein ThiS
3.67586	−181	CTGGTGATAGCCCTCACGCA	–	ML0141	Hypothetical protein
3.67469	−87	GGAGTGACATCGTTCACACG	–	ML0181	Hypothetical protein
3.65712	−149	ACAGTGATACAAATCACAAT	–	ML0154	Hypothetical protein
3.62614	−351	ACCGTGACTAGGGTGACCAA	–	ML0333	Hypothetical protein
3.62494	−261	CGGGTGATAAGAGTCACCAG	–	ML0410	Putative PE-family protein
3.59242	−124	GACGTGAGGGCCATTACGCA	–	ML0229	Hypothetical protein
3.59072	−174	GGCGTGATTTCCCTTACATC	–	ML0240	Hypothetical protein
3.57863	−201	GCTGTGGCTAGTGTCACGTC	rpmF	ML0173	50S ribosomal protein L32
3.52958	−4	GTTGTGAGCAAGTTTACCGA	–	ML0243	acyl-CoA synthase
3.52512	−122	CCAGTGACCGAAGTGACCGA	–	ML0336	Putative ABC-tranporter ATP-binding protein
3.52352	−249	ATTGTCAGAGGCTTTACACG	–	ML0107	Hypothetical protein
3.50566	−338	CGGGTCAACGGGATCACCGA	–	ML0279	Hypothetical protein
3.48974	−189	GCTGTTACCCTAGTGACCCT	pabA	ML0015	Para-aminobenzoate synthase component II
3.48673	−240	GTCGTCGGTTGGGTCACGGT	purN	ML0160	Phosphoribosylglycinamide formyltransferase
3.45768	−81	AATGTCGAGCAGATCACGGA	–	ML0023	Hypothetical protein
3.45074	−329	ATTGTGCGCCGTATCACGGG	rplY	ML0245	50S ribosomal protein L25
3.45049	−248	ACGGTAAGTGGGCTGACGAA	–	ML0383	Hypothetical protein
3.45045	−12	AGGGTCAAAACCATGACCTC	pgi	ML0150	Glucose-6-phosphate isomerase
3.43921	−398	AGTGTCGGCGAAATCACATT	mas	ML0139	Putative mycocerosic synthase
3.43748	−86	GCAGTGGAATTTATCACGAT	–	ML0065	Putative monooxygenase
3.43732	−26	ATGGTCAGTGCATTAACACG	–	ML0162	Hypothetical protein
3.43312	−135	TTTGTCACACCCCTTACCGG	–	ML0314	Putative esterase
3.42825	−91	AATGTGATTTCGCCGACACT	fadD28	ML0138	acyl-CoA synthase
3.42801	−396	TACGTCATCGACGCCACGGA	rpsI	ML0365	30S ribosomal protein S9
3.42317	−94	CGTGTGGTGTTATTCACTAC	fbpC	ML0098	Antigen 85C, mycolyltransferase

Interestingly, most of the novel operons are functionally critical for *Mtb* and several of them are conserved (see below) within the pathogenic mycobacteria. For example *Rv0904* (AccD3) and *Rv0993* (galU) were observed among them and interestingly these enzymes were found as key factors during cell wall biogenesis. Another gene *Rv3645* (membrane

linked adenylyl cyclase) was also found to be conserved in comparative regulon analyses. It is tempting to speculate that this adenylyl cyclase could be acting as a switch during cAMP based signaling in cellular stress and perhaps responsible for maintaining cAMP homeostasis in bacterial cells. *Rv2918* (gln D) and *Rv992* (putative 5-formyltetrahydrofolate cyclo-ligase) were also among the novel predicted boxes that were observed to be conserved in these regulons. These enzymes are also related to vital metabolic pathways of these pathogens.

2.3.2.1 Boxes associated with cell wall biogenesis

One of the novel-binding sites was located upstream of *Rv0993* (gal U), which is actually UTP: alpha-D-glucose-1-phosphate uridylyltransferase. It converts glucose-1-phosphate to UDP-glucose, that is of central importance in the synthesis of the components of the cell envelope of *E. coli* and in both galactose and trehalose metabolism (Weissborn *et al.*, 1994). In case of *Mtb*, galactose is essential for the linking of the peptidoglycan and mycolic acid cell wall layers and is therefore essential for survival of mycobacteria (Weston *et al.*, 1997). Another interesting box observed was the one associated with regulation of *Rv3031*, a conserved hypothetical protein actually belonging to glycoside hydrolase family of proteins. Further, it was observed that *Rv3032* is present in the same operon downstream to *Rv3031,* which was

previously annotated as glycosyl transferase. Members of this family of enzymes transfer activated sugars (UDP, ADP, GDP or CMP linked sugars) to a variety of substrates, including glycogen, fructose-6-phosphate and lipopolysaccharides. These enzymes are directly involved in cell wall biogenesis. A novel site upstream of *Rv0643c,* a methoxy mycolic acid synthase, involved in mycolic acid biosynthesis which is a key component of the mycobacterial cell wall (Kremer *et al.*, 2000) was also found. Yet another novel CRP binding another site was located within *Rv0904c (*AccD3*).* This gene codes for putative acetyl CoA carboxylase carboxyl transferase, which catalyses the initial steps of fatty acid biosynthesis and was reported as a key enzyme in mycolic acid biosynthesis (Gande *et al.*, 2004), an exclusive component of mycobacterial cell wall biogenesis.

Table 2.4: Predicted CRP binding sites in *M. smegmatis* genome

Score	Position	Binding site	Gene	Product
3.98115	−42	AATGTGAGGATCGTCACGCG	MSMEG0397	Conserved hypothetical protein
3.9723	−126	GATGTGATCGTCGTCACGTG	MSMEG0161	Oxalate/formate antiporter
3.9668	−207	ATCGTGATCTGCCTCACGTT	MSMEG4634	Conserved hypothetical protein
3.96558	−144	ATTGTGATGTGTATCACGGT	MSMEG5503	Succinyl-CoA synthetase subunit beta like protein
3.96441	−135	ACTGTGACGCGCATCACGTT	MSMEG6785	Conserved hypothetical protein
3.96317	−45	CTTGTGATGCACGTCACGAC	MSMEG3787	Hypothetical protein
3.96203	−385	GTCGTGATGAATGTCACGTC	MSMEG0953	Hypothetical protein
3.95591	−61	AGTGTGATTTACATCACACC	MSMEG2762	Conserved hypothetical protein
3.95501	−195	AACGTGACGCGCATCACGTC	MSMEG0413	Carboxyphosphonoenolpyruvate phosphonomutase-like protein
3.95431	−327	GGCGTGATGCCGGTCACGGG	MSMEG4836	Oxidoreductase
3.94687	−18	GGTGTGAGCTGTCTCACATG	MSMEG0739	Carbon monoxide dehydrogenase
3.94579	−79	GCTGTGAATCCAGTCACAGC	MSMEG3635	Hypothetical protein
3.94366	−147	ATCGTGATCTGCCTCACACT	MSMEG4561	Aldo/keto reductase
3.94267	−157	CGCGTGACGTGGCTCACGCG	MSMEG4632	Conserved hypothetical protein
3.94199	−125	GGTGTGATGTAAATCACACT	MSMEG2763	Conserved hypothetical protein
3.93621	−177	AGCGTCACCTGCGTCACGGT	MSMEG3816	Universal stress protein family domain protein
3.936	−107	GGCGTGACCCCGATCACGAG	MSMEG5035	Multidrug resistance transporter

3.9321	−233	CTCGTGATGTCGCTCACGCC	MSMEG1818	Phosphoribosylaminoimidazole carboxylase
3.92604	−221	TTCGTGATGGCGGTCACGCT	MSMEG0576	STAS domain protein
3.92588	−240	GCCGTGACATGCGTGACGTC	MSMEG2221	Substrate-CoA ligase
3.92112	−236	GGTGTCACCGAGGTCACGGG	MSMEG3365	Conserved hypothetical protein
3.92019	−42	AGTGTGAACTGTGTCACCTC	MSMEG0349	Ribonucleoside-diphosphate reductase
3.91823	−167	GCTGTGACTGGATTCACAGC	MSMEG3636	Transcription regulator
3.91242	−136	ACCGTGATACACATCACAAT	MSMEG5504	Lipoprotein NlpD
3.91112	−68	GACGTGAGCACCCTCACACG	MSMEG0950	Conserved hypothetical protein
3.91111	−183	GGCGTGAGGGAGCTCACGAA	MSMEG4148	Transcriptional regulator tetR family
3.91015	−54	GCCGTGATGGCAGTCACAAC	MSMEG6453	Hypothetical protein
3.90522	−207	GACGTGACACCCGTGACGGT	MSMEG2272	Conserved hypothetical protein
3.9047	−64	AGAGTGACCTCGGTCACGCT	MSMEG3737	trp-G type glutamine amidotransferase/dipeptidase
3.89664	−158	GGCGTGATCGTCGTGACGCT	MSMEG6098	Hypothetical protein
3.89398	−109	GATGTGACACCTGTGACAGT	MSMEG5447	Conserved hypothetical protein
3.89266	−137	GTGGTGATCTAGATCACGCT	MSMEG0195	Hypothetical protein
3.89196	2	ACTGTGAAGCGAATGACGGT	MSMEG1555	Hypothetical protein
3.89164	−41	AACGTCACGCCCATCACGCC	MSMEG2055	nuoJ
3.88848	−65	TCTGTCACATATCTCACGTT	MSMEG6232	Bacterial NAD-glutamate dehydrogenase superfamily
3.88613	−216	CGCGTGACGATCCTCACATT	MSMEG0396	fadE5
3.88278	−123	CCGGTGACCACGGTCACGCC	MSMEG1565	Glucosamine-fructose-6-phosphate aminotransferase
3.88066	−230	GTCGTAACCTGCGTCACGCG	MSMEG5242	Channel protein
3.87833	−56	AACGTGACCCAGGTCACTTA	MSMEG0158	Oxalate/formate antiporter
3.8772	−109	ACCGTAATCTGCGTCACGTG	MSMEG5382	Dehydrogenase DhgA
3.87575	−371	TGCGTGAAGGCGTTCACACC	MSMEG1123	Probable metalloprotease zinc transmembrane protein
3.87495	−60	CCTGTGAGCCGGGTCACCAC	MSMEG2582	Acetyltransferase
3.87448	−106	GCTGTGACCGCCGTCACCAG	MSMEG1048	Conserved hypothetical protein
3.87374	−47	AATGTGAGCTGCGTAACACC	MSMEG0894	Aldehyde dehydrogenase family protein
3.87358	−63	TCAGTGACCTGGGTCACGTG	MSMEG4216	Uncharacterized BCR
3.87319	−78	GACGTCACCGGGCTCACGAT	MSMEG6192	Acyltransferase
3.87309	−108	GCTGTGATGGAACTGACGGG	MSMEG0722	Hypothetical protein
3.87203	−142	GGCGTCATGCAGCTCACGAT	MSMEG2104	N5IS1096
3.86893	−220	AGCGTGATCTAGATCACCAC	MSMEG0194	Transcriptional regulator
3.86792	−25	TATGTGATCTACGTCACTGG	MSMEG0142	Probable transcription regulator protein
3.86576	−230	TCGGTGAAGCCTGTCACGCG	MSMEG2530	Hypothetical protein agmatinase
3.86496	−16	GGCGTGAAGTTCATGACGAA	MSMEG1056	Conserved hypothetical protein
3.86432	−11	TCGGTGAGGCCCGTCACGTT	MSMEG6429	Domain of unknown function (DUF427) superfamily
3.86424	6	AGTGTGACCGCCATGACACA	MSMEG0299	Aldehyde dehydrogenase
3.86323	−271	GTCGTCATAGCCTTCACGTT	MSMEG2143	Fic protein family
3.86273	−76	GGCGTGACCGTGGTCACCGG	MSMEG1564	Conserved hypothetical protein
3.86087	−15	ACGGTGATCGTGCTCACGTT	MSMEG2733	Conserved hypothetical protein
3.86075	−62	AACGTGACTTGCCTCACTTC	MSMEG5285	dctA
3.86056	−151	GACGTCACGGCGATCACGCA	MSMEG2946	glycosyl transferase 2-hydroxy-3-oxopropionate reductase
3.86051	−237	TTCGTCACGTCGATCACGGC	MSMEG5848	Conserved hypothetical protein
3.86001	−74	AACGTGAAGGCTATGACGAC	MSMEG2144	gp36
3.86	−232	GGCGTCAGGGTGCTCACGCG	MSMEG5860	
3.85957	−316	ACCGTGACGGTTCTGACGAT	MSMEG4183	Hypothetical protein

2.3.2.2 Putative regulatory elements controlling 5'-3' Cyclic Adenosine Monophosphate (cAMP) signaling

A CRP binding box upstream of *Rv0104*, a conserved hypothetical protein, was also identified. Upon sequence analysis, it was found to have a cyclic AMP binding domain at its C-terminal. This domain is very similar to the effector domain of CRP family proteins and cAMP binding domain (regulatory domain) of cAMP dependent protein kinases thereby suggesting a role for this protein in cAMP mediated signaling in *Mtb.* Further it was observed that *Rv0103c* (probable cation-transporter) and *Rv0104* share common CRP binding sites and probably represent a case of divergent gene expression.

A CRP binding box upstream to the gene encoding membrane linked adenylyl cyclase (*Rv3645)* was also identified. Adenylyl cyclases catalyze the production of cAMP, which further acts as a secondary signal. This protein, which was previously speculated to be involved in cAMP mediated signaling in mycobacteria (Linder and Schultz, 2003), is anchored to the membrane in *Mtb.* Gene regulation of adenylyl cyclase could be a case of feed back regulation for activation of CRP.

2.3.2.3 Other potential boxes

Novel binding site was also present in *Rv0520,* a methyl transferase believed to be involved in ubiquinone pathway. The presence of a CRP binding site in *Rv2699*, has already been reported earlier (Bai *et al.*, 2005). Sequence analysis of *Rv2699* revealed similarity to methyl transferase proteins involved in ubiquinone pathway suggesting that the two enzymes with related function may be expressed under the control of the same regulator.

Another important element observed within *Rv2918c* (GlnD-uridyl transferase) regulates the catalytic activity of glutamine synthetase (Garcia and Rhee, 1983). *Rv3113* (phosphatase) has a new CRP binding site regulating the gene *Rv3114* (nucleoside deaminase) involved in salvage pathways of nucleotides. *Rv3505* (fadE27) and *Rv3617* (ephA) also carried new CRP binding sites, possibly involved in regulating probable acyl-CoA dehydrogenase and putatative epoxide hydrolase, respectively.

2.3.3 CRP regulon with operon context in *M. leprae*, *M. avium subsp. paratuberculosis* and *M. smegmatis*

Given the observation that the CRP proteins from all mycobacteria have identical DNA binding domains, the same profile matrix constructed

for *Mtb* to predict CRP binding sites in the genomes of *M. leprae, M. avium* subsp. *paratuberculosis and M. smegmatis* (Tables 2.2, 2.3 and 2.4) was extended. This is the first attempt to interrogate these genomes for CRP regulon signatures. The operon context of these regulons across the genomes is presented in Tables 2.5 to 2.8.

Table 2.5: Predicted CRP regulated operons in *M. tuberculosis*. Bold indicates gene identified in this study

Synonym	Gene	Product
Rv1552	frdA	fumarate reductase
Rv1553	frdB	PROBABLE FUMARATE REDUCTASE [IRON-SULFUR SUBUNIT]
Rv1554	frdC	PROBABLE FUMARATE REDUCTASE [MEMBRANE ANCHOR SUBUNIT]
Rv1555	frdD	fumarate reductase subunit D
Rv0951	sucC	succinyl-CoA synthetase subunit beta
Rv0952	sucD	succinyl-CoA synthetase alpha subunit
Rv1810	-	hypothetical protein
Rv0993	**galU**	**UTP--GLUCOSE-1-PHOSPHATE URIDYLYLTRANSFERASE**
Rv0104	**-**	**hypothetical protein (cAMP binding protein)**
Rv3645	**-**	**membrane linked adenylyl cyclase**
Rv2918c	**glnD**	**PII uridylyl-transferase**
Rv0103c	**ctpB**	**PROBABLE CATION-TRANSPORTER P-TYPE ATPASE B CTPB**
Rv2591	PE_PGRS44	PE-PGRS FAMILY PROTEIN
Rv2700	-	POSSIBLE CONSERVED SECRETED ALANINE RICH PROTEIN
Rv0905	echA6	enoyl-CoA hydratase
Rv0906	-	hypothetical protein
Rv1185c	fadD21	acyl-CoA synthase
Rv0885	-	hypothetical protein
Rv0886	fprB	PROBABLE NADPH:ADRENODOXIN OXIDOREDUCTASE

Rv0884c	serC	phosphoserine aminotransferase
Rv0992c	**-**	**hypothetical protein (5-formyltetrahydrofolate cyclo-ligase)**
Rv1581c	-	Probable phiRv1 phage protein
Rv1580c	-	Probable phiRv1 phage protein
Rv1579c	-	Probable phiRv1 phage protein
Rv0950c	-	hypothetical protein
Rv2699c	-	hypothetical protein
Rv0904c	**accD3**	**PUTATIVE ACETYL-COENZYME A CARBOXYLASE CARBOXYL TRANSFERASE**
Rv1159	**-**	**CONSERVED TRANSMEMBRANE PROTEIN**
Rv2650c	**-**	**POSSIBLE phiRv2 PROPHAGE PROTEIN**
Rv1158c	**-**	**CONSERVED HYPOTHETICAL ALA-, PRO-RICH PROTEIN**
Rv1157c	**-**	**CONSERVED HYPOTHETICAL ALA-, PRO-RICH PROTEIN**
Rv0019c	**-**	**hypothetical protein**
Rv2408	**PE24**	**POSSIBLE PE FAMILY-RELATED PROTEIN**
Rv0643c	**mmaA3**	**METHOXY MYCOLIC ACID SYNTHASE 3**
Rv1230c	**-**	**POSSIBLE MEMBRANE PROTEIN**
Rv1229c	**mrp**	**PROBABLE MRP-RELATED PROTEIN MRP**
Rv1566c	**-**	**Possible inv protein**
Rv1759c	**wag22**	**PE-PGRS FAMILY PROTEIN**
Rv1760	**-**	**hypothetical protein**
Rv3650	**PE33**	**PE FAMILY PROTEIN**
Rv2407	**-**	**ribonuclease Z**
Rv2406c	**-**	**hypothetical protein**
Rv3113	**-**	**POSSIBLE PHOSPHATASE**
Rv3114	**-**	**hypothetical protein (nucleoside deaminase)**
Rv3505	**fadE27**	**PROBABLE ACYL-CoA DEHYDROGENASE FADE27**

Rv3031	**-**	**hypothetical protein**
Rv3032	**-**	**POSSIBLE TRANSFERASE**
Rv3323c	**moaX**	**PROBABLE MOAD-MOAE FUSION PROTEIN MOAX**
Rv3322c	**-**	**POSSIBLE METHYLTRANSFERASE**
Rv0453	**PPE11**	**PPE FAMILY PROTEIN**
Rv3617	**ephA**	**PROBABLE EPOXIDE HYDROLASE**
Rv3618	**-**	**POSSIBLE MONOOXYGENASE**
Rv1386	**PE15**	**PE FAMILY PROTEIN**
Rv1387	**PPE20**	**PPE FAMILY PROTEIN**
Rv0520	**-**	**POSSIBLE METHYLTRANSFERASE/METHYLASE (FRAGMENT)**
Rv0521	**-**	**POSSIBLE METHYLTRANSFERASE/METHYLASE (FRAGMENT)**

Genes which are together are part of an operon

Table 2.6: Predicted CRP regulated operons in *M. avium sub sp. Paratuberculosis*

Synonym	Gene	Product
MAP3737	-	hypothetical protein
MAP3267c	-	hypothetical protein
MAP3266c	-	hypothetical protein
MAP2817	-	hypothetical protein
MAP2816c	-	hypothetical protein
MAP3284c	fadD29	FadD29
MAP3283c	-	hypothetical protein
MAP3282c	-	hypothetical protein
MAP3281c	-	hypothetical protein
MAP3671	-	hypothetical protein
MAP3672	nrdB	ribonucleotide-diphosphate reductase beta subunit
MAP0840	echA6	enoyl-CoA hydratase
MAP0841	-	hypothetical protein
MAP0842	-	hypothetical protein
MAP0843	ctpE	CtpE

MAP0174	-	hypothetical protein
MAP0227c	-	hypothetical protein
MAP2149c	-	hypothetical protein
MAP2220	-	ribonuclease Z
MAP0670	lpqR	LpqR
MAP0727	-	hypothetical protein
MAP0839c	accD3	AccD3
MAP2219c	-	hypothetical protein
MAP0632	pepC	putative aminopeptidase 2
MAP0633	-	hypothetical protein
MAP4094c	recC	RecC
MAP1418c	-	hypothetical protein
MAP1601	-	hypothetical protein

Genes which are together are part of an operon

Table 2.7: Predicted CRP regulated operons in *M. leprae*

Synonym	Gene	Product
ML0201	-	hypothetical protein
ML0185	-	hypothetical protein
ML0182	galU	putative UTP-glucose-1-phosphate uridylyltransferase
ML0155	sucC	succinyl-CoA synthetase subunit beta
ML0156	sucD	succinyl-CoA synthetase alpha subunit
ML0072	sodA	superoxide dismutase
ML0298	-	sulfur carrier protein ThiS
ML0297	thiG	thiazole synthase
ML0141	-	hypothetical protein
ML0181	-	hypothetical protein
ML0154	-	hypothetical protein

ML0333	-	hypothetical protein
ML0410	-	putative PE-family protein
ML0229	-	hypothetical protein
ML0230	panC	pantoate--beta-alanine ligase
ML0231	panD	aspartate 1-decarboxylase precursor
ML0232	-	hypothetical protein
ML0240	-	hypothetical protein
ML0173	rpmF	50S ribosomal protein L32
ML0243	-	acyl-CoA synthase
ML0336	-	putative ABC-tranporter ATP-binding protein
ML0335	-	putative ABC-transporter transmembrane protein
ML0107	-	hypothetical protein
ML0279	-	hypothetical protein
ML0333	-	hypothetical protein
ML0410	-	putative PE-family protein
ML0229	-	hypothetical protein
ML0230	panC	pantoate--beta-alanine ligase
ML0231	panD	aspartate 1-decarboxylase precursor

Genes which are together are part of an operon

Table 2.8: Predicted CRP regulated operons in *M. smegmatis*

Gene	Product
MSMEG0397	conserved hypothetical protein
MSMEG0161	oxalate/formate antiporter
MSMEG4634	conserved hypothetical protein
MSMEG5503	carbamoyl-phosphate synthase
MSMEG6869	Hypothetical transcriptional regulator Rv0043c/MT0049/Mb0044c
MSMEG6785	conserved hypothetical protein
MSMEG3787	hypothetical protein

MSMEG3788	fadE21
MSMEG0953	hypothetical protein
MSMEG2762	conserved hypothetical protein
MSMEG0413	carboxyphosphonoenolpyruvate phosphonomutase-like protein
MSMEG4836	oxidoreductase
MSMEG0739	carbon monoxide dehydrogenase
MSMEG0741	carbon monoxide dehydrogenase
MSMEG0743	ATPase
MSMEG0745	VWA domain containing CoxE-like protein family
MSMEG6786	conserved hypothetical protein
MSMEG3635	hypothetical protein
MSMEG4561	aldo/keto reductase
MSMEG4632	conserved hypothetical protein
MSMEG2763	conserved hypothetical protein
MSMEG3816	universal stress protein family domain protein
MSMEG5035	multidrug resistance transporter
MSMEG1818	phosphoribosylaminoimidazole carboxylase
MSMEG0576	STAS domain protein
MSMEG2221	substrate--CoA ligase
MSMEG3365	conserved hypothetical protein

Genes which are together are part of an operon

2.3.4 Conserved orthologues of CRP regulated genes across mycobacteria

A comparative analysis of CRP target genes in various mycobacteria enabled to identify the common CRP regulated genes across mycobacteria and at least 18 genes were found to be common (Table 2.9). Conservation of these genes in the predicted CRP regulons suggests an important role of their gene products in the life cycle of mycobacteria.

Table 2. 9
Distribution of conserved orthologues of CRP regulated genes across mycobacterial genomes

Gene	Product	*Mtb*	Mavi	Mlep	Msme
–	Hypothetical protein (possible secreted alanine rich protein)	Rv2700	MAP2817		MSMEG2762
–	Hypothetical protein	Rv2699c	MAP2816c		
echA6	enoyl-CoA hydratase	Rv0905	MAP0840		
–	Hypothetical protein (Beta lactamase type Zn dependent hydrolase)	Rv0906	MAP0841		
–	Hypothetical protein (AmpC, Beta-lactamase class C)	Rv0907	MAP0842		
ctpE	CtpE	Rv0908	MAP0843		
accD3	AccD3	Rv0904c	MAP0839c		
–	Hypothetical protein (signaling protein with CBS and cAMP binding domain)	Rv2406c	MAP2219c		
sucC	Succinyl-CoA synthetase subunit beta	Rv0951	MAP0896	ML0155	MSMEG5503
sucD	Succinyl-CoA synthetase alpha subunit	Rv0952	MAP0897	ML0156	
glnD	PII uridylyl-transferase	Rv2918c	MAP2986c		
fadE9	FadE9	Rv0752c	MAP4214c		
mmsB	MmsB	Rv0751c	MAP4213c		
fbpC1	FbpC1 (Ag85C)	Rv3803c	MAP0217	ML0098	
–	Membrane linked adenylate cyclase	Rv3645		ML0201	
galU	Putative UTP-glucose-1-phosphate uridylyltransferase	Rv0993		ML0182	
–	Hypothetical protein (5-formyltetrahydrofolate cyclo-ligase)	Rv0992c		ML0181	
–	Hypothetical protein (similar to peptidase)	Rv0950c	MAP0895c	ML0154	MSMEG5504

Upstream of the corresponding gene do not have CRP box, gene lies with in CRP regulated operon.
Blank space means orthologues are not present.

Most of the genome decay in *M. leprae* is *via* deletion and pseudogenization of its genes. Not all genes however were rendered superfluous by this mechanism of evolution.

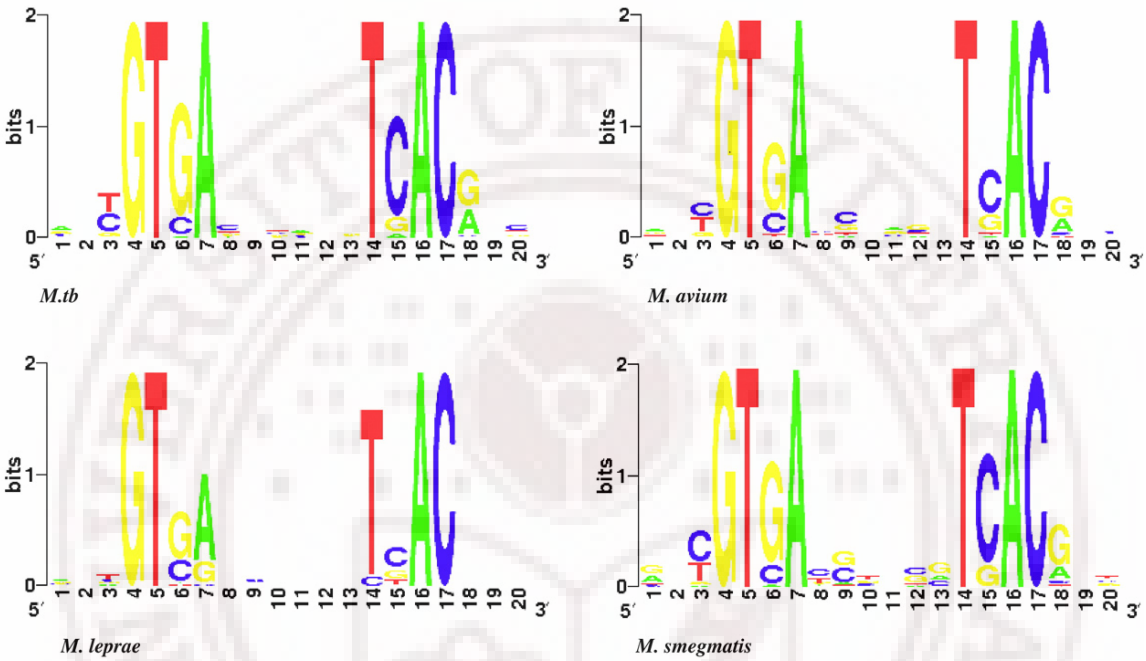


Figure 2.2: Sequence logo of the predicted CRP-binding sites in *Mtb*, *M. avium*, *M. leprae and M. smegmatis*. The height of each stack of letters represents the degree of sequence conservation measured in bits. The height of each letter within a stack is proportional to its frequency at that position in the binding site. The letters are sorted with the most frequent on top. This sequence logo was generated using the online Weblogo programme (http://weblogo.berkeley.edu/).

It has been still a quite successful pathogen in terms of virulence because it retained much of its core genome. Regulation of virulence therefore is still a key issue in *M. leprae* that makes it an interesting pathogen whenever comparative genomics of slow growing mycobacteria is addressed. Homology of most of its pseudogenized genes to 'live' genes of the present day *Mtb* makes the case even more interesting. Pseudogenized chunks still exist within the genome and homology

searches therefore are capable of pointing out those CRP like regions when comparative genomic exercises are carried out. This gives a lot of insight into the organization and arrangement of these genes and about the rate of substitutions and clock rate at which the genes such as CRP like regulons are pseudogenized.

2.3.4.1 Genes related to cAMP signaling in mycobacterium

CRP binding sites from *Rv3645* (membrane linked adenylate cyclase) and *Rv2406c* (cAMP binding protein) were found to be conserved in pathogenic mycobacteria (Table 2.9). The membrane-linked adenylate cyclase catalyzes the production of cAMP and, as discussed earlier may be involved in signaling in mycobacteria [Linder *et al.*, 2003]. *Rv2406c* encodes a hypothetical protein, sequence analyses of which reveals the presence of a CBS domain and a cAMP-binding domain. This protein might also be a part of cAMP regulated signal network in mycobacteria. *Rv2406c* also has CRP-binding element, which is conserved in pathogenic mycobacteria (Table 2.9).

2.3.4.2 Antibiotic resistance operon

Rv0905 (echA6), *Rv0906*, *Rv0907* and *Rv0908* (ctpE) constitute an operon which, in comparative analyses presented here, appeared to be conserved across mycobacterial genomes in terms of the CRP binding site

(Table 2.9). *Rv0906* and *Rv0907* code for two hypothetical proteins and upon sequence analyses and conserved domain search, presented in this chapter, these were found to belong to beta lactamase family of proteins. Rv0906 is a putative beta-lactamase and belongs to a class of Zn^{2+} dependent hydrolase. *Rv0907* belongs to Amp C class of beta lactamase. The beta-lactam class of antibiotics has not been used in the treatment of *Mtb* or other mycobacterial infections as mycobacteria are resistant to these antibiotics and produce beta-lactamases (Kasik, 1979; Kwon *et al.*, 1995). However, the effectiveness of beta-lactam/beta-lactamase inhibitor combinations has been shown *in-vitro* for *Mtb* (Chambers *et al.*, 1995), and also in clinical settings of TB (Chambers *et al.*, 1998; Cynamon *et al.*, 1983; Segura *et al.*, 1998; Sorg and Cynamon, 1987), *M. avium* (Casal *et al.*, 1987), and *M. smegmatis* (Yabu *et al.*, 1985). *Rv0908* (ctpE), which codes for putative cation transporter, is also present in the same operon. Metallo beta lactamase family enzymes are metal dependent and require Zn^{2+} for their activity (Schilling *et al.*, 2005). Both a putative Zn^{2+} dependent beta lactamase and probable Zn^{2+} transporter are part of this operon under same control and seem to have related function.

2.3.4.3 Cell wall Components

Another conserved CRP binding site observed in these analyses was present in *Rv3803* (Ag85c), *MAP0217* and *ML0098* orthologues from *Mtb*, *M. avium* subsp. *paratuberculosis* and *M. leprae*, respectively. The Ag85 complex, family of three proteins of 30 to 32kDa (Ag85A, Ag85B, and Ag85C), forms a major fraction of the secreted proteins in *Mtb* culture filtrate. All of this complex possesses a mycolyl transferase enzyme activity essential for the final stages of mycobacterial cell wall assembly (Belisle *et al.*, 1997) and the same is also involved in the host immune surveillance and defense system (Takayama *et al.*, 2005). These three proteins are coded by three paralogous genes located in distinct regions of the bacterial genome (Content *et al.*, 1991). These genes do not show any resemblance in their 5' upstream region, and are probably regulated independently at the transcriptional level (Content *et al.*, 1991). Interestingly, a CRP box found was found upstream of only *Rv3803* (Ag85c) and this DNA element was conserved across all pathogenic mycobacteria and interestingly, was absent in *M. smegmatis* (Table 2.9). Deletion of the gene encoding antigen 85C protein alters the bacterial cell wall and its permeability, but does not kill the cells (Jackson *et al.*, 1999).

CRP binding site in *Rv0993* (gal U), which encodes UTP: alpha-D-glucose-1-phosphate uridylyltransferase, is also conserved in its *M. leprae* orthologue. As discussed in earlier section the corresponding gene product is important in cell envelope biogenesis of mycobacteria (Weston *et al.*, 1997). *Rv0904c* (AacD3) also has a CRP binding site, which is conserved in pathogenic mycobacteria *Mtb,* and *M. avium* subsp. *paratuberculosis*. *Rv0905* encodes putative Enoyl-CoA hydratase which catalyses the elongation of unsaturated fatty acid chain. *Rv0904c* (AccD3), which codes for putative beta sub-unit of acetyl-coenzyme A carboxylase carboxyl transferase is an enzyme of the mycolic acid biosynthetic pathway (Gande *et al.*, 2004) and is critical for cell wall formation. *Rv0904* shares CRP binding sites with *Rv0905*, an example of divergent regulation. It is therefore, interesting to document divergent regulation of operons amidst a convergently evolving genome.

2.3.4.4 Metabolic Enzymes

The *sucCD* operon comprising of *Rv0951* (sucC) *and Rv0952* (sucD) which encode succinyl coA synthetase beta and alpha respectively, has a CRP binding site that is conserved among mycobacteria (Table 2.9). Succinyl-CoA synthetase is responsible for carrying out two unrelated but vital metabolic functions. One, it catalyzes the substrate-level phosphorylation step of the citric acid cycle (Kaufman *et al.*, 1953), and

two, it replenishes succinyl-CoA for ketone body catabolism (Ottaway *et al.*, 1981) and for porphyrin synthesis (Labbe *et al.*, 1965). The presence of the same locus of an unrelated ORF on opposite strand *Rv0950c* which shares a common CRP binding site with *sucCD* operon and encoding a hypothetical protein was observed which could also be a case of divergent transcription regulation. *Rv0752c* (fadE9) and *Rv0753c* (MmsB) constitute an operon and have a CRP binding box, which is conserved in both *Mtb* and *M. avium* subsp. *paratuberculosis* orthologues. *Rv0752c* encodes a putative Acyl CoA dehydrogenase, which is a flavoprotein catalyzing desaturation of acyl-CoA esters and plays an important role in the oxidation of fatty acyl-CoA esters. The ORF *Rv0753c* encodes a putative 3-Hydroxy isobutyrate dehydrogenase catalyzed oxidation of 3-Hydroxyisobutyrate to methylmalonate semialdehyde.

The search of the CRP-binding sites was broadened in intergenic regions of annotated as well as non-annotated open reading frames. This was then extended to other mycobacterial genomes to pin-point the common CRP regulated genes across the genus.

2.4 Conclusion

Rv3676 (*Mtb*-CRP) was earlier reported to be essential for the survival of mycobacteria inside macrophages and in animal models (Rickman *et al.*, 2005). Further, based on the analyses presented in this chapter, high conservation of these CRP regulated genes among pathogenic mycobacteria than in non-pathogenic mycobacteria was evident. This strengthens the notion that *Mtb*-CRP and its regulated genes are important in pathogenesis of mycobacteria and that these might have co-evolved with the pathogenic branch as a result of genome optimization aimed at devising new survival strategies. That many of these predicted target proteins are critical for the survival of the mycobacterium in the hostile environment of the macrophage adds a new dimension to the understanding of the regulatory complexity in *Mtb*. This computer based predictions of complex networks of *Mtb* genes appear to be a provocative proposition wherein cAMP could be playing a critical role as an effector.

Chapter 3: Biophysical and Biochemical features of *Mtb*-CRP

A part of the work presented in this Chapter has been published as:

ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

International Journal of Medical Microbiology 297 (2007) 451–457

IJMM

www.elsevier.de/ijmm

Novel biochemical properties of a CRP/FNR family transcription factor from *Mycobacterium tuberculosis*

Yusuf Akhter[a], Smanla Tundup[a], Seyed E. Hasnain[a,b,c,d,*]

[a]*Laboratory of Molecular and Cellular Biology, CDFD, Hyderabad 500076, India*
[b]*Jawaharlal Nehru Centre for Advanced Scientific Research, Jakkur, Bangalore 560064, India*
[c]*University of Hyderabad, Hyderabad 500046, India*
[d]*Institute of Life Sciences, Hyderabad 500046, India*

3.1 Introduction

CRP/FNR (cAMP receptor protein/fumarate and nitrate reductase regulator) is one of the members of a family of transcriptional regulators. With over 370 family members, these DNA-binding proteins predominantly function as positive transcriptional regulators and are known to be associated with defense against oxygen stress and starvation, and at the same time respond to other environmental signals. The distinctive features of CRP/FNR superfamily proteins include the presence of a nucleotide-binding domain and a helix-turn-helix motif containing DNA-binding domain at the N- and C-terminal, respectively. The prototype cAMP-binding domain (Schultz *et al.*, 1991) is a versatile structure that has evolved to accommodate different functional specificities in response to a broad range of signals (Green *et al.*, 2001; Körner *et al.*, 2003). *Mtb* H37Rv ORF *Rv3676* codes for a putative CRP/FNR protein (Cole *et al.*, 1998), which is required for virulence in mice and controls transcription of specific genes (Rickman *et al.*, 2005). Recently, Bai *et al.* reported the characterization of *Mtb*-CRP (Rv3676) protein using computational and experimental methods (Bai *et al.*, 2005). Phylogenetically, *Mtb* Rv3676 is nearest to the CooA branch represented by the CO sensor protein of *Rhodospirillum rubrum* (Körner *et al.*, 2003). However, the relative positions of the regulatory and the DNA-binding domain are strikingly different in that the recognition helix of CooA is

rotated 180º away from the position occupied in CRP-cAMP. Further differences between CooA and CRP include an extended N-terminus providing a ligand to the heme of the opposite subunit, an 11-amino-acid extension in the regulatory domain (positions 72–82) to accommodate the heme and a different composition of the hinge region toward the C-terminus, which causes the displacement of the DNA-binding domain (Lanzilotta *et al.*, 2000). Sequence alignment suggests that these 11 residues are not fully conserved in Rv3676.

Escherichia coli CRP and FNR regulate transcription globally in response to glucose starvation and anaerobic conditions, respectively (Kolb *et al.*, 1993). *E. coli* FNR is structurally related to CRP except for the presence of four conserved cysteine residues at the N-terminal extension, which form part of an iron–sulfur cluster and a redox-sensing domain of FNR. This iron–sulfur cluster is absent in Rv3676 similar to other members of the same family from other systems such as *Pseudomonas stutzeri* (Vollack *et al.*, 1999) and *Bradyrhizobium japonicum* (Mesa *et al.*, 2003). Although these proteins do not have an iron–sulfur cluster, they are the regulators of oxygen tension *sensu stricto*. The earlier report by Bai *et al.* (2005) focused on CRP regulon prediction and the experimental validation of the same and provided the first direct evidence for cAMP binding to a transcription factor in *Mtb*, thereby suggesting a role for cAMP-mediated signal transduction in this

bacterium. In this Chapter, the purification and comprehensive characterization of a CRP/FNR regulator from *Mtb* in terms of oligomeric state, cAMP and DNA binding has been described. These results point to some new unusual properties of Rv3676 protein, which could have physiological relevance.

3.2 Materials and Methods

3.2.1 Bacterial strains and plasmids

E. coli DH5a and *E. coli* BL21DE3 bacterial strains were used for cloning and expression purposes, respectively. DNA manipulations were carried out in pET23a plasmid vector using standard techniques. Integrity of the plasmid constructs was confirmed by DNA sequencing.

3.2.2 Cloning, expression and purification of recombinant *Mtb* Rv3676

Mtb ORF Rv3676 was PCR amplified from *Mtb* H37Rv genomic DNA using forward (GGATAT**CATATG**GTGGACGAGATCCTGGCCAGGG) and reverse (CG**CTCGAG**CCTCGCTCGGCGGGCCAGTC) primers with restriction enzyme sites for cloning (shown in bold). The amplicon was cloned into the corresponding sites of pET23a, and recombinant Rv3676

protein was purified as a 6XHis-tagged fusion protein from *E. coli* BL21 (DE3) cells.

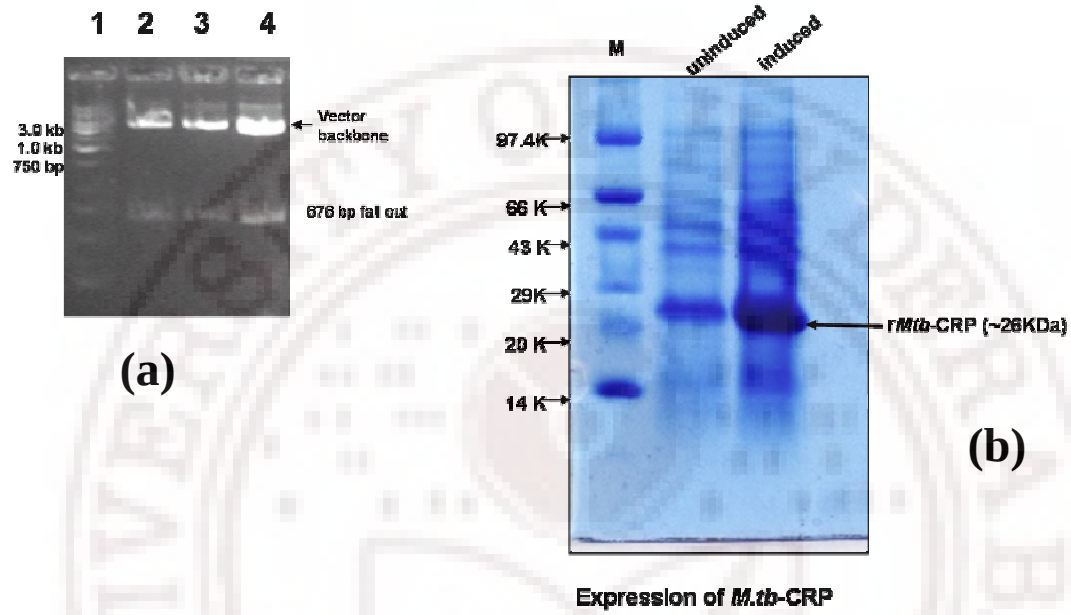


Figure 3.1: Cloning and expression of *Rv3676* (*Mtb*-CRP). (a) Identification of the positive clone of Rv3676 in pET23a vector by restriction digestion with *Nde*I and *Xho*I. Lane 1, 1Kb marker; lane 2, lane 3 and lane 4 shows release of Rv3676 insert cloned in pET23a after. (b) Integrity of positive clones was verified by sequencing and the plasmid carrying the insert was transformed into BL21 (DE3) cells and cultures were grown in 200 ml LB supplemented with 30 µg/ml kanamycin. The culture was induced at an OD600 of 0.4 with 0.4 mM IPTG at 300 K and 200 rev per min to allow protein expression. Protein expression was monitored by SDS–PAGE. Lanes marked "un-induced" and "induced" show expression of *Rv3676* without and with IPTG. Lane marked "M" shows protein molecular size markers.

Protein concentration was estimated using the dye-binding method (Bradford, 1976). To determine suitable storage conditions, aliquots of recombinant Rv3676 (rRv3676) were dialyzed in different buffers, namely phosphate-buffered saline (PBS), 10 mM Tris and 10 mM HEPES. Storage temperature was also optimized, and the conditions under which rRv3676 was most stable were selected.

3.2.3 Analytical size exclusion chromatography

The oligomeric state of native recombinant protein was determined by analytical size exclusion chromatography using a Superose 6 fast protein liquid chromatographic column (BIORAD) at room temperature with PBS as running buffer. A standard curve was prepared according to the instruction manual using standard molecular weight markers. The void volume was determined using Blue Dextran 2000. The elution parameter K_{av} was calculated as follows: $K_{av} = (V_e-V_0)/V_s$, where V_e is the elution volume for the protein, V_0 the column void volume, and V_s the total stationary phase volume. K_{av} was plotted against log molecular weight.

3.2.4 Spectral analyses

To detect the presence, if any, of any associated co-factor, absorption was measured between 200 and 800 nm using a Perkin-Elmer spectrophotometer. Fluorescence spectrometric measurements and ligand-binding assays were carried out using a Perkin-Elmer LS50B luminescence spectrometer and a sample volume of 200 ml with 0.3 cm path length. Tryptophan fluorescence was measured at an excitation wavelength of 295 nm. The slit widths for excitation and emission were 10 and 20 nm, respectively. Emission spectra were recorded between 310 and 500 nm. All spectra measurements were corrected by subtracting

the corresponding buffer backgrounds. Increasing concentrations of urea (1–8 M) and a constant concentration (3 μM) of recombinant protein was used to study the denaturation kinetics of the protein. At 8 M urea the protein was fully unfolded, and the spectrum of fully unfolded protein was further compared with that of 6 μM free tryptophan. The circular dichroism (CD) spectra of recombinant native protein and liganded protein, incubated with different concentrations of cAMP (6–16 μM), were recorded using a JASCO CD spectrometer (Model J-810) between 200 and 250 nm in steps of 0.5 nm with four accumulations in each step. The spectral baseline was corrected by subtracting the respective blanks. Molar ellipticity, expressed in millidegrees, was plotted as a function of wavelength. The secondary structure content of the protein was calculated by using k2d software (www.embl-heidelberg.de/~andrade/k2d/). For CD and fluorimetric spectral analysis, 5 and 3 mM recombinant protein was used, respectively.

3.2.5 Electrophoretic mobility shift assay (EMSA)

Gel retardation assays were carried out as described earlier (Prakash *et al.*, 2005). Complementary synthetic oligodeoxyribonucleotides corresponding to the CRP/FNR-binding site (AATGTGATCTAGGTCACGTG) present upstream of *Rv1552* (frdA) were end labeled with [g-32P]ATP using T4 polynucleotide kinase. One

nanogram labeled oligonucleotide was incubated with 3 mg recombinant protein in binding buffer (10 mM Tris–HCl, 50 mM NaCl, 50 mM MgCl2, 1 mg BSA, 1 mg poly dI:dC, 1 mM EDTA, 1 mM DTT, 1 mM PMSF and 10% glycerol) in 20 ml reaction volume, incubated for 30 min at room temperature and fractionated on a 5% polyacrylamide gel in TBE. After electrophoresis at 200 V at 4° C, the gel was dried and analyzed by autoradiography. To check for the specificity of the complex, unlabeled homologous oligonucleotide or an oligonucleotide carrying a specific mutation (mut) critical for binding (AATTTGATCTAGGTCACGTG, shown as underlined) was used in competition assays.

3.3 Results

3.3.1 Purified rRv3676 exists in dimeric state

Mtb rRv3676 was purified as a 6XHis-tagged protein using affinity chromatography as described in Materials and Methods. Purified rRv3676 was stable at 4° C in PBS while at lower temperatures and in other buffers it formed insoluble aggregates.

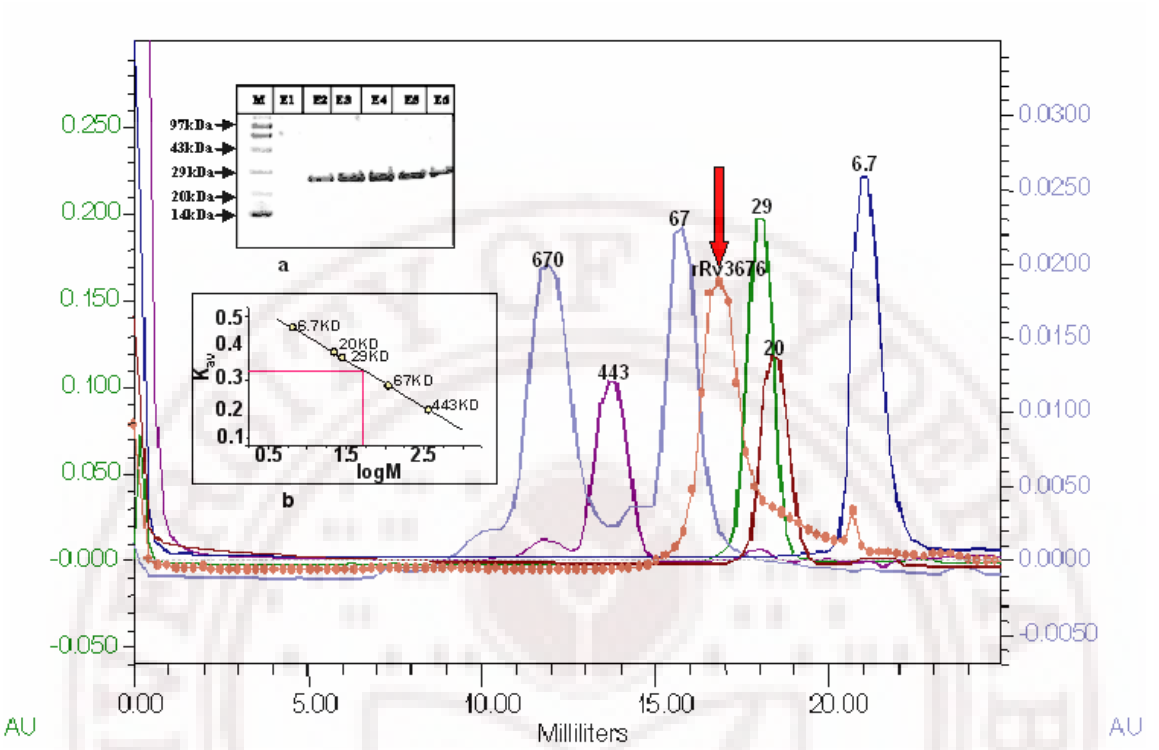


Figure 3.2: Recombinant Rv3676 protein exists as a dimer: (a) Coomassie-stained polyacrylamide gel showing the ORF *Rv3676*-encoded protein of *Mycobacterium tuberculosis* purified from *E. coli*. M represents the protein molecular size marker (Medium range, Genei, India.), E1–E6 show successive TALON column eluted fractions of the recombinant protein. (b) The purified protein was pooled and fractionated on a Superose 6 FPLC column, resulting in a single peak. The calculated molecular mass of the recombinant protein was ~53 kDa corresponding to a dimeric state.

The purified recombinant protein was fractionated by electrophoresis on a 10% polyacrylamide gel and stained with Coomassie Brilliant Blue (Figure 3.2a, inset). Gel filtration analysis was carried out to determine the oligomeric nature, if any, of the rRv3676 protein. rRv3676 exists as a pure dimer of ~53 kDa, as evident from analytical size exclusion chromatography (Figure 3.2).

3.3.2 Purified rRv3676 has no associated cation co-factors

In most oxygen tension-sensing proteins belonging to the same family, transition metals like Fe or Ni are associated with the protein to sense the fluctuations of oxygen availability *via* redox mechanisms (Körner *et al.*, 2003).

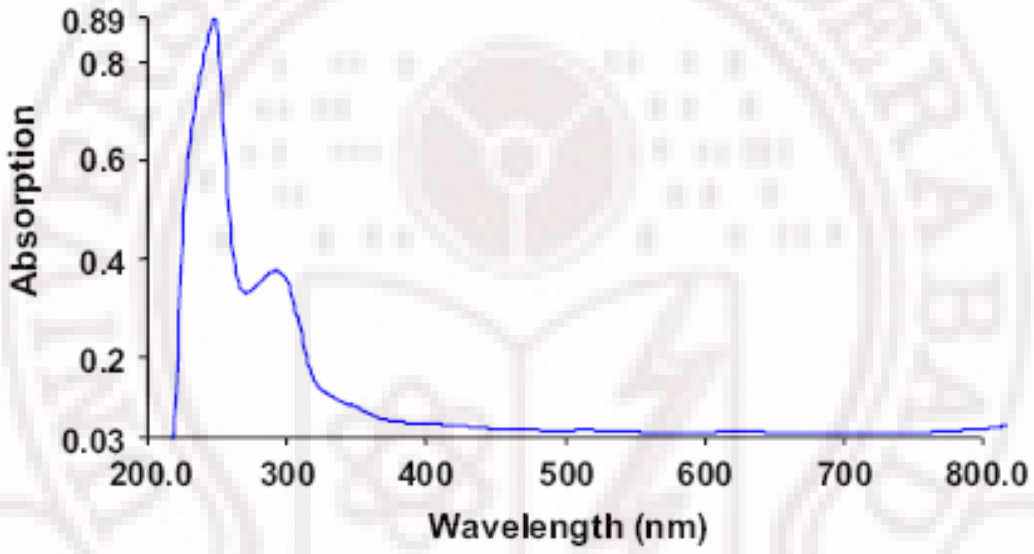


Figure 3.3: Absorption spectrum of purified rRv3676 indicating the absence of any metal ion co-factor. The spectrum shows two prominent peaks, one at 295 nm (corresponds to tryptophan) and the other at 280 nm (corresponds to tyrosine and phenylalanine).

Therefore, the absorption spectrum of purified rRv3676 was scanned to check for the presence, if any, of a metal ion co-factor. Spectral analysis revealed two peaks, one at 295 nm and the other at 280 nm (Figure 3.3). The first peak corresponds to tryptophan, while the other peak is due to phenylalanine and tyrosine. The fact that no any other peak was observed clearly indicates that rRv3676 does not have any other

associated metal ion co-factor. This rather unexpected finding suggests that *Mtb* Rv3676 apparently uses some other mechanism(s) to sense effector signals.

3.3.3 cAMP binds to purified rRv3676 in a concentration dependent manner

Results of protein family searches revealed the presence of a putative cAMP-binding domain at the N-terminal end of rRv3676 protein, thereby raising a strong probability that cAMP may be acting as an effector of Rv3676 protein. Therefore purified rRv3676 was subjected to CD analysis in the presence and absence of cAMP as ligand.

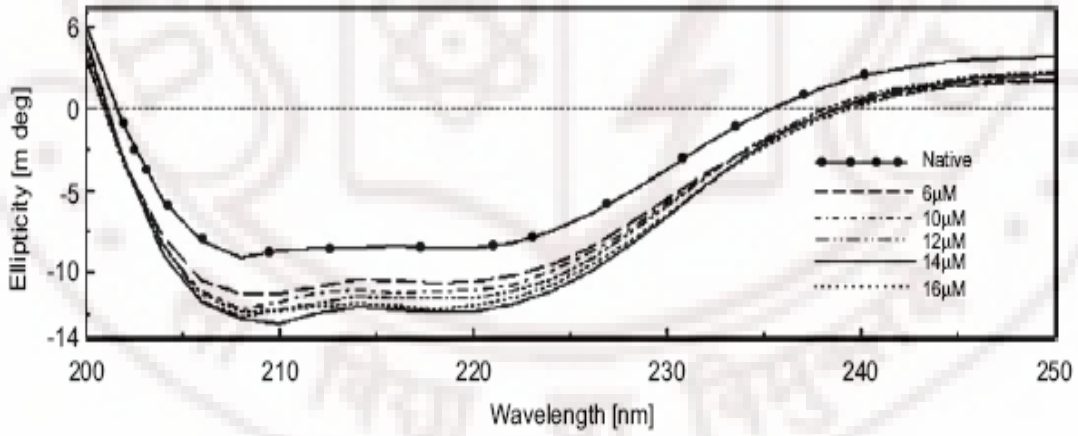


Figure 3.4: Purified rRv3676 protein displays cAMP-binding activity as evident from circular dichroism (CD) spectral analysis (see Materials and Methods). Binding of cAMP to rRv3676 is evident from a change in secondary structure of the native protein upon interaction with cAMP.

The change in secondary structure was calculated using k2d software (http://www.bork.embl-heidelberg.de/$andrade/k2d)based) using a

method developed earlier (Yang *et al.*, 1986). A comparison of CD spectra of cAMP-free and cAMP- bound rRv3676 provides evidence of binding (change in secondary structure of purified rRv3676 upon interaction with cAMP).

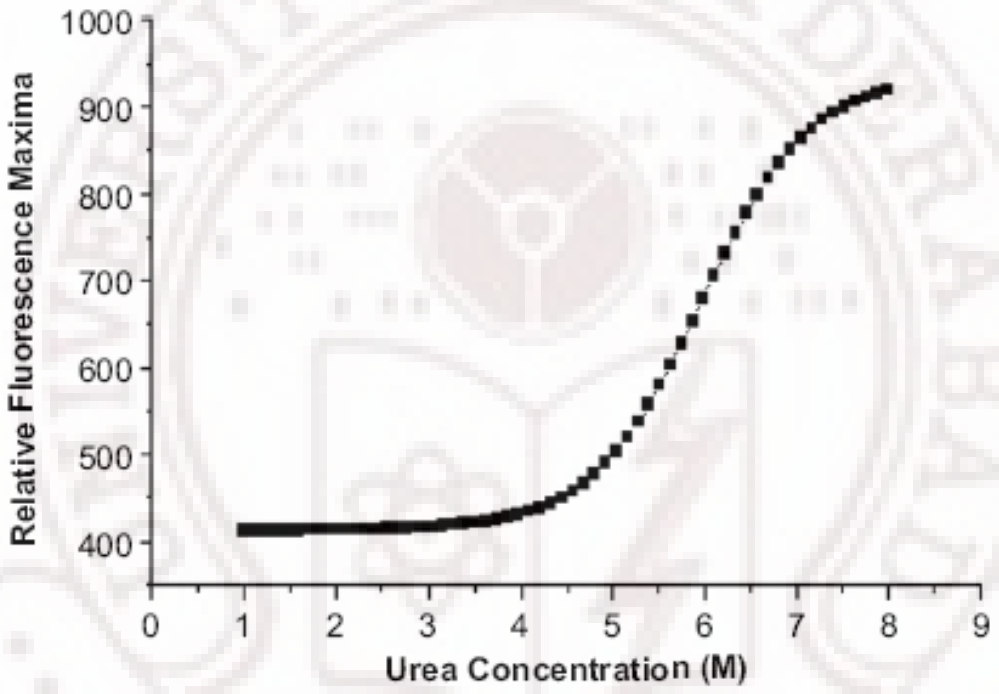


Figure 3.5: Denaturation of recombinant Rv3676 in the presence of urea. Recombinant Rv3676 is completely denatured in the presence of 8 M urea as evident from maximum fluorescence.

This change in secondary structure clearly appears to be a function of increasing concentration of cAMP (Figure 3.4). That cAMP indeed causes concentration-dependent conformational alterations within rRv3676 was further confirmed by tryptophan fluorescence spectrometry.

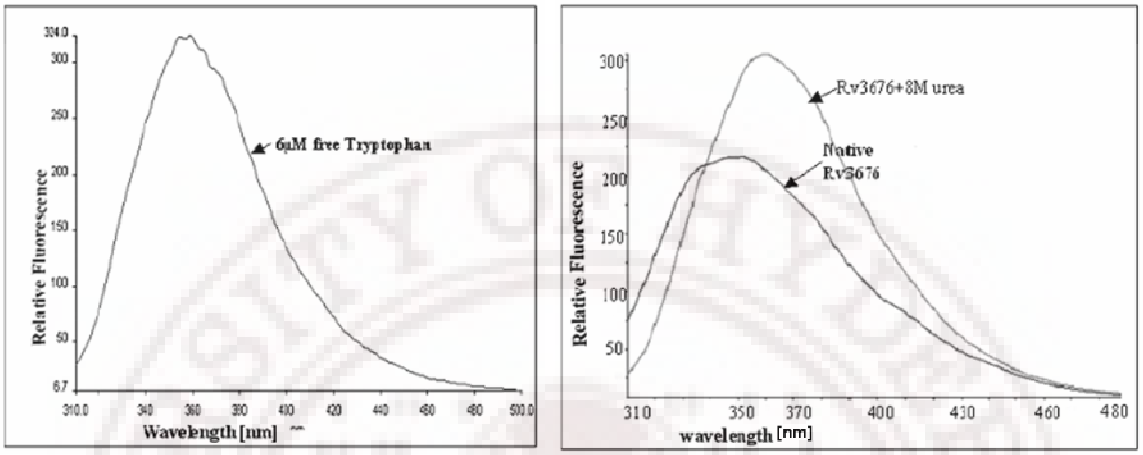


Figure 3.6: Comparison of relative fluorescence intensities contributed by free tryptophan residues and tryptophan residues present in rRv3676 protein. Fluorescence intensities obtained from 6μM of free tryptophan and 3μM rRv3676 (native and unfolded) were compared. Each molecule of rRv3676 contains two tryptophan residues. These comparisons show that at 8 M urea protein is completely unfolded and both the tryptophan residues are fully exposed to the solvent and contribute to equivalent amount of fluorescence intensity as free tryptophan.

The two tryptophan residues (Trp112 and Trp203) present in Rv3676 protein were used as probe to study conformational changes in the protein in solution upon urea-induced denaturation. Purified rRv3676 unfolds completely in the presence of 8 M urea without any further increase in fluorescence (Figure 3.5), indicating the presence of fully unfolded protein molecules. The maximum wavelength of absorbance of denatured rRv3676 is approximately 360 nm, which is equal to the maximum wavelength of absorbance of 6 mM free tryptophan (Figure 3.6). As protein unfolds (relaxed) tryptophan residues are exposed to the solvent, resulting in an increase in relative

fluorescence. Therefore, fluorescence method was used to assay whether increasing concentrations of cAMP have any effect on the unfolding of rRv3676 protein.

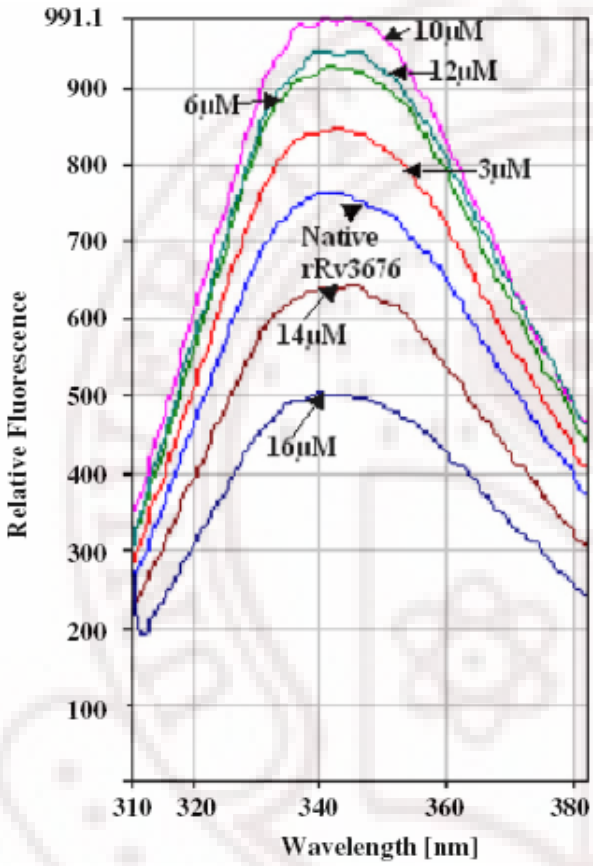


Figure 3.7: Fluorescence emission spectra of rRv3676 as a function of cAMP concentration (6–16 mM). The fluorescence maximum of the protein increases steadily up to 10 mM cAMP and later drops as a function of increasing cAMP concentration (12–16 mM).

PBS was used as solvent, which has a physiological pH and an ionic strength similar to the intracellular milieu of the bacilli. Physiological cAMP levels are in the range of 0–10 mM. At lower concentrations (6–10 mM) the binding shows positive cooperativity, and at 10 mM cAMP the protein is in the most open conformation. This is evident from the increase in tryptophan fluorescence (Figure 3.7). With further increase of cAMP (12–16 mM), the relative tryptophan

fluorescence decreases, suggesting that the protein is getting compacted. This protein compaction could be a reflection of a feedback regulation.

3.3.4 Purified rRv3676 binds *in vitro* to the CRP/FNR cognate nucleotide sequence motif present upstream of *Rv1552*

Having shown that Rv3676 is a likely member of the CRP/FNR family of DNA-binding proteins, purified rRv3676 was tested whether it indeed displays such an activity.

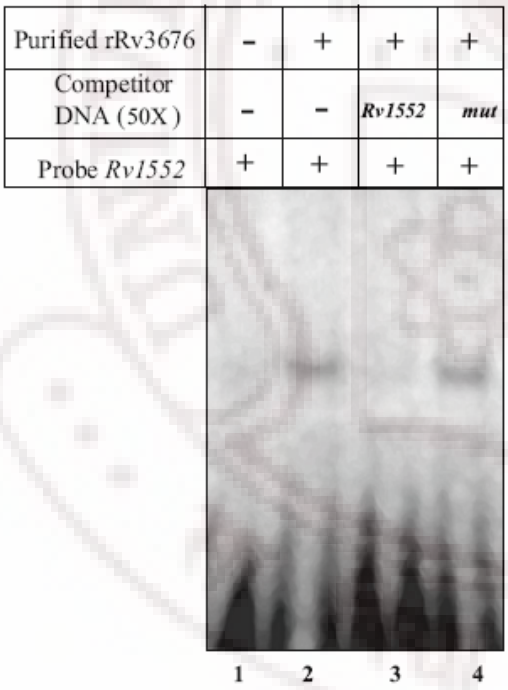


Figure 3.8: Recombinant Rv3676 binds to the CRP/FNR-binding element present upstream of *Rv1552* (*frdA*). Specificity of binding was confirmed by homologous competition with a 50-fold excess of unlabeled ligand (lane 3). The complex is unaffected when a mutant oligonucleotide (mut) carrying an alteration in the binding sequence is used in the competition assay (lane 4).

Mtb ORF *Rv1552* was selected, which is putatively regulated by the CRP/FNR family of transcriptional regulators. Synthetic oligodeoxyribonucleotide corresponding to the CRP/FNR cognate DNA-binding element present in ORF *Rv1552* (*frdA*) was radiolabeled and used

as probe in EMSA using purified rRv3676. The results clearly show a shift in the mobility of the CRP/FNR probe upon incubation with rRv3676 protein (Figure 3.8, lane 2). The specificity of the binding is evident from the disappearance of the complex in a competition assay using a 50-fold excess of homologous unlabeled CRP/FNR oligonucleotides (Figure 3.8, lane 3). The specificity of this DNA–protein interaction is further evident from the absence of any competition when a mutated version of the oligonucleotide (mut in Figure 3.8, lane 4) is used. These results demonstrate that rRv3676 indeed specifically interacts with the CRP/ FNR cognate DNA sequence motifs.

3.4 Discussion

Mtb harbors a single member of the CRP/FNR superfamily, i.e. Rv3676. That this gene is important is evident from knockout studies. An Rv3676 knock-out strain is impaired in growth under *in vitro* conditions, in bone marrow-derived macrophages and also in an animal model (Rickman *et al.*, 2005). ORF *Rv3676* was therefore selected for further analyses of its protein product. Purified rRv3676 exists in a single oligomeric state as a homodimer (Figure 3.2) and is active in terms of DNA binding. It is interesting to note that, despite an only weak DNA-binding activity, the interaction is very specific as could be seen from the inability of the mutant oligonucleotide to compete for binding. Most other

proteins of this family are active as dimers (Körner *et al.*, 2003) but, unlike Rv3676, contain metal cations such as iron and nickel as co-factors. Interestingly, Rv3676 does not carry a metal-binding motif, and the absence of any metal co-factor is indeed evident from the spectral features of rRv3676. *Mtb* Rv3676 thus appears to be different from other oxygen-sensing proteins in terms of non-requirement of a metal cationic co-factor. To investigate the ability of rRv3676 to bind to its cognate DNA motif, EMSA was carried out using purified rRv3676 and a radiolabeled oligonucleotide carrying the CRP/FNR-binding site present upstream of the *frd* (*Rv1552*) gene encoding the fumarate reductase enzyme. This binding site was identified as a putative binding site in recent reports (Bai *et al.*, 2005; Spreadbury *et al.*, 2005). In *in silico* regulon prediction studies, this motif elicited the highest score (Chapter 2; Akhter *et al.*, 2008), and therefore was selected for EMSA. It has been reported that Rv3676 senses oxygen (Bai *et al.*, 2005; Spreadbury *et al.*, 2005) indirectly by controlling the expression of genes such as *frd*. Fumarate serves as an alternative electron acceptor in the absence of oxygen, and this is mediated by a membrane-linked fumarate reductase enzyme complex (Lambden and Guest, 1976). The putative CRP/FNR binding site, present upstream of the *frd* operon, was recognized by rRv3676 protein as evident from EMSA.

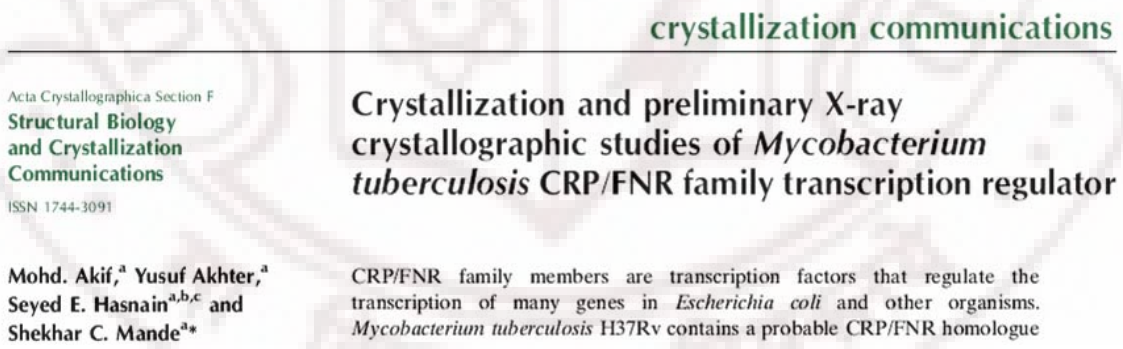
The *Mtb* genome encodes as many as 15 adenylate cyclases, suggesting that cAMP may have an important role in mycobacteria. It has indeed been reported that cAMP can alter the gene expression profile of *Mtb* during anaerobic conditions (Gazdik and McDonough, 2005). The predicted cAMP-binding site in rRv3676 indeed shows binding to cAMP leading to conformational changes in the protein as evident from spectral analyses. The extent of change in secondary structure is maximal in the presence of 10 mM cAMP. The effect of cAMP binding on the DNA-binding efficiency of Rv3676 has already been reported earlier (Bai *et al.*, 2005). cAMP, acting as effector, is known to modulate the regulation of a large number of target genes, and it is likely that Rv3676 is involved in this process. While these *in vitro* findings point to the importance of cAMP, it remains to be experimentally demonstrated whether cAMP is actually involved in regulating gene expression by recruiting Rv3676 protein. While the biophysical features of purified *Mtb* Rv3676 protein described here are physiologically relevant, experimental validation *in vivo* will be required to dissect the complete network of *Mtb* genes regulated by Rv3676 and cAMP.

Chapter 4: Structural studies on *Mtb*-CRP

While a part of the work presented in this Chapter has been communicated as:

Akhter Y, Pogenberg V, Hasnain SE and Wilmanns M. Crystal Structure of cAMP receptor protein from *Mycobacterium tuberculosis* in complex with DNA and cAMP revealed a novel cAMP pocket: implication on DNA binding (Communicated).

A 3-D structure of the apo-CRP including preliminary X-ray crystallography data have also been published as



crystallization communications

Acta Crystallographica Section F
Structural Biology and Crystallization Communications
ISSN 1744-3091

Crystallization and preliminary X-ray crystallographic studies of *Mycobacterium tuberculosis* CRP/FNR family transcription regulator

Mohd. Akif,[a] Yusuf Akhter,[a] Seyed E. Hasnain[a,b,c] and Shekhar C. Mande[a]*

CRP/FNR family members are transcription factors that regulate the transcription of many genes in *Escherichia coli* and other organisms. *Mycobacterium tuberculosis* H37Rv contains a probable CRP/FNR homologue

Kumar P, Joshi DC, Akif M, **Akhter Y**, Hasnain SE, and Mande SC. 2009. Crystal Structure of apo-Cyclic AMP Receptor Protein of *M. tuberculosis* and Normal Mode Analyses reveal an Elegant Mechanism of Allostery induced upon cAMP binding. ***Biophysical Journal*** (In press)

4.1 Introduction

Cyclic AMP receptor protein (CRP) belongs to the CRP/ Fumarate Nitrate Reductase Regulator (FNR) family of transcription regulators. It is well described how different environmental conditions can change the cAMP levels and induce CRP dependent transcription (Körner *et al.*, 2003; Green *et al.*, 2001). cAMP is a central secondary messenger in all living cells. It is currently being investigated how CRP dependent transcription plays a critical role in a successful adaptation of a pathogen to its host (Alspaugh *et al.*, 2002; Caler *et al.*, 2000; D'Souza and Heitman, 2001). CRP/FNR transcriptional regulators are actively involved in response to various kinds of stresses like lower oxygen, redox stress and starvation (Körner *et al.*, 2003).

Previously, in comparative proteomics studies of *Mtb* and *M. bovis* BCG (BCG), differences in electrophoretic mobility of CRP protein were observed (Mattow *et al.*, 2001). Subsequently, such a mobility shift was attributed to point mutations in both the DNA and cAMP binding domains in BCG-CRP. These mutations were implicated to the impaired DNA binding activity of BCG-CRP in comparison to *Mtb*-CRP (Spreadbury *et al.*, 2005) thereby suggesting that these mutations could be one of the contributing factors in attenuation of virulence of BCG.

The mechanism of transcription modulation by CRP is well characterized in *E. coli* and has served as a model for how small

molecules can regulate protein-DNA interactions (Harman, 2001; Tutar, 2008). Structurally, CRP and FNR form homodimers with two distinct domains, an N-terminal effector domain (NTD) and a C-terminal DNA binding domain (DBD). Alteration in the NTD results in change in DNA binding properties. For FNR the regulation is dependent on the redox states of a Fe-S cluster located in NTD while for CRP it depends on the binding of cAMP molecule in the NTD (Spiro, 1994). There is a gap in the understanding of how in the case of *E. coli* CRP the binding of cAMP to the NTD can control DNA binding in the DBD which is ~30Å apart from the binding site of cAMP.

The levels of cAMP act as signal and upon increasing cAMP concentrations, the cAMP-CRP complex is formed. The binary complex can thereafter bind to promoter elements containing variants of the CRP consensus binding motif: TGTGANNNNNNTCACA (Berg and von Hippel, 1988). This binding regulates the transcription of the downstream genes. Therefore, the DNA binding of CRP *via* its C-terminal DBD (residues 138-209) is allosterically regulated by cAMP binding at the NTD (residues 1-137). The NTD predominantly consists of beta sheets and is basically responsible for dimerization and cAMP binding to this domain results in the activation of CRP. Upon activation, CRP recognizes DNA binding elements by the C-terminal helix-F, which forms a prototype Helix-turn-Helix (HTH) motif, together with helix-E (McKay and Steiz, 1981).

CRP from *E. coli* (EC-CRP) was the first transcription regulator protein for which crystal structure was determined (McKay and Steitz, 1981). The crystal structure of EC-CRP has been determined in complex with cAMP (Passener and Steiz,) and in complex with cAMP and DNA (McKay and Steiz, 1981). Recently, the first structural account of apo-EC-CRP was reported using NMR spectroscopy (Popovych *et al.*, 2009). Still the mechanism of regulation by cAMP is not fully understood. As yet there is no single protein from this family for which the crystal structures for all states apo protein, cAMP-CRP (binary complex) and CRP-cAMP-DNA (ternary complex) are known. Structural features of CRP have not been studied extensively in other organisms. Apo-CRP crystal structure from *Mtb* has been reported (Gallagher *et al.*, 2009). Very recently *Mtb*-CRP-cAMP structure has been also reported (Reddy *et al.*, 2009).

How cAMP activates CRP and allows the binary complex to bind the DNA is still a puzzle, and cannot be answered with the present apo and liganded structures of *Mtb*-CRP. In this part of work crystal structure of *Mtb*-CRP in complex with its DNA binding element and cAMP is presented. Interestingly, in addition to the well characterized canonical cAMP binding pockets a third novel cAMP binding pocket was observed in *Mtb*-CRP. Finally, functional implications of this secondary cAMP binding pocket are presented and it was found that it regulates the extent and specificity of DNA interaction.

4.2 Materials and Methods

4.2.1 Bacterial strains and genetic manipulations

In this study *Escherichia coli* strain DH5α from Stratagene and BL21 (DE3) from Novagen were used. For generating *Mtb*-CRP mutant (Asn67Met, Asn137Lys), site-directed mutagenesis kit (from Phusion finzyme) and standard DNA techniques were used as described in the manual.

4.2.2 Purification of *Mtb*-CRP-DNA complex

Mtb-CRP (Rv3676) was expressed and purified as described in the previous Chapter 3. Several single stranded 22-24 base pairs synthetic deoxyribonucleic acid oligos with either blunt end or AT overhangs for crystallization trials were purchased (Metabion International AG, Germany). Single strands of corresponding complementary sequences were annealed as described in Chapter 3. The CRP DNA binding element from upstream of *Rv1552* (*frdA)* was used (Bai *et al.*, 2005; Rickman *et al.*, 2005) as also reported (Chapter 3). For complex formation, suitable DNA to protein ratio was determined by Elecrophoretic Mobility Shift Assay (EMSA), using a constant amount of DNA and an increasing amount of protein. The samples were analyzed by native-PAGE using an 8% TBE gel and stained with ethidium bromide and commassie brilliant blue for analyzing DNA and protein contents, respectively. A DNA to protein ratio of 1 to 3 was used to generate *Mtb*-CRP-DNA complex.

Further purification was performed by size-exclusion chromatography to ensure a highly pure sample using a Superdex 75 16/60 column (Amersham Biosciences) pre-equilibrated in 10 mM Tris-HCl pH 8, 50 mM NaCl. Peak fractions of the protein-DNA stable complex were collected and analyzed by SDS-PAGE and native PAGE prior to crystallization.

4.2.3 Crystallization

Purified *Mtb*-CRP-DNA complex was concentrated to 10 mg/ml and concentration was determined by dye binding method (Bradford, 1976). cAMP (purchased from Sigma) was dissolved in milli Q water. cAMP was added before the crystallization set-up giving a final concentration of 500μM. This purified DNA-protein complex with cAMP was used to initially screen for crystallization conditions using the high-throughput crystallization facility operated by EMBL Hamburg, Germany (Mueller-Dieckmann, 2006). The initial crystallization screens were performed using the sitting-drop vapor-diffusion method with Crystal Screen and Crystal Screen Cryo (Hampton Research, Aliso Viejo, California, USA) based on 288 different conditions. From initial hits the *Mtb*-CRP-cAMP-DNA crystals were manually reproduced at 289 K using the hanging-drop vapor-diffusion method and the crystallization condition was optimized.

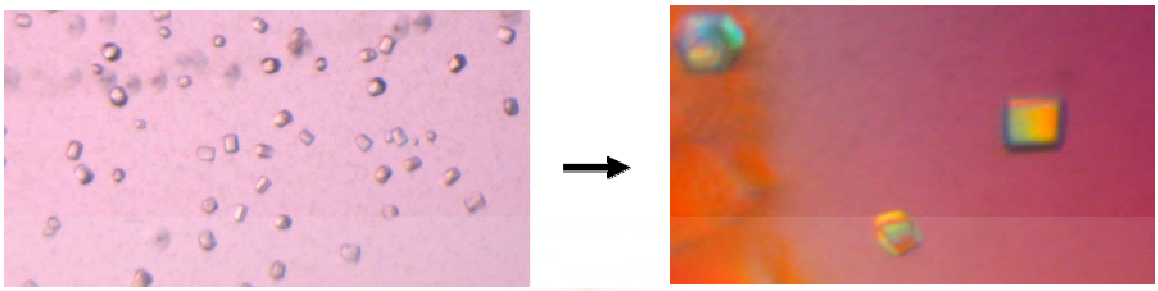


Figure 4.1: Crystals of *Mtb*-CRP-DNA-cAMP ternary complex grew using 0.2 M magnesium chloride, 0.1 M Tris-Cl pH 8.5 and 30 %(w/v) PEG 4000. First panel shows crystals obtained from screens and second panel shows the finally optimized crystals. The size of each optimized crystal is approximately 100X100X150 µm.

The volume of the drop was 4 µl and consisted of a mixture of 2 µl protein solution and 2 µl mother liquor solution. The drop was equilibrated over a reservoir filled with 0.5ml mother liquor (0.2 M magnesium chloride, 0.1 M Tris-Cl pH 8.5 and 30 %(w/v) PEG 4000). Cubic-shaped crystals grew to 100-200 µm in their longest dimension (Figure 4.1).

4.2.4 Data Collection, structure determination and refinement

Prior to data collection, single crystals of *Mtb*-CRP-cAMP-DNA were cryoprotected in 25% glycerol for 30 s and flash-cooled in liquid nitrogen. Diffraction data sets were collected on beamline ID23 at ESRF (Grenoble, France) using an ADSC Q315r (Area Detector Systems Cooperation) CCD detector and a wavelength of 1.0066Å. The EMBL/ESRF/BM14 robotic sample changer (SC3) (Cipriani *et al.*, 2006) at the beamline allowed extensive screening of more than 50 crystals to obtain a crystal was

diffracting beyond 2.9 Å resolutions. The final data were indexed and integrated using the XDS program package (Kabsch, 1993).

The relevant data-collection and processing parameters are given in Table 4.1. The structure was solved using the Molecular Replacement (MR) protocol of the automated structure determination platform Autorickshaw (Panjikar *et al.*, 2005). Within the software pipeline, the Crp/Fnr family protein structure from *Porphyromonas gingivalis* (pdb entry: 2GAU) as top priority model for the MR, which was successful. After MR the electron density allows building of the double stranded DNA in iterative steps of refining and model building using the programs phenix.refine (Adams *et al.*, 2002) and COOT (Emsley and Cowtan, 2004) respectively. TLS (Translation/Libration/Screw) refinement using Chain A and Chain B as TLS groups and simulated annealing strategies as implemented in phenix.refine were essential in the refinement process (Adams *et al.*, 2002).The final refinement statistics achieved are listed in Table 4.1.

Table 4.1 Data collection and refinement statistics

Crystal	*Mtb* CRP–DNA-cAMP
Space group	$P2_1$
a, b, c (Å)	65.74 60.82 89.64
α, β, γ (°)	90.00 104.33 90.00
Solvent content (%, v/v)	55.75
A. Data collection	
Resolution (Å)	50.0–2.90 (2.98–2.90)[a]
Unique reflections	14777 (1088)
Multiplicity	2.28 (2.30)
I/σ(I)	7.66 (2.2)
Completeness (%)	95.4 (97.5)
R-merge (%)	17.3 (49.2)
B. Refinement	
Refinement resolution	19.74–2.90
R_{cryst} (%)	25.20
Rfree[b](%)	28.00
C. No of atoms in asu	4254
Protein	3368
DNA	840
cAMP	66
Average B-factor (Å2)	
Protein	59.77
DNA	65.57
cAMP	70.96
r.m.s. deviations from ideal	
Bond lengths (Å)	0.011
Bond angles (°)	1.6
Ramachandran plot	
Most favored regions (%)	87.4
Allowed regions (%)	12.3
Generously allowed (%)	0.3
Disallowed regions (%)	0.0

[a] Values for highest resolution shell are written in parentheses.
[b] Rfree was calculated using 6.54% of data, which was omitted from the refinement

4.2.5 Isothermal titration calorimetry

Recombinant *Mtb*-CRP, the DNA duplex and cAMP stock solutions were prepared as described above. All three components were dialyzed overnight against 10 mM tris-HCl, pH 8.0, 50 mM NaCl. During this experiment same buffer was used for dilution purposes. ITC measurements were carried out in duplicate on a MicroCal VP-ITC instrument at 25°C. DNA (80 µM) was injected in steps of 10 µl into 1.4 ml of *Mtb*-CRP (16 µM) or *Mtb*-CRP Mutant (Asn67Met, Asn137Lys) (16 µM) with and without pre-incubation with cAMP (500µM). The kcal/mole

per injection was fitted by a one-step reaction using the Origin software (MicroCal). Data were fitted by a one to one interaction model using the Origin software 7 (MicroCal), after substracting -1.10 kcal/mol corresponding to the average of the control experiment, namely titrating DNA into buffer.

4.3 Results and discussion

4.3.1 *Mtb*-CRP ternary Complex formation

The *Mtb*-CRP-DNA complex was obtained by mixing recombinant purified 6XHis tagged apo-*Mtb*-CRP with *frd* DNA motif present upstream of *frd* (fumarate dehydroenase*)* operon. This DNA element was able to bind *Mtb*-CRP specifically as reported earlier (Chapter 3; Bai *et al.*, 2005; Rickman *et al.*, 2005). It has been reported that *Mtb*-CRP activate the expression of genes like *frd* which helps it to survive the hypoxic environment in macrophages (Bai *et al.*, 2005; Spreadbury *et al.*, 2005) Fumarate could serve as an alternative electron acceptor in the absence of oxygen (hypoxia), and this is mediated by a membrane-linked fumarate reductase enzyme complex (Lambden and Guest, 1976).

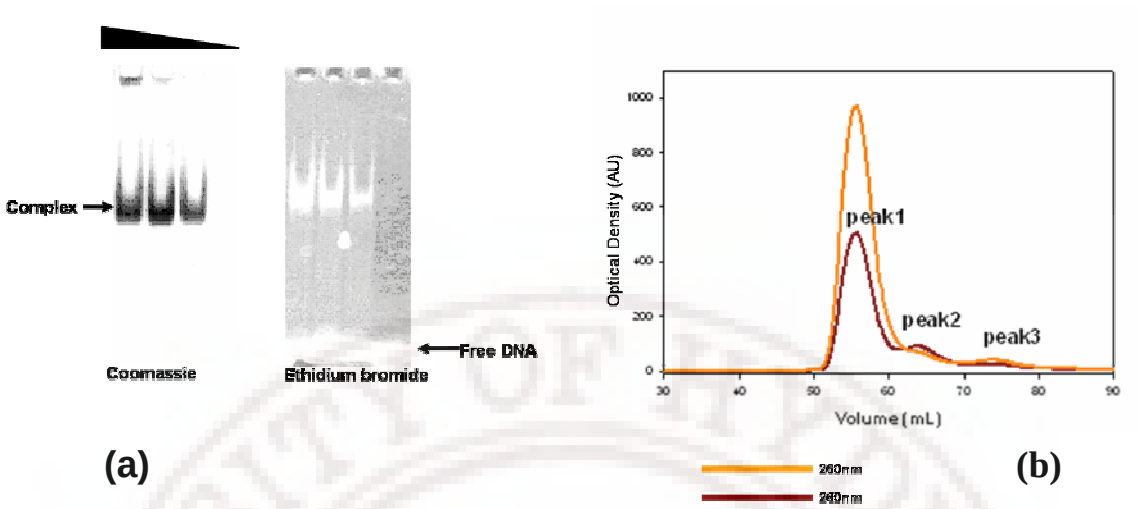


Figure 4.2: Purification of *Mtb*-CRP-DNA complex. (a) DNA:protein stoichiometry suitable for complex formation was determined by Elecrophoretic Mobility Shift Assay (EMSA) . *Mtb*-CRP-DNA complex was determined by mixing constant amount of DNA mixing with increasing amounts of protein. The samples were analyzed by native-PAGE using an 8% TBE gel and stained with ethidium bromide and commassie brilliant blue for analyzing DNA and protein respectively. (b) Size exclusion chromatography was performed Superdex 75 16/60 (Amersham Biosciences) column pre-equilibrated in 10 mM Tris-HCl pH 8, 50 mM NaCl. the eluates were monitored at 280nm and 260nm indicated in brown and yellow, respectively. Chromatogram showing three peaks labeled as peak 1, peak 2 and peak 3. Peak 1 is a major eluate peak corresponding to the protein-DNA complex, peak 2 corresponds to the non-complexed protein and free DNA elutes as peak 3, as interpreted by monitoring absorbance at 280 and 260.

The suitable stoichiometry for large scale protein-DNA complex purification was determined by native-PAGE and visualized by dual staining method using commassie brilliant blue and ethidium bromide for protein and for the DNA respectively as described in Materials and Methods (Figure 4.2). The elution profile shows three peaks. The main peak, labeled as “peak 1”, contains the protein-DNA complex (*Mtb*-CRP-DNA-cAMP), while the second peak contains pure apo-*Mtb*-CRP and the third peak pure DNA (Figure 4.2). Elution fractions from peak 1 were analyzed on a native-PAGE to reconfirm the integrity of the protein-DNA

complex. Fractions constituting peak 1 were pooled and used for further crystallization trials.

4.3.2 Structure determination and refinement of the *Mtb*-CRP-DNA-cAMP complex

The structure of *Mtb*-CRP-DNA-cAMP was solved in the $P2_1$ spacegroup at a 2.9Å resolution using the CRP/FNR family protein structure (pdb entry: 2GAU) from *Porphyromonas gingivalis* as MR search model, The model was refined at 2.9Å to R_{cryst}/R_{free} of 0.25/0.28.The refined model of protein/DNA complex contains 224 residues of the protein including an additional N-terminal histidine from the 6XHis-tag, 23 basepairs of DNA and three cAMP molecules. The overall structure shows acceptable geometry with all residues in the allowed regions of Ramachandran plot (Table. 4.1). The homodimeric *Mtb*-CRP-DNA complex shows similar structural features to that of EC-CRP (Chen *et al.*, 2001) (Figure 4.3).

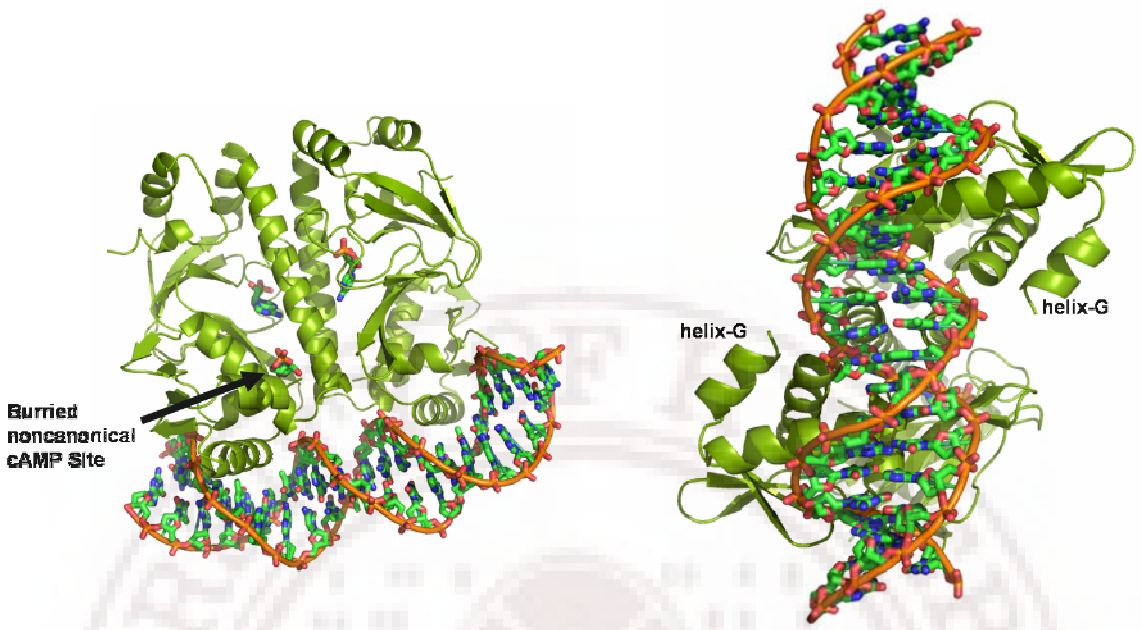


Figure 4.3: Overall structure of *Mtb*-CRP-cAMP-DNA. Overall secondary and tertiary structure is similar to EC-CRP structures reported earlier (pdb entries: 1ZRC and 1O3S). *Mtb*-CRP homodimer interact with 23 basepair DNA. Only novel features are labeled in *Mtb*-CRP-DNA complex structure. Arrow shows non-canonical cAMP binding site containing cAMP molecule. The helix G is the additional helix not present in EC-CRP but present in the *Mtb*-CRP structure.

The two CRP molecules within the complex have structurally similar architecture and open conformations as seen for the EC-CRP-DNA structure with a RMSD of 2.5 Å (comparison with EC-CRP-DNA complex structure, pdb code 1O3S). Each *Mtb*-CRP molecule contains six helices (denoted A to F, according to EC-CRP) and 11 beta-strands. However, in comparison to EC-CRP, *Mtb*-CRP contains a short helix at the very beginning of C-terminus (Ser215 to Arg224), which is not present in EC-CRP. Structure based sequence alignment reveals that the prolonged C-terminus is likely to be unique to myobacteria.

After building the complete model of the protein-DNA complex, unambiguous density calculations allowed the positioning of the three cAMP molecules. The positive electron density (mfo-nfc) allowed positioning of three cAMPs molecules unambiguously. The *Mtb*-CRP-DNA model was refined in the absence of cAMP molecules to produce the Fo-Fc map. This map is unbiased as the atomic coordinates used for phase calculations have never been refined together with the cAMP molecules. The cAMP molecules were finally inserted to structure the final coordinates after some further refinement. Two of these were located at Non Crystallographic Symmetry (NCS) positions consistent with the cAMP binding sites of EC-CRP (Weber and Steitz, 1987). Density supporting a third cAMP molecule was observed in subunit B facing the helixes C, D and E (Figure 4.3, more below). The two cAMP molecules are fully bound into the conserved binding pockets whereas the third cAMP is partially (50%) occupied. This suggests that binding of cAMP to the secondary binding pocket displays lower affinity. All the three cAMP molecules were found in anti-conformation (Figure 4.3) state.

4.3.3 C-terminal helix-G

In both of the subunits of the *Mtb*-CRP-DNA structure it was observed that the very C-terminus residues forms an alpha-helix. In the apo-*Mtb*-CRP structure only one of the subunits have a structured C-terminus whereas the second subunit has a completely disordered C-

terminus. The differences in the two C-termini of the apo-*Mtb*-CRP complex likely explain why this homodimer is surprisingly asymmetric (Gallagher *et al.*, 2009).

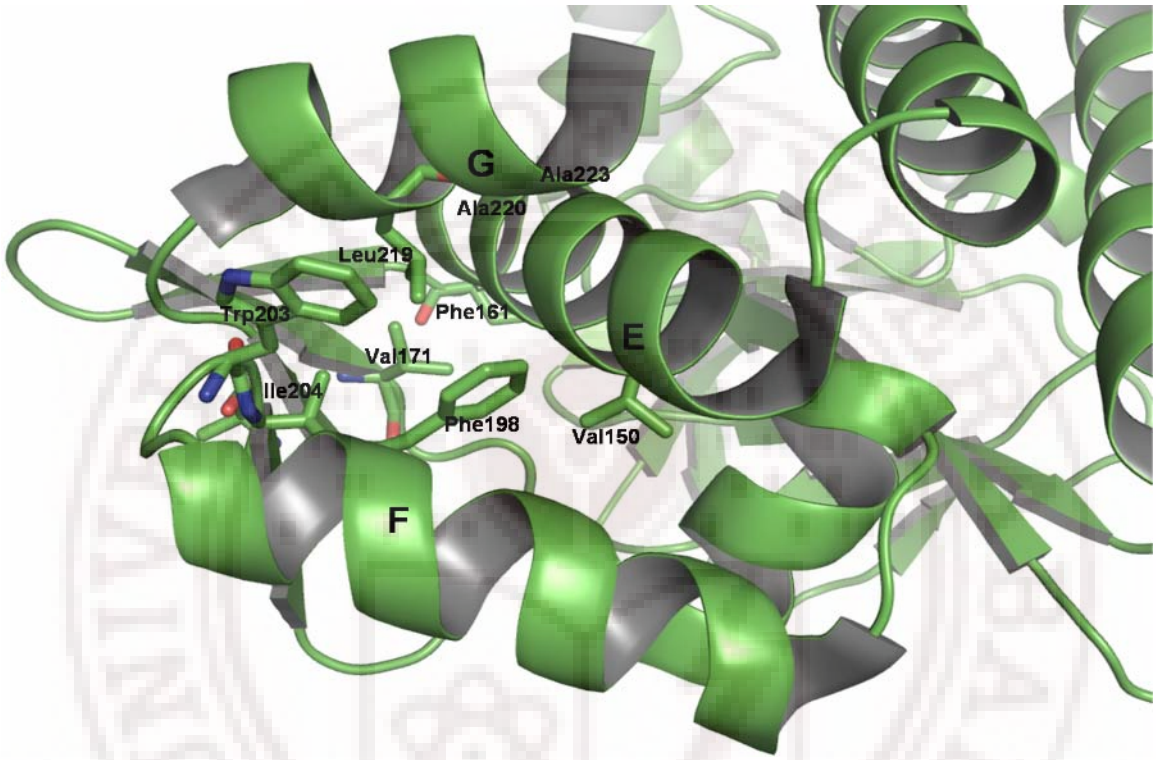


Figure 4.4: The role of helix-G: a flap for covering hydrophobic core and to prevent solvent access: (a) Hydrophobic core formed between DNA interacting helix-F and helix-E and helix-G is serving as a flap to prevent any solvent access to it. Hydrophobic side chains are shown in sticks.

Therefore it was investigated that if this terminal could play a functional role in the ability of CRP to form complex. Deletion of the C-terminal part of the protein (residues 214-224) leads to an insoluble form. This indicated that the C-terminus is involved in solubilization and correct folding of the protein. These findings could be further explained by the environment surrounding the C-terminal helix-G. Residues from the helix-G which could function as a flap to cover highly hydrophobic

surface formed at the interface of helices E and F and lead to formation of a unique compact sub-domain with a hydrophobic core (Figure 4.4).

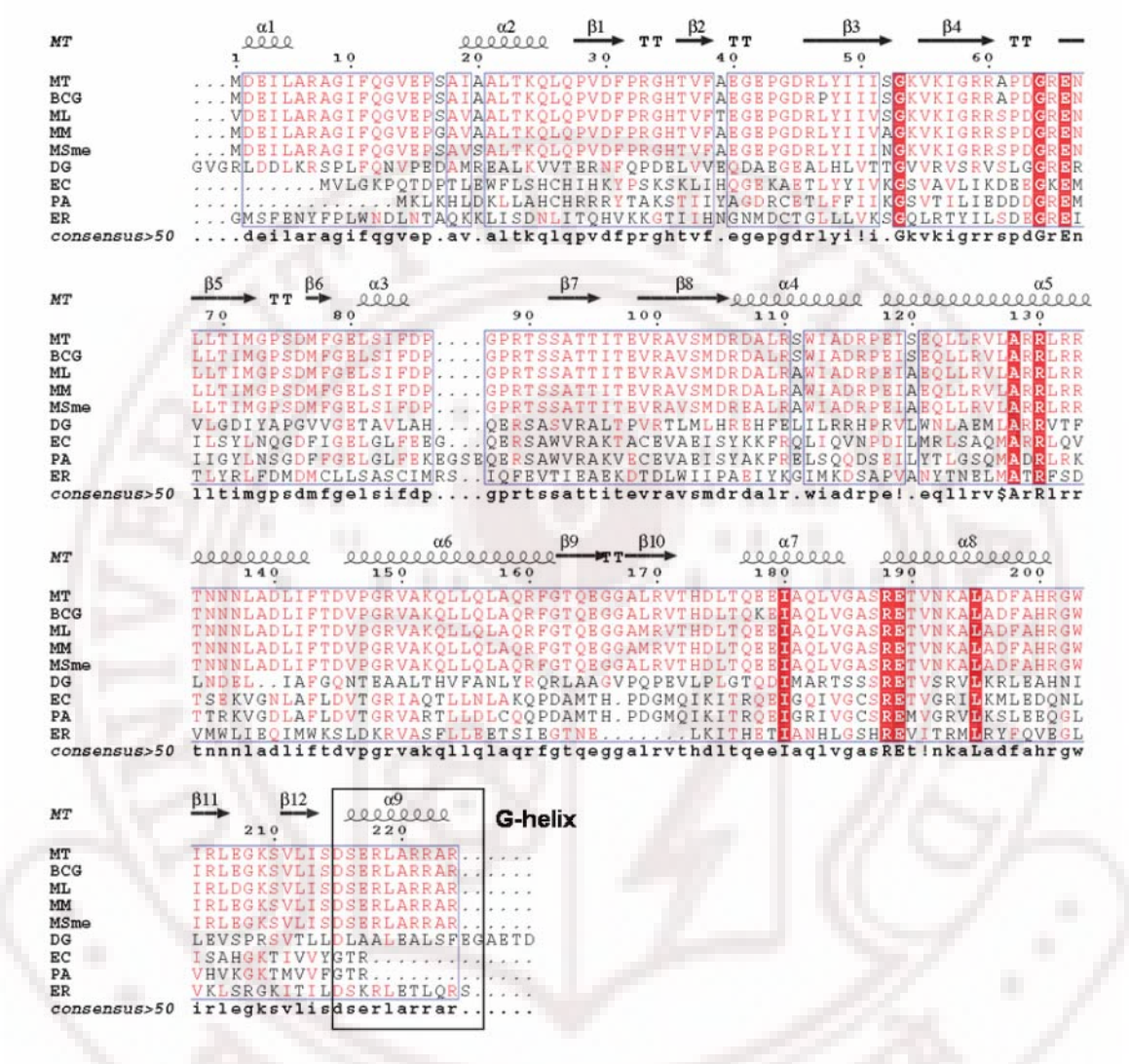


Figure 4.5: Multiple sequence alignment of CRP from different species. Helix-G is a specific feature of mycobacterial CRPs. An alignment of array of CRP aminocid sequences from different bacterial species clearly showed that helix-G is conserved in mycobacterial CRPs while in other bacterial species it is either truncated or completely absent. Abbreviations used are MT, *Mtb*; BCG, *M.bovis* BCG; ML, *M.leprae*; MM, *M.marinum*; DG, *Deinococcus geothermalis*; EC, *E. coli*; PA, *Pseudomonas aeruginosa* and ER, *Eubacterium rectale.*

Indeed, three residues (namely Leu219, Ala220 and Ala223) from helix-G are present which are in close hydrophobic contact with Trp203, Ile204,

Phe161, Val171, and Phe198and Val150. All of these hydrophobic side chains are facing the centre of this sub-domain.

Interestingly, there is no existence of such a hydrophobic core in the homologous structures. For example in EC-CRP Phe161 (hydrophobic, aromatic) is replaced by Pro (less hydrophobic, aliphatic), Val171 is replaced by Met, Phe198 (highly hydrophobic) is replaced by Leu (less hydrophobic) and Trp203 is replaced by Leu (less hydrophobic). Multiple alignment clearly showed that this helix is only conserved in mycobacterium related species and in others it is either absent or not conserved (Figure 4.5). A bigger amino acid sequence alignment confirmed these observations for other homologues also. It could be seen that the residues corresponding to helix-G are not present in other homologous protein nor the hydrophobic core (Figure 4.5).

4.3.4 Novel cAMP binding site: structural features

The finding of a third cAMP molecule was unexpected, although secondary cAMP binding sites in EC-CRP have been reported earlier (Passner and Steitz, 1997), the results presented in this Chapter are different (Figure 4.6). In EC-CRP the site is surface exposed and located near the DNA protein interaction area interacting with Arg180 (located on DNA interacting helix-F), Glu58 (in nearby beta-sheet) and Gly173 and Gly177 (both on helix-E) (Passner and Steitz, 1997; pdb entry: 2CGP). In

Mtb-CRP the secondary pocket is buried and located at the interface of three helices: helix-D (dimer interface helix), helix-E and helix-F (constituent of DNA interacting helix-turn-helix).

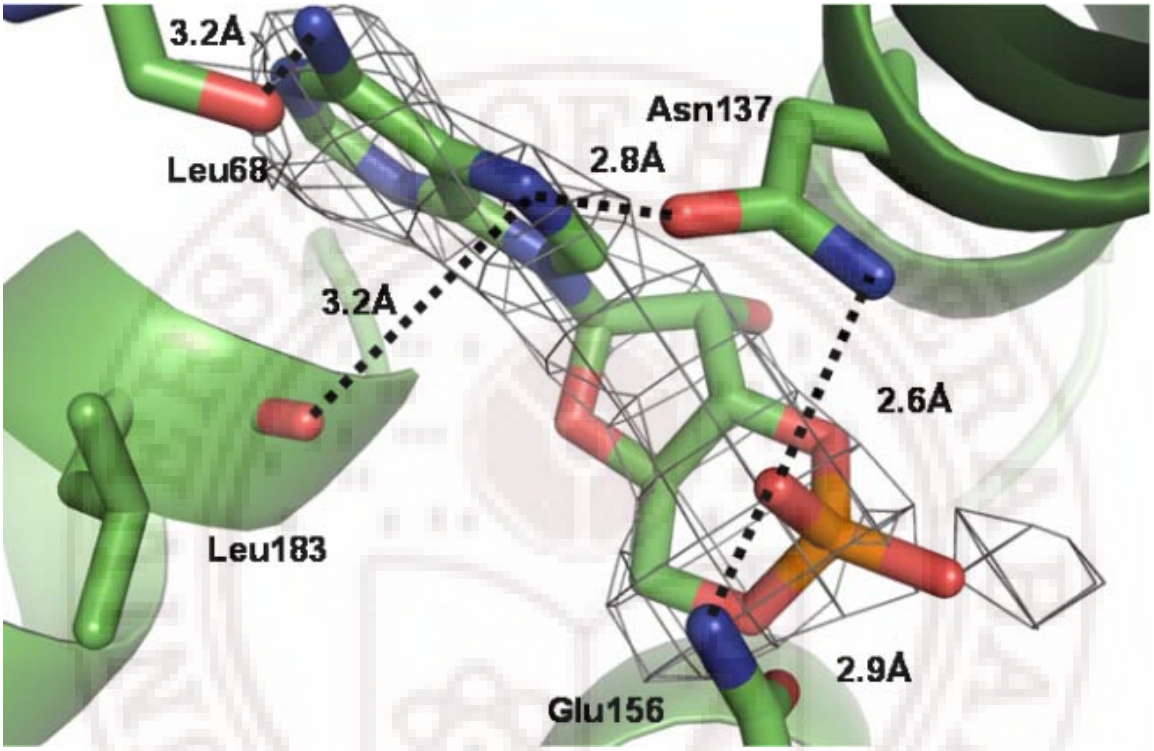


Figure 4.6: The binding of a cAMP to *Mtb*-CRP in non-canonical pocket. The picture shows the mode of binding of cAMP in its omit-density of an (Fo-Fc)αc map. The *Mtb*-CRP-DNA structure, after omitting the cAMP molecule (contoured at 2.0 sigma). Upon the maps have been calculated after omit refinement, leaving out the highlighted molecule from the model. Different side chains including the distances in angstrom which are interacting to cAMP are labeled.

cAMP directly interacts with six of the side chains of residues Asn67, Leu 68, Asn137, Asp140, Glu156 and Leu156 surrounding the pocket *via* several H-bonds and salt bridges. When the least square superimposition analysis with *Mtb*-CRP-cAMP structure (PDB code: 3i54)

was performed it was not found to be much different than *Mtb*-CRP-DNA structure (RMSD=1.16 Å). This pocket environment was found to be conserved including the side-chains interacting with cAMP. For unknown reasons there was no evidence of binding of secondary cAMP molecule to the *Mtb*-CRP-cAMP complex structure secondary cAMP molecule was not bound. In case of EC-CRP it is known that interaction with DNA could give rise to secondary cAMP binding (Lin *et al.*, 2002, Scott and Jarjous, 2005). This might be similar in case of *Mtb*-CRP. This position seems to be interesting in terms of DNA binding regulatory function.

For this non-canonical pocket, cAMP molecule could be built only in sub-unit B of *Mtb*-CRP with 50% occupancy. This indicates that this is a secondary binding site and is not occupied fully in all molecules. On the other hand some residual density on the similar NCS related position in sub-unit A was also observed, although it was not sufficient to clearly position in the cAMP. Taken together, this indicates that there could be concentration dependent sequential multi-step binding events for each sub-unit. Second non-canonical pocket has probably a lower affinity than canonical one as it only has 50% occupancy. This observation is consistent with biophysical and biochemical studies for EC-CRP (Heyduk and Lee, 1989; Leu *et al.*, 1999; Takahashi *et al.*, 1980). In surface-potential analysis a cleft for entry of cAMP to the pocket could be seen. The surface around this cleft was found positively charged (Figure 4.7).

This positive charge could facilitate the entry of cAMP to the non canonical pocket.

Figure 4.7: Cleft for entry of secondary cAMP molecule into non-canonical pocket. Surface potential representation shows a cleft for ligand entry to secondary site. Positively charged surface represented in blue while negatively charged in red colors. Positively charged surface patches around the cleft may facilitate entry of cAMP molecule to the pocket.

Upon comparison with the EC-CRP-DNA structure (pdb: 1O3S) a different position of the helix-D (DBD) and helix-F (DNA interacting) of more than 3 Å was observed. This could be due to the presence of the secondary binding pocket. This movement in helix-D translated as a motion in helix-F which finally gives rise to more interaction with DNA (Figure 4.8). A careful analysis revealed that the helix rearrangement could be mediated by H-bonding between phosphate of cAMP and side-chain amino groups from Asn137 and Gln156 (Figure 4.8).

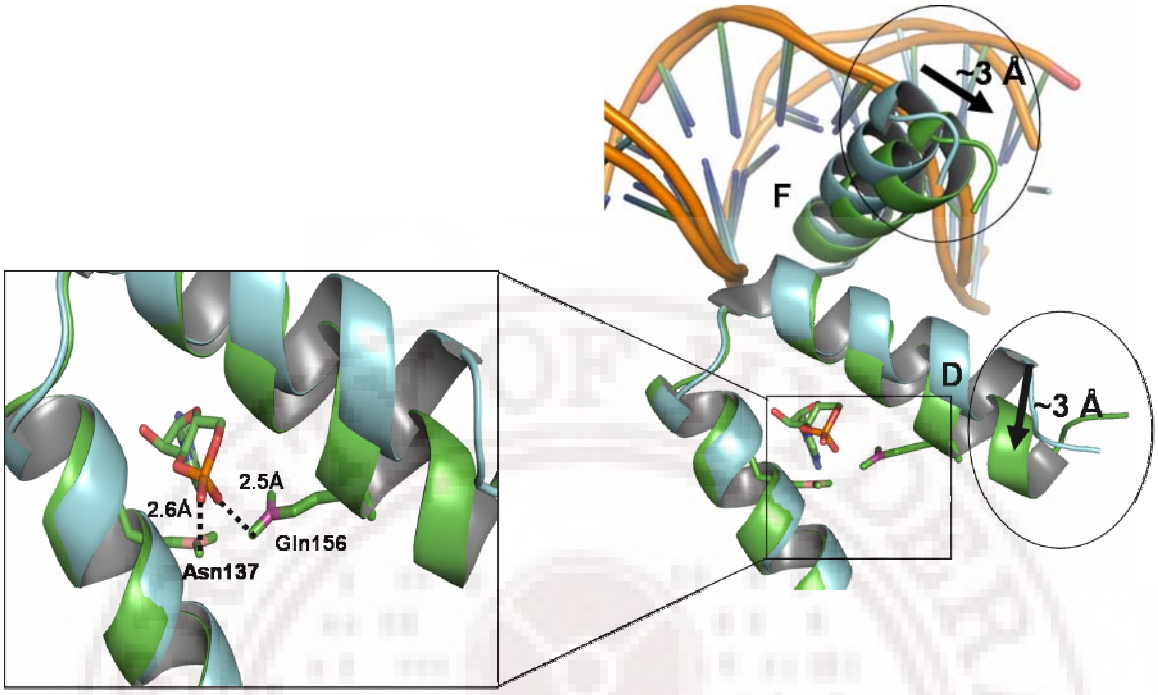


Figure 4.8: Re-arrangement of helices as implication of cAMP binding to non-canonical site. Least square superposition of EC-CRP with *Mtb*-CRP revealed movement of helices, DNA interacting helix-F and helix-D. There are 3 Å of movement in both of the helices as indicated by the arrows and site of movement is encircled. EC-CRP is shown in cyan while *Mtb*-CRP in green color cartoon representation using pymol.

When least square superimposition of the structure of *Mtb*-CRP-DNA with EC-CRP (PDB entry: 1O3S) was performed it was found that the side chains of Met 59 and Lys129 occupy the binding pocket of CRP in *E. coli*, whereas the corresponding residues, Asn67 and Asn137 have lighter side chains in *Mtb*-CRP and are positioned to allow cAMP binding or to specifically interact with the bound cAMP molecule (Figure 4.9).

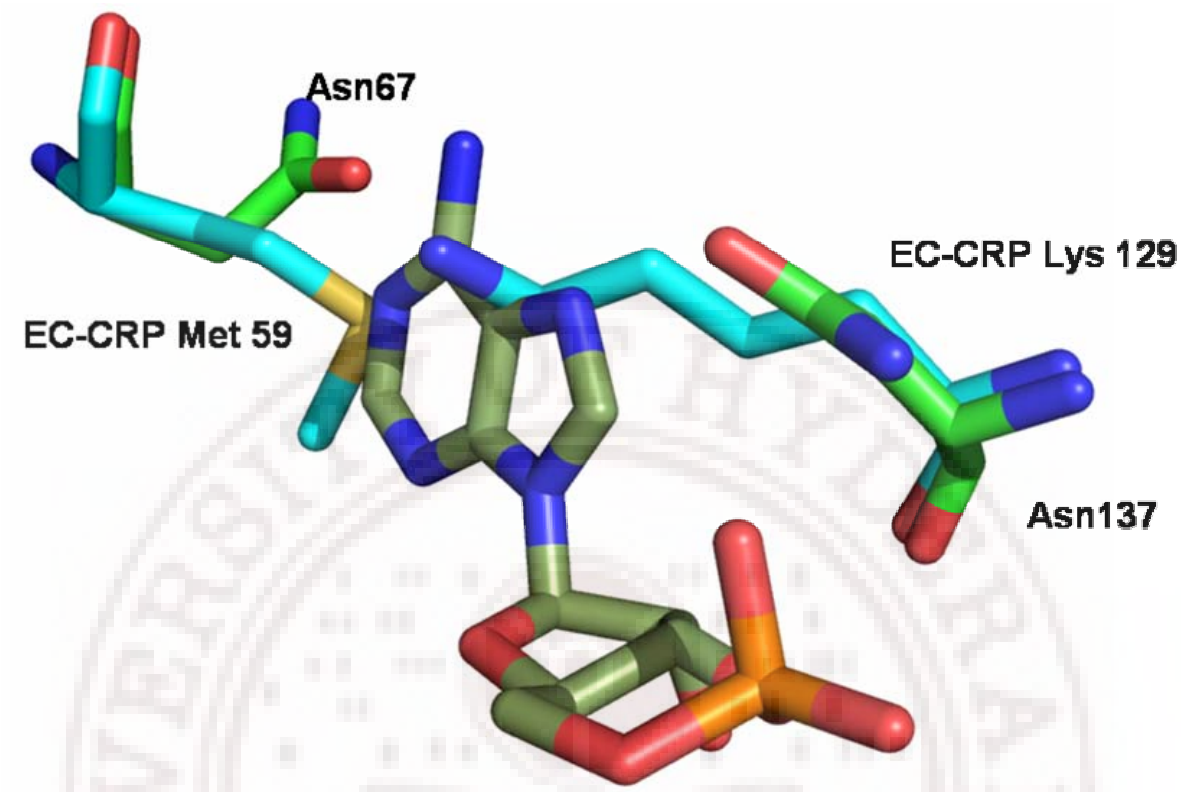


Figure 4.9: Comparison of non canonical cAMP binding pocket of *Mtb*-CRP with non pocket environment of EC-CRP in superposition of both proteins. Bulkier side chains (Met59, Lys12) from EC-CRP showed in cyan create steric hindrance and make non-pocket environment while equivalent residues in *Mtb*-CRP have been replaced by smaller side chains (Asn67, Asn137) showed in green, making room for cAMP pocket.

Similar comparative analyses for the position of DNA interacting helix-F in *Mtb*-CRP-cAMP (PDB code: 3i54) were also performed but no major changes in its position could be seen. This might suggest that *Mtb*-CRP-cAMP form reported earlier, in which only one cAMP is occupied in each sub-unit, could also be the active form of *Mtb*-CRP ready to bind DNA and could start activation of transcription (Reddy *et al.*, 2009). It may be noted that secondary cAMP is not bound to *Mtb*-CRP in this structure. In this case seemingly, second cAMP binding could further

regulate the transcription in a feed forward type of mechanism. These observations indicate that the third binding pocket is specific to cAMP, but not conserved among all bacteria. Sequence alignment with other mycobacterial CRPs showed that these aminoacids are conserved in all of them (Figure 4.5).

4.3.5 Secondary cAMP pocket: functional implications

To investigate the importance of this secondary pocket *Mtb*-CRP mutants were designed, which can mimic the non-pocket environment found in EC-CRP. The two side chains, Asn67 and Asn137 from *Mtb*-CRP were replaced with the corresponding side chains Met67 and Lys137 from EC-CRP (Figure 4.9). The generated mutant was expected not to bind cAMP in this pocket as the larger residues are expected to sterically hinder the binding of cAMP as seen in EC-CRP. The mutant allowed the investigation of the effect of the secondary binding pocket on the ability of *Mtb*-CRP to bind the DNA. The role of secondary cAMP binding pocket on modulating DNA binding events was then evaluated.

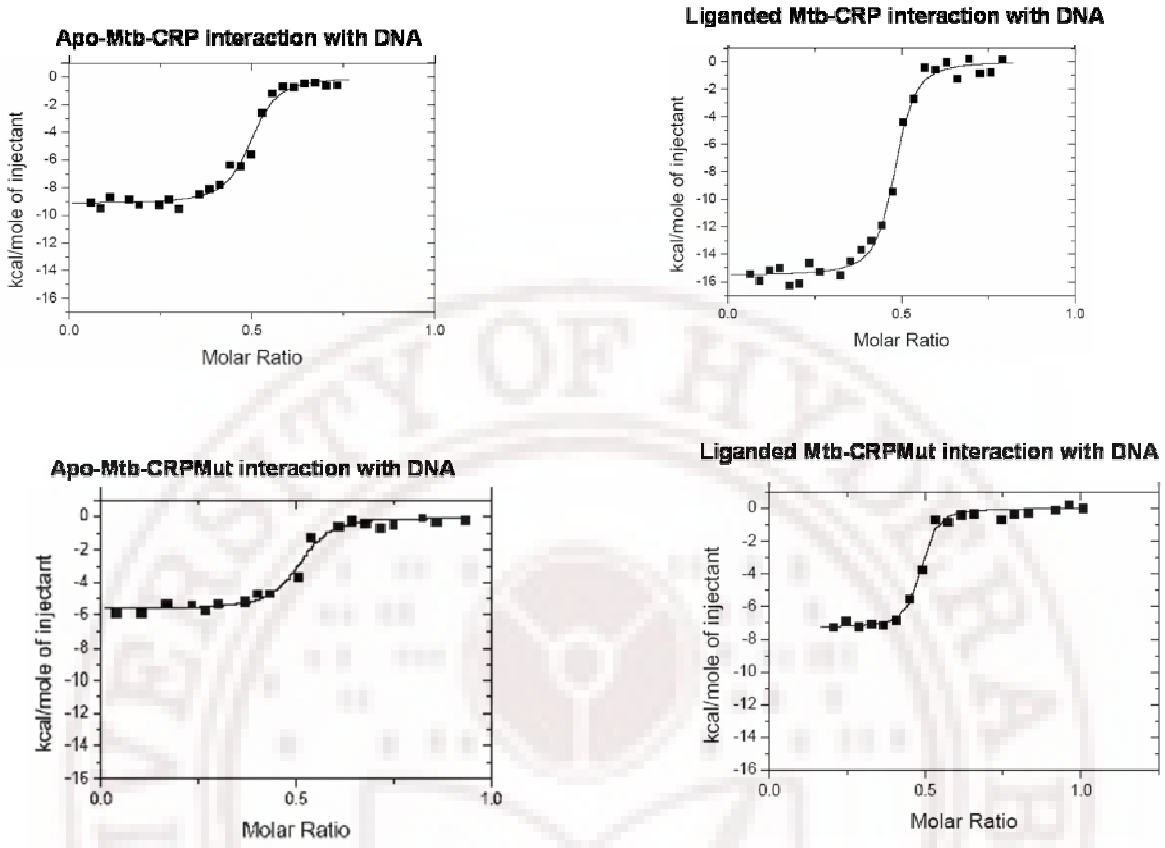


Figure 4.10: DNA-protein interaction in presence and absence of cAMP. DNA binding curves for the double stranded *frd*-DNA element (23 bp) and *Mtb*-CRP/ *Mtb*-CRPMut accumulated by isothermal titration microcalorimetric analysis showing importance of novel secondary cAMP binding pocket in *Mtb*-CRP. Trace of the calorimetric titration of 30 X 2.5-µl aliquots of 463 µM DNA injected into 1.8 ml of 15.8 µM *Mtb*-CRP or *Mtb*-CRPMut in liganded and unliganded forms as described in Material and Method. Solid lines in the lower plot represent the best fit to the data given by the parameters in Table 4.2.

Four parallel sets of experiments for determining DNA binding curves in different thermodynamic events were performed :1) Wild type *Mtb*-CRP and the DNA motif; 2) Wild type *Mtb*-CRP saturated with 500uM cAMP prior to addition of DNA; 3) *Mtb*-CRPMut with DNA; and finally 4) *Mtb*-CRPMut saturated with cAMP (500µM). To make sure that cAMP would occupy both binding sites of each monomer 500uM of cAMP was used.

Table 4.2: Thermodynamic parameters for DNA and *Mtb*-CRP/*Mtb*-CRPMut interactions calculated from isothermal microcalorimetric titrations

	CRP	**CRP+cAMP**	**CRPMut**	**CRPMut+cAMP**
ΔH (kcal/mol)	9.15±0.16	15.53±0.2	5.6±0.143	7.27±0.156
ΔS (cal/mol/K)	4.85	-15	15.6	12.1
Kd (nM)	16.9±5.9	0.16±0.03	30.2±7.5	10.6±2.8
No of sites	0.48±0.004	0.47±0.003	0.49±0.007	0.47±0.004

The Kd values for the DNA binding events for *Mtb*-CRP with and without cAMP saturated states show the positive effect on DNA binding efficiency. The dissociation constant (Kd) dropped by almost 100 fold on addition of cAMP (from 16nM to 0.16nM). When a parallel experiment with *Mtb*-CRPMut was carried out, devoid of novel non-canonical pocket, the massive drop in Kd was not observed, and just ~3 fold (from 30nM to 10nM) drop was seen (Figure 4.10, Table 4.2). DNA binding efficiency of mutant protein did not increase as much as that of wild type. Interestingly, in earlier reports EC-CRP was reported to have very poor DNA binding when it was treated with higher concentrations of cAMP (Harman, 2001, Adhya *et al.*, 1995). This clearly indicates that the effects of cAMP binding on DNA interactions are different in EC-CRP and *Mtb*-CRP and novel non-canonical cAMP pocket have a critical role in regulation of DNA binding by *Mtb*-CRP.

Further results showed that energetics of DNA binding are also very distinct in *Mtb*-CRP. The change in enthalpy for cAMP liganded-*Mtb*-CRP (15.5 kcal/mol) on DNA interaction is rather higher than the energy

released in unliganded-*Mtb*-CRP (9.15 kcal/mol) DNA interaction (Figure 4.10, Table 4.2). This high energy release is also compensated by high amount of change in entropy (-15cal/mol/K) in liganded interaction. Entropy change difference between liganded and unliganded interactions is almost 20cal/mol/K. When a similar experiment was carried out with *Mtb*-CRPMut no such big difference in release of energy was observed. High amount of enthalpy changes could be reflection of the specificity of a binding reaction. According to present accepted model for EC-CRP, binding of first cAMP favors the binding of cAMP molecules at secondary cAMP sites. Binding of extra cAMPs to the secondary binding site decreases its specificity for DNA (Lin *et al.*, 2002; Scott and Jarjous, 2005). It should be noted that the secondary cAMP binding site in *Mtb*-CRP is at different location to that EC-CRP. It is therefore likely that the two binding site have different mechanism. These results clearly show that cAMP binding to secondary sites increases the specificity of CRP for DNA instead of decreasing its specificity.

4.3.6 *Mtb*-CRP and DNA interactions

Currently, the only CRP-DNA complex structures available are from *E.coli.* Therefore, a comparative analysis of the DNA-protein interactions between EC-CRP and *Mtb*-CRP was carried out. A predicted helix-turn-helix (176-LTQEEIAQLVGASRETVNKALA-196) motif which is

involved in DNA binding for *Mtb*-CRP was reported in the earlier Chapter 2.

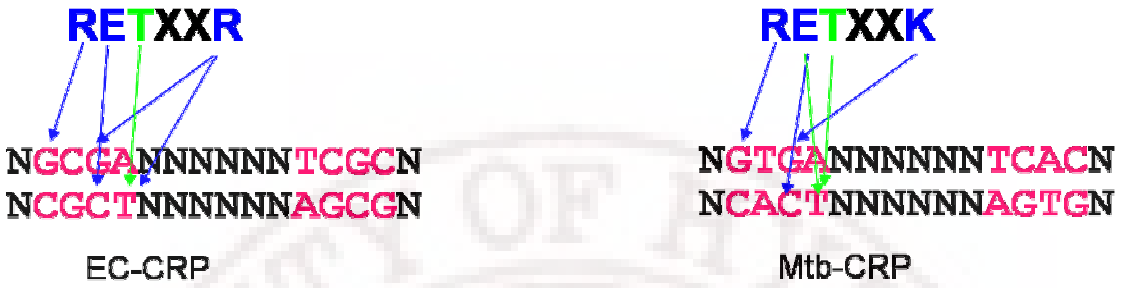


Figure 4.11: Minimal signature motifs of DNA and protein required for interaction. Minimal DNA and amino acid residues for both EC-CRP and *Mtb*-CRP were calculated for comparative analysis. Electrostatic base specific interactions are shown with blue lines while hydrophobic base specific interactions are shown with green lines.

Here, with the help of DNA-protein complex crystal structure was further narrowed down to exact minimal signature motif responsible for specific interaction. 1ZRC and 1O3S pdbs (EC-CRP-DNA complex structures) were used to identify the amino acid residues. Minimal signature nucleic acid sequence required for specific interaction with protein for both EC-CRP and *Mtb*-CRP were also obtained. In EC-CRP, Arg-Glu-thr-Xaa-Xaa-Arg (residues180-185) are critical while for *Mtb*-CRP the Arg-Glu-thr-Xaa-Xaa-Lys are involved (residues188-193) (Figure 4.11). At the last position in case of *Mtb*-CRP the lysine residue is substituted by arginine. The DNA motif for the EC-CRP is “GCGA” while for *Mtb*-CRP it is “GTGA”. The overall pattern of interactions was found to be similar except for two interactions: the lysine in *Mtb*-CRP could interact only to one base specifically while equivalent arginine in EC-CRP could make one more interaction with thymine. On the other hand glutamate in *Mtb*-CRP

signature motif could interact with thymine. It thus appears that the overall interactions are conserved and DNA and aminoacid residues for both EC-CRP and *Mtb*-CRP could be documented.

Except these base specific interactions some non-specific DNA backbone related H-bonds were also observed. Val146, Asn192 and Arg65 could interact with DNA backbone *via* H-bonds. Surprisingly, Ser187 and Glut178 in subunit B were able to form H-bond with the DNA backbone but not in subunit A. These obvious differences were observed between two subunits because two subunits are not identical in all respect as a least square superposition between two subunits yields RMSD of 0.44Å. This could be an effect of binding of more molecules of cAMP to secondary pocket in subunit B than subunit A as seen in the structure. More of DNA-protein interactions in subunit B were observed in comparison to subunit A.

Although, *Mtb*-CRP is 32% identical to EC-CRP in terms of aminoacid sequence, the mode of cAMP binding to the protein is unique in *Mtb*-CRP. The implications of such binding are completely different quantitatively and qualitatively as compared to EC-CRP. In *Mtb*-CRP secondary cAMP pocket is completely unrelated. It would be interesting to address what is the cause of the formation of non-canonical cAMP binding site in case of *Mtb*-CRP. To answer these queries kinetics of ligand and DNA binding need to be studied in detail.

Mtb-CRP is essential for survival of TB bacilli inside host cells (Rickman *et al.*, 2005). These small unique but important structural differences between *Mtb*-CRP and *E.coli* and other bacteria could be targeted for drug discovery and could lead to the design of effective therapeutic interventions against *Mtb* in future.

4.3.7 Protein Data Bank accession code

The coordinates and structure factors have been deposited in the PDB under accession code 3HE9.

Chapter 5: Summary and Conclusions

Approximately eight million people develop active tuberculosis (TB) every year, with two million dying from the disease. In addition to this already huge burden of disease, it is estimated that up to two billion people have been infected with the causative agent, *Mycobacterium tuberculosis* (*Mtb*) (Dye *et al.*, 1999). Most people control the initial infection by mounting a cell-mediated immune response that prevents disease but can leave a residual population of viable mycobacteria. Between 5–10% of individuals who become infected subsequently develop clinical disease (Bloom and Murray, 1992). Primary TB develops within 1 or 2 years after an initial infection and, particularly in children, is often associated with disseminated disease. Post-primary TB develops later in life, and can be caused either by reactivation of bacteria remaining from the initial infection or by failure to control a subsequent reinfection. Post-primary TB is predominantly a pulmonary disease, involving extensive damage to the lungs and efficient aerosol transmission of bacteria. The risk of disease is highly dependent on the immune status of the host; co-infection with HIV markedly increases the incidence of both forms of disease. The latency attribute of the pathogen represents a significant obstacle to the worldwide control and eradication of tuberculosis because the non-replicating bacilli may be in a state of 'drug indifference' wherein they are not killed by the drugs. In contrast to individuals with active tuberculosis, individuals with latent tuberculosis do not transmit the disease. The harsh environment faced by the bacillus

inside the human macrophages and dendritic cells include depletion of nutrients, shift in pH, production of growth limiting products and/or depletion of oxygen. On the other hand these extreme conditions give some how signal to the *Mtb* cells to slow-down this metabolism and go into dormancy/latency. Maintenance of dormancy within the host cells and granuloma for many years or decades is the outcome of a balanced war between the host immune response and the pathogen's tactics to overcome it (Wayne and Sohaskey, 2001).

Transcriptional regulation in response to environmental changes encountered during infection is a common theme in bacterial pathogenesis, and similar concepts can be applied to dormant/latent bacteria. *Mtb* cells have to reprogram their transcriptional expression profiles to survive within the harsh environment presented by the macrophages. cAMP receptor protein/fumarate nitrate reductase (CRP/FNR) regulators is a family of eubacterial transcriptional factors associated with a variety of stress responses like hypoxia, nutrition depletion and redox regulation within the bacteria (Korner *et al.*, 2003). *Mtb* ORF *Rv3676* encodes cAMP receptor protein and knock out mutants have been reported to be defective for growth in animal models and macrophages (Rickmann *et al.*, 2005).

The present thesis is an attempt to study the properties of cAMP receptor protein from *Mtb* and also regulation by CRPs of other

mycobacterial species with an eventual goal to use this as drug candidate for intervention against latent and active TB.

This work started with the computer based bioinformatics approach aimed to identify the operons which could be regulated by CRP in mycobacteria. In earlier computational predictions of CRP-regulon, 73 binding elements in *Mtb* genome were observed (Bai *et al.*, 2005). Of the top 44 binding sites identified in the present study, 25 have been reported previously (Bai *et al.*, 2005). These analyses thus highlighted 19 new *Mtb*-CRP binding sites upstream of various operons in *Mtb* genome. The CRP-binding element from *fumarate reductase* (*Rv1552*) received the highest score. While this DNA element was also reported as a potential CRP-binding site in earlier reports (Bai *et al.*, 2005; Rickman *et al.*, 2005, Spreadbury *et al.*, 2005), in the work presented here, experimental evidence for specific DNA: protein interaction between this DNA element and recombinant purified *Mtb*-CRP was provided. Also, Spreadbury *et al.* (2005) proposed some potential genes as members of the CRP-regulon. There are several overlaps between the results from this work and others. While previous studies (Bai *et al.*, 2005, Spreadbury *et al.*, 2005) utilized information from *E. coli* CRP-regulon, in the present study only the available information from *Mtb* CRP-regulon was used. Recently, a comprehensive comparative account of all *Mtb*-CRP regulons has also been published (Krawczyk *et al.*, 2009). Out of these novel identified

members predicted in the present analysis of CRP regulon in *Mtb* genome many are critical for pathogenesis and general life cycle of the bacterium. It includes genes related to cell wall biogenesis, 5'-3' Cyclic Adenosine Monophosphate (cAMP) signaling, aminoacid biosynthesis pathways and recycling machinery of the cell.

CRP regulators from different species of mycobacteria have very similar DNA binding domains when compared with *Mtb*-CRP in terms of amino acid sequences (for *M. avium*, *M. leprae* and *M. smegmatis* are 96%, 96% and 97% identical respectively). Given the observation that the CRP proteins from all mycobacteria have identical DNA binding domains, the same profile matrix constructed for *Mtb* was extended to predict CRP binding sites in the genomes of *M. leprae, M. avium subsp. paratuberculosis* and *M. smegmatis.* This represented the first such attempt to interrogate other mycobacterial genomes. The operon context of these regulons across the genomes was also described.

A comparative analysis of CRP target genes in various species enabled the identification of the common CRP regulated genes across mycobacteria and at least 18 genes were found to be common. Conservation of these genes in the predicted CRP regulons suggests an important role of their cognate gene products in the mycobacterial life cycle. *Mtb*-CRP was earlier reported to be essential for the survival of mycobacteria inside macrophages and in animal models (Rickman *et al.*, 2005). Further, in these analyses, a high conservation of these CRP

regulated genes among pathogenic mycobacteria than in non-pathogenic mycobacteria was found. This strengthened the notion that *Mtb*-CRP and its regulated genes are important for pathogenesis of mycobacteria and that these might have co-evolved with the pathogenic branch as a result of genome optimization.

In the next part of study *Mtb*-CRP was cloned, expressed and the recombinant protein was purified from *E.coli*. It was characterized in terms biophysical and biochemical properties. Analytical size exclusion chromatography was carried out to determine the apparent oligomeric nature, if any, of the purified *Mtb*-CRP protein. *Mtb*-CRP was found to exist as a dimer of ~53 kDa (apparent molecular weight).

In most of oxygen tension-sensing proteins belonging to the CRP/FNR family of proteins, transition metals like Fe or Ni are associated with the protein to sense the fluctuations of oxygen availability *via* redox mechanisms (Korner *et al.*, 2003). The absorption spectrum of purified *Mtb*-CRP was scanned to check for the presence of metal ion cofactor. Resulting spectra revealed only two characteristic peaks of proteins, one at 295 nm and the other at 280 nm. No bound associated metal cofactor was found. This suggested that *Mtb*-CRP apparently uses some other mechanism(s) to sense effector signals.

Further, the cAMP binding properties of purified recombinant *Mtb*-CRP were determined. Results of protein family search (pfam) revealed the presence of a putative cAMP-binding domain at the N-terminal end of

Mtb-CRP protein, thereby raising a strong probability that cAMP may be acting as an effector of *Mtb*-CRP. Purified *Mtb*-CRP was subjected to CD analysis in the presence and absence of cAMP as ligand. A comparison of CD spectra of these two forms provided evidence of binding as readout of change in secondary structure. The two tryptophan residues (Trp112 and Trp203) present in *Mtb*-CRP were used as probe to study the effect of cAMP concentration on conformational changes. This change in secondary structure clearly appeared to be a function of increasing concentration of cAMP. That cAMP indeed caused concentration dependent conformational alterations within *Mtb*-CRP was evident from tryptophan fluorescence spectrometry data. Physiological cAMP levels are in the range of 0–10 mM. At lower concentrations (6–10 mM), the binding showed positive cooperativity, and at 10 mM cAMP the protein existed in the most open conformation. With further increase of cAMP (12–16 mM), the protein was compacted which could be a reflection of a feedback regulation. To investigate the ability of *Mtb*-CRP to bind to its cognate DNA motif, EMSA was carried out using purified *Mtb*-CRP and a radiolabeled oligonucleotide carrying the CRP/FNR-binding site present upstream of the *frd* (*Rv1552*) gene encoding the fumarate reductase enzyme. This binding site was identified as a putative binding site in previous reports (Bai *et al.*, 2005; Spreadbury *et al.*, 2005). In the present work, during *in silico* regulon prediction studies, this motif elicited the highest score, and therefore it was selected for EMSA. It has been

reported that *Mtb*-CRP senses oxygen (Bai *et al.*, 2005; Spreadbury *et al.*, 2005) indirectly by controlling the expression of genes such as *frd*. Fumarate serves as an alternative electron acceptor in the absence of oxygen, and this is mediated by a membrane-linked fumarate reductase enzyme complex (Lambden and Guest, 1976). The putative CRP/FNR binding site, present upstream of the *frd* operon, was recognized by purified *Mtb*-CRP protein and this was evident from EMSA. The predicted cAMP-binding site in *Mtb*-CRP, indeed showed binding to cAMP leading to conformational changes in the protein as evident from spectral analyses. The extent of change in secondary structure was maximal in the presence of 10 mM cAMP.

To investigate the possible mechanistic role of *Mtb*-CRP in transcription, X-ray crystal structure of *Mtb*-CRP-cAMP-DNA ternary complex was determined. The structure of *Mtb*-CRP-DNA-cAMP was solved in the $P2_1$ spacegroup at a 2.9Å resolution. Using the CRP/FNR family protein structure (pdb entry: 2GAU) from *Porphyromonas gingivalis* as MR search model, it was possible to build complete protein structure as well as 23 basepairs of DNA and three cAMP molecules in the complex. In both of the subunits of the *Mtb*-CRP-DNA structure, it was observed that the very C-terminus residues forms an alpha-helix. It was investigated if this terminal could play a functional role in the ability of CRP to form complex. Results of deletion by side directed mutagenesis indicated that the helix-G was playing an active role in the folding of the

protein. Residues from the helix-G which could function as a flap to cover highly hydrophobic core formed at the interface of these three helices (E, F and G). This hydrophobic core consisted of Trp203, Ile204, Phe161, Val171, and Phe198and Val150. These all hydrophobic side chains were facing the centre of the core. There were three residues (namely Leu219, Ala220 and Ala223) from helix-G which were facing this hydrophobic core, also hydrophobic in nature, and interacting with the core side chains *via* non bonding interactions. Interestingly, neither the residues corresponding to helix-G were present in homologous proteins nor the hydrophobic core. These results indicated that this helix could be a unique feature of mycobacterial CRPs and related proteins.

In the structure there was one novel non-canonical cAMP binding site. The finding of a third cAMP molecule was unexpected, although secondary cAMP binding sites in EC-CRP, reported earlier (Passner 1997) was different from the one presented in this work. In EC-CRP the site is surface exposed and located near the DNA protein interaction area and interacting with Arg180 (located on DNA interacting helix-F), Glu58 (in nearby beta-sheet) and Gly173 and Gly177 (both on helix-E) (Passner 1997; pdb entry: 2CGP). In *Mtb*-CRP the secondary pocket was buried and located at the interface of three helices: helix-D (dimer interface helix), helix-E and helix-F (constituent of DNA interacting helix-turn-helix). This cAMP could directly interact with six of the side chains surrounding the pocket (Asn67, Leu 68, Asn137, Asp140, Glu156 and

Leu156) *via* several H-bonds and salt bridges. Least square superimposition analysis with *Mtb*-CRP-cAMP structure (PDB code: 3i54) did not reveal much difference with *Mtb*-CRP-DNA structure (RMSD=1.16 Å). This pocket environment was found to be conserved including the side-chains interacting with cAMP. For unknown reasons in *Mtb*-CRP-cAMP complex structure third cAMP was not bound. In case of EC-CRP it was known that interaction with DNA could give rise to secondary cAMP binding (Lin *et al.*, 2002, Scott and Jarjous, 2005). This might be similar for *Mtb*-CRP. This position seems interesting in terms of regulatory function for DNA interactions.

For this non-canonical pocket cAMP molecule was predicted to be present in only subunit B of *Mtb*-CRP with 50% occupancy. This indicated that this is seemingly a secondary binding site and not occupied fully in all molecules. On the other hand, some residual density on the similar NCS related position in sub-unit A was also seen although it was not sufficient to clearly position the cAMP. Taken together, this gave indication of concentration dependent sequential multi-step binding events for each sub-unit. Second non-canonical pocket has lower affinity than canonical one as it only had 50% occupancies. This observation was consistent with biophysical and biochemical studies for EC-CRP (Heyduk and Lee, 1989; Lee *et al.*, 1999; Takahashi *et al.*, 1980). In a surface-potential analysis a cleft for entry of cAMP to the pocket could be seen. It was also observed that the surface is highly positively charged

near this cleft. This positive charge could facilitate the entry of cAMP to the non canonical pocket.

Upon comparison analysis with the EC-CRP-DNA structure (pdb: 1O3S) different position of the helix-D (DBD) and helix-F (DNA interacting) of more than 3 Å was observed. This could be due to the presence of secondary binding pocket. This movement in helix-D translated as movement in helix-F which finally gives rise to more interaction with DNA. On looking in minute details it was observed that this helix rearrangement could be mediated by H-bonding between phosphate of cAMP and side-chain amino groups from Asn137 and Gln156. When the structure of *Mtb*-CRP-DNA was superimposed onto EC-CRP (PDB entry: 1O3S) it was found that the side chains of Met 59 and Lys129 occupied the binding pocket of CRP in *E. coli*, whereas the corresponding residues, Asn67 and Asn137 have lighter side chains in *Mtb*-CRP and were positioned to allow cAMP binding or specifically interact with the bound cAMP molecule. This indicated that the third binding pocket was specific to cAMP, but not conserved among all bacteria. Similar comparative analyses for the position of DNA interacting helix-F in *Mtb*-CRP-cAMP (PDB code: 3i54) were also performed but not any big movement in position could be seen. This might suggest that *Mtb*-CRP-cAMP form reported earlier in which only one cAMP is occupied in each sub-unit could be also the active form of *Mtb*-CRP ready to bind DNA and could start activation of transcription (Reddy *et al.*, 2009). In

this case second cAMP binding could further regulate the transcription in a feed forward type of mechanism.

To investigate further, the importance of this secondary pocket, *Mtb*-CRP mutants which could mimic the non-pocket environment found in EC-CRP were designed. The generated mutants were expected not to bind cAMP as the larger residues were expected to sterically hinder the binding of cAMP as seen in EC-CRP. The mutant allowed investigating the effect of the secondary binding pocket on the ability of *Mtb*-CRP to bind the DNA. Four sets of experiments were performed: 1) Wild type *Mtb*-CRP and the DNA motif; 2) Wild type *Mtb*-CRP saturated with 500uM cAMP prior to addition of DNA; 3) *Mtb*-CRPMut with DNA; and, finally 4) *Mtb*-CRPMut saturated with cAMP (500uM). The Kd values for the DNA binding events for *Mtb*-CRP with and without cAMP saturated states showed positive effect on DNA binding efficiency. The dissociation constant (Kd) dropped by almost 100 folds upon addition of cAMP (from 16nM to 0.16nM). In a parallel experiment with *Mtb*-CRPMut which was devoid of novel non-canonical pocket no huge drop in Kd was evident rather it was only ~3 times (from 30nM to 10nM). Interestingly, EC-CRP showed very poor DNA binding when it was treated with higher concentrations (0.5mM) of cAMP (Harman, 2001; Garges and Adhya, 1995). This clearly indicated that the effects of cAMP binding on DNA interactions were different in EC-CRP and *Mtb*-CRP and novel non-

canonical cAMP pocket have a critical role in regulation of DNA binding by *Mtb*-CRP.

Further results showed that energetics of DNA binding were also very distinct in *Mtb*-CRP. The change in enthalpy for cAMP liganded-*Mtb*-CRP (15.5 kcal/mol) and DNA interaction was rather higher than the energy released in unliganded-*Mtb*-CRP (9.15 kcal/mol). This high energy release was also compensated by high amount of change in entropy (-15cal/mol/K) in liganded interaction. Entropy change difference between liganded and unliganded interactions was almost 20cal/mol/K. While in similar experiment with *Mtb*-CRPMut big difference in release of energy was not seen. High amount of enthalpy changes could be a reflection of specificity of binding reaction. According to present accepted model for EC-CRP, binding of first cAMP favors the binding of cAMP molecules at secondary cAMP sites. Binding of extra cAMPs to the secondary binding site decreases its specificity for DNA (Lin, 2002; Scott, 2005). It should be noted that the secondary cAMP binding site in *Mtb*-CRP is at a different location than that in EC-CRP. It is therefore likely that the two binding site have different mechanism. ITC DNA binding data on *Mtb*-CRP and *Mtb*-CRPMut clearly showed that secondary cAMP increases the specificity of CRP for DNA instead of decreasing its specificity.

Currently, the only CRP-DNA complex structures available are from *E.coli*. Therefore a comparative analysis of the DNA-protein interactions between EC-CRP and *Mtb*-CRP was carried out. The minimal

motif in DNA as well as in protein which is necessary for interaction was defined. A predicted helix-turn-helix (176-LTQEEIAQLVGASRETVNKALA-196) motif which is involved in DNA binding for *Mtb*-CRP was identified. With the help of DNA-protein complex crystal structure it was further narrowed down to the exact minimal signature motif responsible for interaction. In EC-CRP this is Arg-Glu-thr-Xaa-Xaa-Arg (residues180-185) while for *Mtb*-CRP it was Arg-Glu-thr-Xaa-Xaa-Lys (residues188-193). At the last position in case of *Mtb*-CRP the lysine residue was substituted by arginine. While for DNA motif in EC-CRP the motif is "GCGA", in case of *Mtb*-CRP it was "GTGA". Except these base specific interactions some non-specific DNA backbone specific H-bonds were also observed. Val146, Asn192 and Arg65 could also interact with DNA backbone *via* H-bonds and salt bridges. Surprisingly, Ser187 and Glut178 in subunit B were able to form H-bond with the DNA backbone but not with subunit A. These obvious differences were observed between two subunits because two subunits were not identical in all respect as a least square superposition between two subunits yielded RMSD of 0.44Å. This could be an effect of binding of cAMP to secondary pocket in more molecules in subunit B than subunit A as seen in the structure. More DNA-protein interaction in subunit B comparison to subunit A was observed. Overall pattern of interactions was found to be similar except for two interactions, the lysine in *Mtb*-CRP could interact to only one base specifically while equivalent arginine in EC-CRP could make one

more interaction with thymine. On the other hand glutamate in *Mtb*-CRP signature motif could interact with thymine. So, in conclusion the overall interactions were conserved.

Although, *Mtb*-CRP is 32% identical to EC-CRP in terms of aminoacid sequence, the minimal signature motifs for DNA-protein interaction were also very similar. The fashion of cAMP binding to the protein was unique in *Mtb*-CRP. Further implications of binding which lead to fate and extent of DNA binding were completely different than EC-CRP. In *Mtb*-CRP secondary cAMP pocket was completely unrelated.

The *Mtb* genome encodes as many as 15 adenylate cyclases, suggesting that cAMP may have an important role in mycobacteria. It has indeed been reported that cAMP can alter the gene expression profile of *Mtb* during anaerobic conditions (Gazdik and McDonough, 2005).

While these *in vitro* findings point to the importance of cAMP, it remains to be experimentally demonstrated whether cAMP is actually involved in regulating gene expression by recruiting *Mtb*-CRP. While the biophysical features of purified *Mtb*-CRP described here are physiologically relevant, experimental validation *in vivo* will be required to dissect the complete network of *Mtb* genes regulated by *Mtb*-CRP and cAMP. cAMP, acting as effector, is known to modulate the regulation of a large number of target genes, and it is likely that *Mtb*-CRP will be involved in this process.

Mtb-CRP is essential for survival of TB bacilli inside host cells (Rickman *et al.*, 2005). Small unique but important structural features of *Mtb*-CRP described in this thesis could be targeted for drug discovery approaches and could lead to design of effective therapeutic interventions against *Mtb* in future.

Bibliography

Adams PD, Grosse-Kunstleve RW, Hung LW, Ioerger TR, McCoy AJ, Moriarty NW, Read RJ, Sacchettini JC, Sauter NK, and Terwilliger TC. 2002. Phenix: building new software for automated crystallographic structure determination. ***Acta Crystallogr D Biol Crystallogr*** 58:1948–1954.

Adams, DO. 1976. The granulomatous inflammatory response. A review, ***Am J Pathol*** 84:164–192.

Adhya S, Ryu S, Garges S. 1995. Chapter 10. Subcellular Biochemistry, Biswas S, Roy S (eds). Plenum Press: New York, 303–321. ***Adv. Microb. Physiol.*** 44:1-34.

Agarwal N, Lamichhane G, Gupta R, Nolan S and Bishai WR. 2009. Cyclic AMP intoxication of macrophages by a *Mycobacterium tuberculosis* adenylate cyclase. ***Nature*** 460:98-102.

Akhter, Y., Tundup, S., Hasnain, S.E., 2007, Novel biochemical properties of a CRP/FNR family transcription factor from *Mycobacterium tuberculosis*. ***Int. J. Med. Microbiol.*** 297:451–457.

Akhter, Y, Yellaboina S, Farhana A, Ranjan A, Ahmed N and Hasnain, SE. 2008. Genome scale portrait of cAMP Receptor Protein-Regulons in mycobacteria points to their role in pathogenesis. ***Gene*** 407:148-58.

Akif, M, Akhter Y, Huanain SE and Mande, SC. 2006. Crystallization and preliminary X-ray crystallographic studies of *Mycobacterium tuberculosis* CRP/FNR family transcriptional regulator. ***Acta Crystallography F*** 62:873-875.

Alspaugh, JA, Pukkila-Worley R, Harashima T, Cavallo LM, Funnell D, Cox GM, Perfect JR, Kronstad JW, and Heitman J. 2002. Adenylyl cyclase functions downstream of the G protein Gpa1 and controls mating and pathogenicity of *Cryptococcus neoformans*. ***Eukaryot. Cell*** 1: 75–84.

Anes E, Khnel MP, Bos E, Moniz-Pereira J, Habermann A, and Griths G. 2003. Selected lipids activate phagosome actin assembly and maturation resulting in killing of pathogenic mycobacteria. ***Nat Cell Biol.*** 5:793–802.

Anes E, Peyron P, Staali L, Jordao L, Gutierrez M G, Kress H, Hagedorn M, Maridonneau-Parini I, Skinner MA, Wildeman AG, Kalamidas SA, Kuehnel M and Griths G. 2006. Dynamic life and death interactions

between *Mycobacterium smegmatis* and J774 macrophages., ***Cell Microbiol.*** 8:939–960.

Aziz MA, Wright A, Laszlo A, Muynck AD, Portaels F, Deun AV, Wells C, Nunn P, Blanc L, Raviglione M, WHOUA 2006. Tuberculosis, and LDGP on Antituberculosis Drug Resistance Surveillance, Epidemiology of antituberculosis drug resistance (the Global Project on Anti-tuberculosis Drug Resistance Surveillance): an updated analysis. ***Lancet***, 368:2142–2154.

Bai, G, McCue, LA, McDonough, KA. 2005. Characterization of *Mycobacterium tuberculosis* Rv3676 (CRPMt), a cyclic AMP receptor protein-like DNA binding protein. ***J. Bacteriol.*** 187:795–8004.

Barnard AM, Green J and Busby SJ, 2003. Transcription regulation by tandem-bound FNR at *Escherichia coli* promoters. ***J. Bacteriol.*** 185:5993–6004.

Belisle JT, Vissa VD, Sievert T, Takayama K, Brennan PJ, Besra GS, 1997. Role of the major antigen of *Mycobacterium tuberculosis* in cell wall biogenesis. ***Science*** 276:1420–1422.

Benoff BH, Yang CL, Lawson, Parkinson G, Liu J, Blatter E, Ebright YW, Berman HM and Ebright RH. 2002. Structural basis of transcription activation: the CAP–αCTD–DNA complex, ***Science*** 297:1562–1566.

Berg, OG and von Hippel PH. 1988. Selection of DNA binding sites by regulatory proteins. II. The binding specificity of cyclic AMP receptor protein to recognition sites. ***J Mol Biol.*** 200:709–723.

Bloom BR and Murray CJ. 1992. Tuberculosis: commentary on a re-emergent killer. ***Science*** 257:1055–1064.

Borjigin M, Li, Lanz H, Kerby ND, GP Roberts and TL Poulos. 2007. Structure-based hypothesis on the activation of the CO-sensing transcription factor CooA, ***Acta Crystallogr. D Biol. Crystallogr.*** 63:282–287.

Botsford JL and Harman JG. 1992. Cyclic AMP in prokaryotes. ***Microbiol. Rev.*** 56:100–122.

Bradford MM. 1976. A rapid and sensitive method for quantification of microgram quantities of protein utilizing the principle of protein-dye binding. ***Anal Biochem*** 72:248-254.

Brosch, R, , Gordon SV, Marmiesse M, Brodin P, Buchrieser C, Eiglmeier K, Garnier T, Gutierrez C, Hewinson G, Kremer K, Parsons LM, Pym A S, Samper S, Van Soolingen D and Cole ST, 2002. A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. ***Proc Natl Acad Sci U S A*** 99:3684–3689.

Caler EV, Morty RE, Burleigh BA and Andrews NW. 2000. Dual role of signaling pathways leading to Ca2 and cyclic AMP elevation in host cell invasion by *Trypanosoma cruzi*. ***Infect. Immun.*** 68:6602–6610.

Casal MJ, Rodriguez FC, Luna MD and Benavente MC. 1987. In vitro susceptibility of *Mycobacterium tuberculosis*, *Mycobacterium africanum, Mycobacterium bovis, Mycobacterium avium, Mycobacterium fortuitum, and Mycobacterium chelonae* to ticarcillin in combination with clavulanic acid. ***Antimicrob. Agents Chemother.*** 31:132–133.

Chambers HF *et al.*, 1995. Can penicillins and other beta-lactam antibiotics be used to treat tuberculosis? ***Antimicrob. Agents Chemother.*** 39:2620–2624.

Chambers HF, Kocagoz T, Sipit T, Turner J and Hopewell PC. 1998. Activity of amoxicillin/clavulanate in patients with tuberculosis. ***Clin. Infect. Dis.*** 26, 874–877.

Chan MK. CooA, CAP and allostery. 2000. ***Nat. Struct. Biol.*** 7:822–824.

Chen S, Gunasekera A, Zhang X, Kunkel TA, Ebright RH and Berman HM. 2001. Indirect readout of DNA sequence at the primary-kink site in the CAP-DNA complex: alteration of DNA binding specificity through alteration of DNA kinking. ***J Mol Biol*** 314: 75-82

Chu SY, Tordova M, Gilliland GL, Gorshkova I, Shi Y, Wang S and Schwarz FP. 2001. The structure of the T127L/S128A mutant of cAMP receptor protein facilitates promoter site binding, ***J. Biol. Chem.*** 276:11230–11236.

Cipriani F, Felisaz F, Launer *et al.*, 2006. Automation of sample mounting for macromolecular crystallography. ***Acta Crystallogr D Biol Crystallogr.*** 62:1251-9.

Cole ST, Brosch R, Parkhill J, *et al.* 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. ***Nature*** 393:537–544.

Content J, de la Cuvellerie A, De Wit L, Vincent-Levy-Frebault V, Ooms J and De Bruyn J. 1991. The genes coding for the antigen 85 complexes of *Mycobacterium tuberculosis* and *Mycobacterium bovis* BCG are members of a gene family: cloning, sequence determination, and genomic organization of the gene coding for antigen Ag85C of *M. tuberculosis*. ***Infect. Immun.*** 59:3205–3212.

Converse PJ, Karakousis PC, Klinkenberg LG, Kesavan AK, Ly LH, Allen SS, Grosset JH, Jain SK, Lamichhane G, Manabe YC, McMurray DN, Nuermberger EL, and Bishai WR. 2009. Role of the dosR-dosS two-component regulatory system in *Mycobacterium tuberculosis* virulence in three animal models. ***Infect Immun.*** 77:1230–1237.

Corbett EL, Watt CJ, Walker N, Maher D, Williams BG, Raviglione MC and Dye C. 2003. The growing burden of tuberculosis: global trends and interactions with the HIV epidemic. ***Arch Intern Med,*** 163:1009–1021.

Crooks GE, Hon G, Chandonia JM and Brenner SE. 2004. WebLogo: a sequence logo generator. ***Genome Res.*** 14:1188–1190.

Cynamon MH and Palmer GS. 1983. In vitro activity of amoxicillin in combination with clavulanic acid against *Mycobacterium tuberculosis.* ***Antimicrob. Agents Chemother.*** 24:429–431.

D'Souza CA, and Heitman J. 2001. Conserved cAMP signaling cascades regulate fungal development and virulence. ***FEMS Microbiol. Rev.*** 25: 349–364.

Davis IW, Leaver Fay A, Chen VB, Block JN, Kapral GJ, Wang X, Murray, LW, Arendal III W B, Snoeyink J, Richardson JS, and Richardson DC. 2007. MolProbity: all atom contacts and structure validation for proteins and nucleic acids. ***Nucleic Acids Research*** 35:W375:W383.

Dye C, Scheele S, Dolin P, Pathania V and Raviglione MC. 1999. Consensus statement. Global burden of tuberculosis: estimated incidence, prevalence, and mortality by country. WHO Global Surveillance and Monitoring Project. ***JAMA*** 282:677–686.

Eiting MG, Hagelüken WD, Schubert and Heinz DW. 2005. The mutation G145S in PrfA, a key virulence regulator of *Listeria monocytogenes*, increases DNA-binding affinity by stabilizing the HTH motif. ***Mol. Microbiol.*** 56:433–446.

Emsley P and Cowtan K. 2004. Coot: model-building tools for molecular graphics., ***Acta Crystallogr D Biol Crystallogr*** 60:2126–2132.

Fischer HM. 1994. Genetic regulation of nitrogen fixation in rhizobia. ***Microbiol. Rev.*** 58:352-386.

Flynn JL. 2006. Lessons from experimental *Mycobacterium tuberculosis* infections., ***Microbes Infect***, 8:1179–1188.

French, S. and Wilson, K. 1978. ***Acta Cryst.*** A34:517-525.

Frieden TR, Sterling TR, Munsi SS, Watt CJ and Dye C. 2003. Tuberculosis., ***Lancet***, 362:887–899.

Gallagher DT, Smith N, Kim SK, Robinson H and Reddy PT. 2009. Profound Asymmetry in the Structure of the cAMP-free cAMP Receptor Protein (CRP) from *Mycobacterium tuberculosis*. ***J Biol Chem.*** 284:8228-32.

Gande R, *et al.*, 2004. Acyl-CoA carboxylases (accD2 and accD3), together with a unique polyketide synthase (Cg-pks), are key to mycolic acid biosynthesis in Corynebacterianeae such as *Corynebacterium glutamicum* and *Mycobacterium tuberculosis*. ***J. Biol. Chem***. 279:44847–44857.

Garcia E and Rhee SG. 1983. Cascade control of *Escherichia coli* glutamine synthetase. Purification and properties of PII uridylyltransferase and uridylyl-removing enzyme. ***J. Biol. Chem.*** 258:2246–2253.

Gazdik MA and McDonough KA 2005. Identification of cyclic AMP-regulated genes in *Mycobacterium tuberculosis* complex bacteria under low-oxygen conditions. ***J. Bacteriol.*** 187:2681–2692.

Green J, Scott C and Guest J. 2001. Functional versatility in the CRP-FNR superfamily of transcription factors: FNR and FLP. ***Adv. Microb. Physiol.*** 44:1–34.

Hagedorn M, Rohde KH, Russell DG and Soldati T. 2009. Infection by tubercular mycobacteria is spread by non-lytic ejection from their amoeba hosts. ***Science*** 323:1729–1733.

Harman JG. 2001. Allosteric regulation of the cAMP receptor protein. ***Biochim Biophys Acta.*** 1547:1-17.

Heyduk T and Lee JC. 1989. *Escherichia coli* cAMP receptor protein: evidence for three protein conformational states with different promoter binding affinities. ***Biochemistry*** 28:6914–6924.

Heyduk T, Lee JC, Ebright YW, Blatter EE, Zhou Y and Ebright RH. 1993. CAP interacts with RNA polymerase in solution in the absence of promoter DNA, ***Nature*** 364:548–549.

Hollands K, Busby SJW and Lloyd GS. 2007. New targets for the cyclic AMP receptor protein in the *Escherichia coli* K-12 genome, ***FEMS Microbiol. Lett.*** 274:89–94.

Hudson JM, Crowe LG, Fried MG. 1990. A new DNA binding mode for CAP. ***J Biol Chem*** 265:3219-25.

Hunt DM, Saldanha JW, Brennan JF, Benjamin P, Strom M, Cole JA, Spreadbury CL and Buxton RS. 2008. Single nucleotide polymorphisms that cause structural changes in the cyclic AMP receptor protein transcriptional regulator of the tuberculosis vaccine strain *Mycobacterium bovis* BCG alter global gene expression without attenuating growth. ***Infect Immun.*** 76:2227-34.

Jackson M, *et al.*, 1999. Inactivation of the antigen 85c gene profoundly affects the mycolate content and alters the permeability of the *Mycobacterium tuberculosis* cell envelope. ***Mol. Microbiol.*** 31:1573–1587.

Joyce, MG, Levy C, Gábor K, Pop SM, Biehl BD, Doukov TI, Ryter JM, Mazon H, Smidt H, van den Heuvel RHH, Ragsdale SW, van der Oost J and Leys D. 2006. CprK crystal structures reveal mechanism for transcriptional control of halorespiration, ***J. Biol. Chem.*** 281:28318–28325.

Kabsch W. 1993. ***J. Appl. Cryst.*** 26:795-800.

Kasik JE. 1979. Mycobacterial beta-lactamases. In: Hamilton-Miller, J.M.T., Smith, J.T. (Eds.), ***beta-Lactamases.*** Academic Press, New York, 339–350.

Kaufman S, Gilvarg C, Cori O and Ochoa S. 1953. Enzymatic oxidation of alpha-ketoglutarate and coupled phosphorylation. ***J. Biol. Chem.*** 203:869–888.

Kaufmann SHE. and McMichael AJ. 2005. Annulling a dangerous liaison: vaccination strategies against AIDS and tuberculosis. ***Nat Med***.11:S33–S44.

Kimerling ME, Kluge H, Vezhnina N, Iacovazzi T, Demeulenaere T, *et al.*1999. Inadequacy of the current WHO re-treatment regiment in central Siberian prisons: treatment failure and MDR-TB. ***Int. J. Tuberc. Lung Dis.*** 3:451–53.

Kinchen JM and Ravichandran KS. 2008. Phagosome maturation: going through the acid test., ***Nat Rev Mol Cell Biol***. 9:781–795.

Kolb A, Busby S, Garges S and Adhya S. 1993. Transcriptional regulation by cAMP and its receptor protein. ***Annu. Rev. Biochem.*** 62:749–795.

Komori H, Inagaki S, Yoshioka S, Aono S and Higuchi Y. 2007. Crystal structure of CO-sensing transcription activator CooA bound to exogenous ligand imidazole, ***J. Mol. Biol.*** 367:864–871.

Körner H, Sofia HJ and Zumft WG. 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.

Krawczyk J, Kohl TA, Goesmann A, Kalinowski J and Baumbach J. 2009. From *Corynebacterium glutamicum* to *Mycobacterium tuberculosis*--towards transfers of gene regulatory networks and integrated data analyses with MycoRegNet. ***Nucleic Acids Res***. 37:e97.

Kremer L, Baulard AR and Besra GS. 2000. Molecular Genetics of Mycobacteria; Eleventh chapter: Genetics of Mycolic Acid Biosynthesis. ASM Press, 173–190.

Krueger SI, Gorshkova J, Brown J, Hoskins KH and McKenney FP Schwarz. 1998. Determination of the conformations of cAMP receptor protein and its T127L, S128A mutant with and without cAMP from small angle neutron scattering measurements. ***J. Biol. Chem.*** 273:20001–20006.

Kumar P, Joshi DC, Akif M, Akhter Y, Hasnain SE and Mande SC. 2009. Crystal Structure of apo-Cyclic AMP Receptor Protein of *M. tuberculosis* and Normal Mode Analyses reveal an Elegant Mechanism of Allostery induced upon cAMP binding. ***Biophysical Journal*** (In press)

Kwon HH, Tomioka H and Saito H. 1995. Distribution and characterization of beta-lactamase of mycobacteria and related organisms. ***Tuber Lung Dis*** 76:141–148.

Labbe RF, Kurumada T and Onisawa, J. 1965. The role of succinyl-CoA synthetase in the control of heme biosynthesis. ***Biochim. Biophys. Acta.*** 111:403–415.

Lambden PR and Guest JR. 1976. Mutants of *Escherichia coli* K12 unable to use fumarate as an anerobic electron acceptor. ***J. Gen. Microbiol.*** 97:145–160.

Lanzilotta, WN, Schuller DJ, Thorsteinsson MV, Kerby RL, Roberts GP and Poulos T.L. 2000. Structure of the CO sensing transcription activator CooA. ***Nat. Struct. Biol.*** 7:876–880.

Laskowski RA, MacArthur MW, Moss DS and Thornton JM. 1993. PROCHECK: a program to check the stereochemical quality of protein structures. ***J. Appl. Crystallog.*** 26:283:291.

Lawson CL, Swigon D, Murakami KS, Darst SA, Berman HM and Ebright RH. 2004. Catabolite activator protein: DNA binding and transcription activation, ***Curr. Opin. Struct. Biol.*** 14:10–20.

Leu SF, Baker CH, Lee EJ and Harman JG. 1999. Position 127 amino acid substitutions affect the formation of CRP:cAMP:lacP complexes but not CRP:cAMP:RNA polymerase complexes at lacP. ***Biochemistry*** 38:6222–6230.

Li J, Cheng X and Lee JC. 2002. Structure and dynamics of the modular halves of *Escherichia coli* cyclic AMP receptor protein. ***Biochemistry*** 41:14771–14778.

Lillebaek, T, Dirksen A, Baess I, Strunge B, Thomsen V and Andersen AB. 2002. Molecular evidence of endogenous reactivation of *Mycobacterium tuberculosis* after 33 years of latent infection. ***J Infect Dis*** 185:401–404.

Lin PL, Pawar S, Myers A, Pegu A, Fuhrman C, Reinhart A T, Capuano, SV, Klein E and Flynn JL. 2006. Early events in *Mycobacterium tuberculosis* infection in cynomolgus macaques. ***Infect Immun*** 74:3790–3803.

Lin SH, Kovac L, Chin AJ, Chin CC and Lee JC. 2002. Ability of *E. coli* cyclic AMP receptor protein to differentiate cyclic nucelotides: effects of single site mutations. ***Biochemistry*** 41:2946–2955.

Linder JU and Schultz JE. 2003. The class III adenylyl cyclases:multi-purpose signalling modules. ***Cell. Signal.*** 15:1081–1089.

Małecki JA, Polit Z, Wasylewski. 2000. Kinetic studies of cAMP-induced allosteric changes in cyclic AMP receptor protein from *Escherichia coli*, ***J. Biol. Chem.*** 275:8480–8486.

Marchler-Bauer, A., Panchenko, A.R., Shoemaker, B.A., Thiessen, P.A., Geer, L.Y. and Bryant, S.H., 2002. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. ***Nucleic Acids Res.*** 30:281–283.

Mattow J, *et al.*, 2001. Identification of proteins from *Mycobacterium tuberculosis* missing in attenuated *Mycobacterium bovis* BCG strains. ***Electrophoresis*** 22:2936–2946.

McKay DB and Steitz TA. 1981. Structure of catabolite gene activator protein at 2.9 A resolution suggests binding to left-handed B-DNA. ***Nature*** 290:744-9.

Mesa S, Bedmar EJ, Chanfon A, Hauke H and Fischer H. 2003. *Bradyrhizobium japonicum* NnrR, a denitrification regulator, expands the FixLJ-FixK2 regulatory cascade. ***J. Bacteriol.*** 185:3978–3982.

Mueller-Dieckmann, J. 2006. ***Acta Cryst. D*** 62:1446-1452.

Murphy D, Corner LAL and Gormley E. 2008. Adverse reactions to *Mycobacterium bovis* Bacille Calmette-Gurin (BCG) vaccination against tuberculosis in humans, veterinary animals and wildlife species.***Tuberculosis (Edinb)***. 88:344–357.

Mwandumba HC, Russell DG, Nyirenda MH, Anderson J, White SA, Molyneux ME and Squire SB. 2004. *Mycobacterium tuberculosis* resides in nonacidified vacuoles in endocytically competent alveolar macrophages from patients with tuberculosis and HIV infection. ***J Immunol***. 172:4592-8.

Nathan C and Shiloh MU. 2000. Reactive oxygen and nitrogen intermediates in the relationship between mammalian hosts and microbial pathogens. ***Proc Natl Acad Sci U S A*** 97:8841–8848.

Ottaway JH, McClellan JA, Saunderson CL, 1981. Succinic thiokinase and metabolic control. ***Int. J. Biochem.*** 13:401–410.

Panjikar S, Parthasarathy V, Lamzin VS, Weiss, MS and Tucker PA. 2005. ***Acta Cryst. D*** 61:449-457.

Parkinson G, Wilson C, Gunasekera A, Ebright YW, Ebright RE and Berman HM. 1996. Structure of the CAP–DNA complex at 2.5 angstroms resolution: a complete picture of the protein–DNA recognition. ***J. Mol. Biol.*** 260:395–408.

Passner JM, and Steitz TA. 1997. The structure of a CAP-DNA complex having two cAMP molecules bound to each monomer. ***Proc Natl Acad Sci USA*** 94: 2843–2847.

Popovych N, Tzeng SR, Tonelli M, Ebright RH and Kalodimos CG. 2009. Structural basis for cAMP-mediated allosteric control of the catabolite activator protein.***Proc Natl Acad Sci. USA*** 106:6927-32.

Prakash P, Yellaboina S, Ranjan A and Hasnain SE. 2005. Computational prediction and experimental verification of novel IdeR binding sites in the upstream sequences of *Mycobacterium tuberculosis* ORFs. ***Bioinformatics*** 21:2161–2166.

Ray JCJ., Flynn JL, and Kirschner DE. 2009. Synergy between individual TNF-dependent functions determines granuloma performance for controlling *Mycobacterium tuberculosis* infection. ***J Immunol.*** 182:3706–3717.

Reddy MC, Palaninathan SK, Bruning JB, Thurman C, Smith D, Sacchettini JC. 2009. Structural insights into the mechanism of the allosteric transitions of the *Mycobacterium tuberculosis* cAMP receptor protein. ***J Biol Chem.*** (In press)

Rickman L, Scott C, Hunt DM, Hutchinson T, Mene´ndez MC, Whalan R, Hinds J, Colston MJ, Green J and Buxton RS. 2005. A member of the cAMP receptor protein family of transcription regulators in *Mycobacterium tuberculosis* is required for virulence in mice and controls transcrip-tion of the rpfA gene coding for a resuscitation promoting factor. ***Mol. Microbiol.*** 56:1274–1286.

Russell DG. 2007. Who puts the tubercle in tuberculosis? ***Nat Rev Microbiol***. 5:39–47.

Sacchettini JC, Rubin EJ, and Freundlich JS. 2008. Drugs versus bugs: in pursuit of the persistent predator *Mycobacterium tuberculosis*. ***Nat Rev Microbiol***. 6:41–52.

Schaible UE and Kaufmann SHE. 2007. Malnutrition and infection: complex mechanisms and global impacts. ***PLoS Med.*** 4:e115.

Schilling O *et al.*, 2005. Zinc-and iron-dependent cytosolic metallo-beta-lactamase domain proteins exhibit similar zinc-binding affinities, independent of an atypical glutamate at the metal-binding site. ***Biochem. J.*** 385:145–153.

Schluger NW and Rom WN. 1998. The host immune response to tuberculosis. ***Am J Respir Crit Care Med.*** 157:679–691.

Schultz, SC, Shields GC and Steitz TA. 1991. Crystal structure of a CAP–DNA complex: the DNA is bent by 90 degrees. ***Science*** 253:1001–1007.

Scott SP, Jarjous S. 2005. Proposed structural mechanism of Escherichia coli cAMP receptor protein cAMP-dependent proteolytic cleavage protection and selective and nonselective DNA binding. ***Biochemistry*** 44:8730–8748.

Segura C, Salvado M, Collado I, Chaves J and Coira A. 1998. Contribution of beta-lactamases to beta-lactam susceptibilities of susceptible and multidrug-resistant *Mycobacterium tuberculosis* clinical isolates. ***Antimicrob. Agents Chemother.*** 42, 1524–1526.

Shaw DJ, Rice DW and Guest JR. 1983. Homology between CAP and Fnr, a regulator of anaerobic respiration in Escherichia coli. ***J. Mol. Biol.*** 166:241-247.

Sorg TB, and Cynamon MH. 1987. Comparison of four beta-lactamase inhibitors in combination with ampicillin against *Mycobacterium tuberculosis*. ***J. Antimicrob. Chemother.*** 19:59–64.

Spiro S.1994. The FNR family of transcription regulators. ***Antonie van Leiwenhoek*** 66:23–26.

Spreadbury CL, Pallen MJ, Overton T, Behr MA, Mostowy S, Spiro S, Busby SJ and Cole JA. 2005. Point mutations in the DNA- and cNMP-binding domains of the homologue of the cAMP receptor protein (CRP) in *Mycobacterium bovis* BCG: implications for the inactiva-tion of a global regulator and strain attenuation. ***Microbiology*** 151:547–556.

Stewart GR, Robertson BD and Young DB. 2003. Tuberculosis: a problem with persistence. ***Nat Rev Microbiol***, 1:97–105.

Takahashi M, Blazy B and Baudras A. 1980. An equilibrium study of the cooperative binding of adenosine cyclic 3',5'-monophosphate and

guanosine cyclic 3',5'-monophosphate to the adenosine cyclic 3',5'-monophosphate receptor protein from *Escherichia coli*. ***Biochemistry*** 19: 5124–5130.

Takayama K, Wang C and Besra GS. 2005. Pathway to synthesis and processing of mycolic acids in *Mycobacterium tuberculosis.* ***Clin. Microbiol. Rev.*** 18:81–101.

Tobin DM and Ramakrishnan L. 2008. Comparative pathogenesis of *Mycobacterium marinum* and *Mycobacterium tuberculosis.* ***Cell Microbiol***. 10:1027–1039.

Toyama A, Kurashiki E, Watanabe Y, Takeuchi H, Harada I, Aiba H, Lee BJ and Kyogoku Y. 1991. Ultraviolet resonance Raman spectra of cyclic AMP receptor protein: structural change induced by cyclic AMP binding and the conformation of protein bound cyclic AMP. ***J. Am. Chem. Soc.*** 113:3615–3616.

Tuberculosis Program. http://www.who. int/gtb/

Tundup S, Akhter Y, Thiagarajan D and Hasnain SE. 2006. Clusters of PE and PPE genes of *Mycobacterium tuberculosis* are organized in operons: evidence that PE Rv2431c is co-transcribed with PPE Rv2430c and their gene products interact with each other. ***FEBS Lett.*** 580:1285–1293.

Tutar Y. 2008. Syn, anti, and finally both conformations of cyclic AMP are involved in the CRP-dependent transcription initiation mechanism in *E. coli* lac operon. ***Cell Biochem Funct***. 26:399-405.

Ulrichs T, and Kaufmann SH. 2006. New insights into the function of granulomas in human tuberculosis. ***J. Pathol.*** 208:261–269.

Vaney MC, Gilliland GL, Harman JG, Peterkofsky A, Weber IT. 1989. Crystal structure of a cAMP-independent form of catabolite gene activator protein with adenosine substituted in one of two cAMP-binding sites. ***Biochemistry*** 28:4568–4574.

Vergne I, Chua J, Singh SB, and Deretic V. 2004. Cell biology of *Mycobacterium tuberculosis* phagosome. ***Annu Rev Cell Dev Biol***. 20:367–394.

Vollack KU, Hartig E, Körner H and Zumft WG. 1999. Multiple transcription factors of the FNR family in denitrifying *Pseudomonas*

stutzeri: characterization of four fnr-like genes, regulatory responses and cognate metabolic processes. ***Mol. Microbiol.*** 31:1681-1694.

Wayne LG and Sohaskey CD. 2001. Non-replicating persistence of *Mycobacterium tuberculosis*. ***Annu Rev Microbiol.*** 55:139-63.

Weber IT and Steitz TA. 1987. Structure of a complex of catabolite gene activator protein and cyclic AMP at 2.5 A resolution ***J.Mol.Biol.***. 198: 311- 26.

Weissborn AC, Liu Q, Rumley MK and Kennedy EP. 1994. UTP: alpha-D-glucose-1-phosphate uridylyltransferase of *Escherichia coli*: isolation and DNA sequence of the galU gene and purification of the enzyme. ***J. Bacteriol.*** 176:2611–2618.

Weston A, Stern RJ, Lee RE, Nassau PM, Monsey D, Martin SL, Scherman MS, Besra GS, Duncan K and McNeil MR. 1997. Biosynthetic origin of mycobacterial cell wall galactofuranosyl residues. ***Tubercle and Lung Disease*** 78:123–131.

WHO Report 2008, Global tuberculosis control - surveillance, planning, financing.

Won HS, Seo MD, Ko HS, Choi WS and Lee BJ. 2008. Interdomain interaction of cyclic AMP receptor protein in the absence of cyclic AMP. ***Biochem. J.***143:163–167.

Won HS, Lee YS, Lee SH and Lee BJ. 2009. Structural overview on the allosteric activation of cyclic AMP receptor protein. ***Biochim Biophys Acta.*** 1794:1299-308

World Health Organization, 2006. Tuberculosis Fact Sheet No. 104 – Global and Regional Incidence. World Health Organization, Geneva.

World Health Organization, 2009. Tuberculosis Fact Sheet, Global and Regional Incidence. World Health Organization, Geneva. http://www.who. int/gtb/

World Health Organization. 1993. The World Health Organization Global

World Health Organization. 2003. The World Health Organization Global

Yabu K, Kaneda S and Ochiai T. 1985. Relationship between beta-lactamase activity and resistance to beta-lactam antibiotics in *Mycobacterium smegmatis*. ***Microbiol. Immunol.*** 29:803–809.

Yang JT, Wu CS and Martinez HM. 1986. Calculation of protein conformation from circular dichroism. ***Methods Enzymol.*** 130:208–269.

Yellaboina S, Ranjan S, Chakhaiyar P, Hasnain SE, 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.

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

Youn H, Kerby RL, Koh J and Roberts GP. 2007. A C-helix residue, Arg-123, has important roles in both the active and inactive forms of the cAMP receptor protein, ***J. Biol. Chem.*** 282:3632–3639.

Youn, H, Kerby RL, Conrad M, Roberts GP. 2006. Study of highly constitutively active mutants suggests how cAMP activates cAMP receptor protein. ***J. Biol. Chem.*** 281:1119–1127.

Vitae of Candidate

YUSUF AKHTER
European Molecular Biology Laboratory,
c/o DESY, Notkestraße 85,
22603 Hamburg, Germany
E-mail: yusuf.akhter@embl-hamburg.de
yusuf.akhter@gmail.com

EDUCATIONAL QUALIFICATION-

Ph.D. (Department of Biochemistry, University of Hyderabad) 2004-till date
Work performed at: Center for DNA Fingerprinting & Diagnostics, Hyderabad, INDIA / European Molecular Biology Laboratory, Hamburg, Germany
Title of thesis: Studies on cAMP Receptor Proteins from mycobacteria
Supervisors: Prof Seyed E. Hasnain, University of Hyderabad, Hyderabad, India/Dr. Matthias Wilmanns, EMBL, Hamburg.

Master of Science (M.Sc), 2004
Department of Biotechnology,
Jamia Hamdard (Hamdard University), New Delhi, INDIA
Specialization: Biotechnology
First Division with 79.40% Marks: Second rank in University

Bachelor of Science (B.Sc), 2002
University of Lucknow, Lucknow, INDIA
Subjects: Botany, Zoology and Chemistry
First Division with 67.83% Marks: First rank in college

Higher Secondary school (10+2), 1998
Central Board of Secondary Education (CBSE), New Delhi, India
Subjects: Physics, Chemistry, Mathematics, Biology, English
First Division with 79.20% marks: First rank in school

AWARDS, HONOUR AND NATIONAL TESTS-

1. **The Bill and Melinda Gates Foundation Global Health Travel Award** to attend the Keystone Symposia **"Tuberculosis: Biology, Pathology and Therapy"** to be held at Keystone Resort, Keystone, Colorado, USA on Jan 25 - Jan 30, 2009. (55 awardees worldwide for year 2009)

2. **DAAD (German Academic Exchange Service) long-term Doctoral Fellowship** to work at European Molecular biology Laboratory, Hamburg, Germany, 2007–2009. (22 awardees from India in 2007)

3. Selected in worldwide competition to attend **"MEETING OF NOBEL LAUREATES (Physiology & Medicine) AND Young researchers at Lindau, Germany during 1-6 July 2007"** (23 awardees from India for 2007 meeting)

4. **Department of Science and technology (Govt. of India) travel award** to attend Lindau Meeting of Nobel laureates and Young researchers 2007, Germany.

5. **ISBC travel grant winner** to attend **"International School of Biological Crystallization 2009 at Granada, Spain (18-22 May 2009)"**.

6. Qualified **National Eligibility Test (NET),** for junior research fellowship conducted by **CSIR/ UGC** June. 2004 (stood in top 20% qualified students)

7. Qualified **GATE 2004 (Graduate Aptitude Test for Engineering)** conducted by **IIT Delhi**, New Delhi India with **91.43** percentile.

8. Recipient of **Tasmia Merit Scholarship** for best academic performance at Masters level university exams in Jamia Hamdard (2003-2004)

9. Recipient of **"Certificate of Merit"** by Central Board for Secondary Examination, New Delhi for securing 0.1 Percent top ranker position in CHEMISTRY (97%) at 10+2 level, 1998.

Publications

Manuscripts Published

1. Pramod Kumar, Dhananjay C. Joshi, Mohd Akif, **Yusuf Akhter,** Seyed E. Hasnain and Shekhar C. Mande. 2009. Crystal Structure of apo-Cyclic AMP Receptor Protein of *M. tuberculosis* and Normal Mode Analyses reveal an Elegant Mechanism of Allostery induced upon cAMP binding. ***Biophysical Journal*** (In press)

2. Yusuf Akhter, Sailu Yellaboina, Aisha Farhana, Akash Ranjan, Niyaz Ahmed and Seyed E Hasnain (2008). Genome scale portrait of cAMP Receptor Protein-Regulons in Mycobacteria points to their role in pathogenesis. ***Gene*** 407:148-58.

3.Yusuf Akhter, Smanla Tundup and Seyed E. Hasnain (2007). Novel Biochemical Properties of a CRP/FNR Family Transcription Factor from *Mycobacterium tuberculosis*. ***Int. J. Med. Micro.*** 297: 451–457

4.Yusuf Akhter, Irshad Ahmed, S. Manju Devi and Niyaz Ahmed (2007). The co-evolved *Helicobacter pylori* and gastric cancer: Trinity of bacterial virulence, host susceptibility and lifestyle. ***Infect Agent Cancer,*** 4, 2:2 (Highly Accessed Article)

5. S Manjulata Devi, Irshad Ahmed, Paolo Francalacci, M Abid Hussain, **Yusuf Akhter**, Ayesha Alvi, Leonardo A Sechi, Francis Megraud and and Niyaz Ahmed (2007). Ancestral European roots of *Helicobacter pylori* in India. ***BMC Genomics.*** 8**:**184

6.Mohd. Akif, **Yusuf Akhter,** Seyed E. Husnain and Shekhar C. Mande. (2006). Crystallization and preliminary X-ray crystallographic studies of *Mycobacterium tuberculosis* CRP/FNR family transcriptional regulator. ***Acta Crystallography F,*** 62:873-5

7.Smanla Tundup, **Yusuf Akhter,** D Thiagarajan, and Seyed E. Hasnain. (2006). Clusters of PE and PPE genes of Mycobacterium tuberculosis are organized in operons: evidence that PE Rv2431c is co-transcribed with PPE Rv2430c and their gene products interact with each other. ***FEBS Lett.,*** 580: 1285-93.

Manuscripts under communication

1. Yusuf Akhter*, Christian Poulsen* et al, Proteome wide analysis of Prokaryotic Ubiquitin like Protein (PUP) substrates and its diverse functional implication (Under revision ***Molecular Systems Biology***)

2. Yusuf Akhter et al, 2.9A structure of cAMP receptor protein-cAMP-DNA from *Mycobacterium tuberculosis* reveals novel cAMP binding site (Communicated)

***Equal contribution**

PERSONAL DATA

Father's Name	:	AKHTER HUSSAIN KHAN
Date of Birth	:	7th JUNE 1979
Nationality	:	INDIAN
Sex	:	MALE
Marital Status	:	SINGLE
Languages Known	:	ENGLISH, GERMAN, HINDI AND URDU
Permanent Address	:	HASHMI MANZIL SAKRAWAL (WEST), TANDA, AMBEDKAR NAGAR, 224190 UTTAR PRADESH (INDIA)

Publications

Available online at www.sciencedirect.com
ScienceDirect
Gene 407 (2008) 148–158
GENE
www.elsevier.com/locate/gene

Genome scale portrait of cAMP-receptor protein (CRP) regulons in mycobacteria points to their role in pathogenesis

Yusuf Akhter [a,e,1], Sailu Yellaboina [b], Aisha Farhana [a], Akash Ranjan [b], Niyaz Ahmed [c], Seyed E. Hasnain [d,e,*]

[a] *Laboratory of Molecular and Cellular Biology, CDFD, Hyderabad, 500076, India*
[b] *Laboratory of Computational and Functional Genomics and CDFD-Sun Microsystems Centre of Excellence in Medical Bioinformatics, CDFD, Hyderabad, 500076, India*
[c] *Pathogen Evolution Laboratory, CDFD, Hyderabad, 500076, India*
[d] *Jawaharlal Nehru Centre for Advanced Scientific Research, Jakkur, Bangalore, 560012, India*
[e] *University of Hyderabad, Hyderabad, 500046, India*

Received 23 March 2007; received in revised form 2 October 2007; accepted 5 October 2007
Available online 22 October 2007
Received by G. Pesole

Abstract

cAMP Receptor Protein (CRP)/Fumarate Nitrate Reductase Regulator (FNR) family proteins are ubiquitous regulators of cell stress in eubacteria. These proteins are commonly associated with maintenance of intracellular oxygen levels, redox-state, oxidative and nitrosative stresses, and extreme temperature conditions by regulating expression of target genes that contain regulatory cognate DNA elements. We describe the use of informatics enabled comparative genomics to identify novel genes under the control of CRP regulator in *Mycobacterium tuberculosis* (*M.tb*). An inventory of CRP regulated genes and their operon context in important mycobacterial species such as *M. leprae, M. avium* subsp. *paratuberculosis and M. smegmatis* and several common genes within this genus including the important cellular functions, mainly, cell-wall biogenesis, cAMP signaling and metabolism associated with such regulons were identified. Our results provide a possible theoretical framework for better understanding of the stress response in mycobacteria. The conservation of the CRP regulated genes in pathogenic mycobacteria, as opposed to non-pathogenic ones, highlights the importance of CRP-regulated genes in pathogenesis.

Keywords: Rv3676; *Mycobacterium tuberculosis*; cAMP signaling; CRP-regulated genes

1. Introduction

Cyclic AMP receptor proteins (CRP)–FNR superfamily of transcription factors regulate a diversity of physiological processes in bacteria. These proteins predominantly regulate intracellular reactions related to carbon, sulfur and nitrogen metabolism, denitrification, nitrogen fixation, aerobic and anaerobic respiration and expression of virulence genes in response to a variety of environmental and metabolic signals (Korner et al., 2003; Green et al., 2001). With more than 370 family members, these DNA-binding proteins primarily function as positive regulators except that a few from this family also act as negative regulator of transcription. FNR protein from *E. coli* acts both as activator as well as repressor of transcription depending upon the distance between the binding site and the transcription start-points (Barnard et al., 2003). Some of the distinguishing features of CRP/FNR members include the presence of a nucleotide binding domain and a helix turn helix DNA binding

Abbrevations: CRP, Cyclic AMP Receptor Protein; FNR, Fumarate Nitrate Reductase Regulator; *M.tb*, *Mycobacterium tuberculosis*; *M. bovis*, *Mycobacterium bovis*; *M. leprae*, *Mycobacterium leprae*; *M. smegmatis*, *Mycobacterium smegmatis*, RPS-BLAST, Reversed Position Specific Basic Local Alignment Tool.

* Corresponding author. University of Hyderabad, Gachibowli, Hyderabad, 500 046, India. Tel.: +91 40 23010121; fax: +91 40 23011090.
E-mail address: vc@uohyd.ernet.in (S.E. Hasnain).
[1] Present address: EMBL Hamburg c/o DESY, Notkestraße 85, 22603 Hamburg, Germany.

0378-1119/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2007.10.017

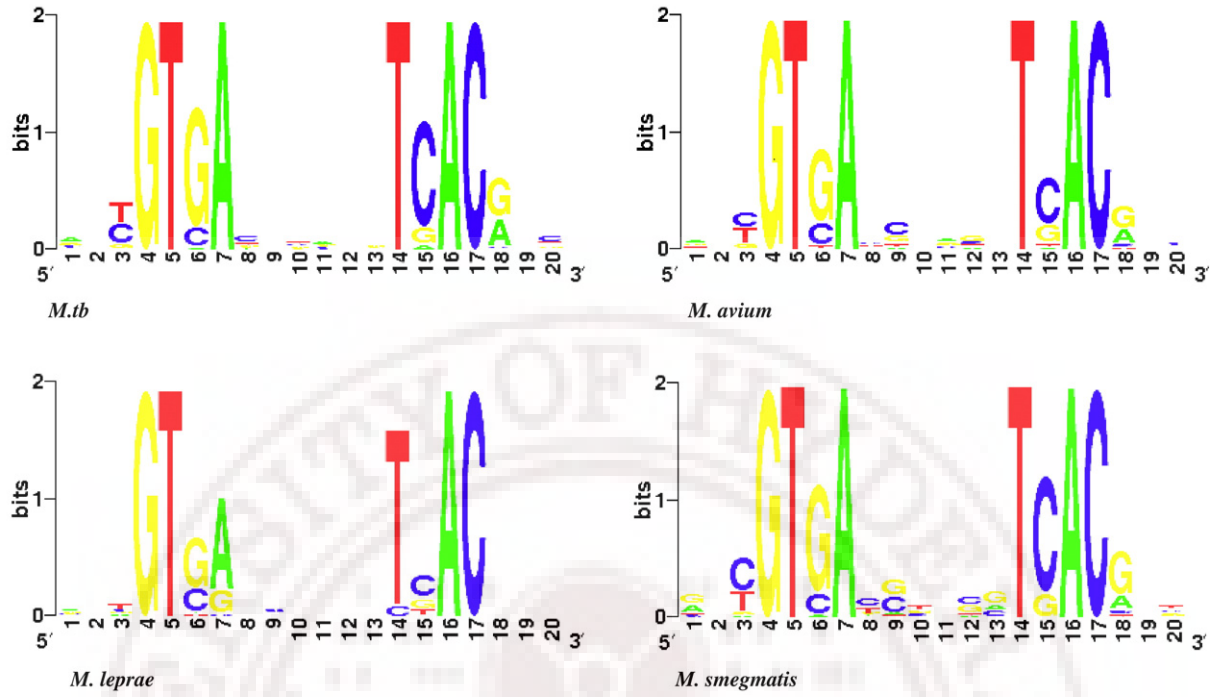


Fig. 1. Sequence logo of the predicted CRP-binding sites in *M.tb*, *M. avium*, *M. leprae* and *M. smegmatis*.The height of each stack of letters represents the degree of sequence conservation measured in bits. The height of each letter within a stack is proportional to its frequency at that position in the binding site. The letters are sorted with the most frequent on top. This sequence logo was generated using the online Weblogo programme (http://weblogo.berkeley.edu/).

motif at the N-and C-terminal, respectively. The classical cAMP-binding domain (Schultz et al., 1991) is a versatile structure that has evolved to accommodate different functional specificities in response to a wide range of signals (Korner et al., 2003; Green et al., 2001). The best-studied prototype proteins from this superfamily are CRP and FNR of *E. coli* that regulate expression of numerous genes in response to starvation and hypoxia, respectively.

In case of *E. coli* CRP, cAMP acts as a secondary signal and binds to the CRP protein to form cAMP-CRP complex. This complex then binds to the promoters carrying specific DNA elements related to the consensus motif TGTGANNNNNNT-CACA (Berg and von Hippel, 1988) thereby regulating the expression of the downstream sequence. However, in case of FNR protein, redox states of bound metals act as the signal and activate the FNR that further binds to the promoter containing specific DNA elements (consensus TTGATNNNNATCAA) and regulate the expression of the genes (Spiro, 1994).

M.tb H37Rv contains a single CRP/FNR homologue coded by ORF *Rv3676* (Cole et al., 1998). Orthologue of *Rv3676* (*M. tb*-CRP) is present in all sequenced and unfinished mycobacterial genomes. Recently, the DNA binding and cAMP binding properties of recombinant Rv3676 were described based on computational predictions (Bai et al., 2005). This study was extended by us (Akhter et al., 2007) to provide insights into unusual and novel biochemical properties of Rv3676 protein. We also reported the crystallization and preliminary X-ray diffraction data of this CRP/FNR regulator (Akif et al., 2006). Previously, Mattow et al. (2001), while comparing the proteome profiles of *M.tb* and *M. bovis* BCG, observed differences in electrophoretic mobility of CRP proteins (Mattow et al., 2001). Subsequently, such a mobility shift was attributed to point mutations in both the DNA and cAMP binding domains in CRP of *M. bovis* BCG. These mutations were ascribed to the impaired DNA binding activity of *M. bovis* BCG-CRP in comparison to *M.tb*-CRP (Spreadbury et al., 2005). This might suggest these mutations to be one of the contributing factors in attenuation of the virulence of *M. bovis* BCG. In other studies, *M.tb* deletion mutants corresponding to Rv3676 (*M.tb*-CRP) revealed growth defects in laboratory cultures of bone marrow derived macrophages and in mouse models of tuberculosis (Rickman et al., 2005).

Apart from some initial studies not much is known about the population wide repertoire of CRP-regulated genes *per-se* in different clinical settings and their cellular functions. We describe CRP regulated novel genes of *M.tb* and predict their abundance and operon context in the genomes of *M. leprae, M. avium* subsp. *paratuberculosis* and *M. smegmatis.* We also attempted to identify common genes across mycobacteria, which could be regulated by CRP.

2. Materials and methods

2.1. Source of genome sequence

Published and annotated genome sequences of *M.tb, 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/

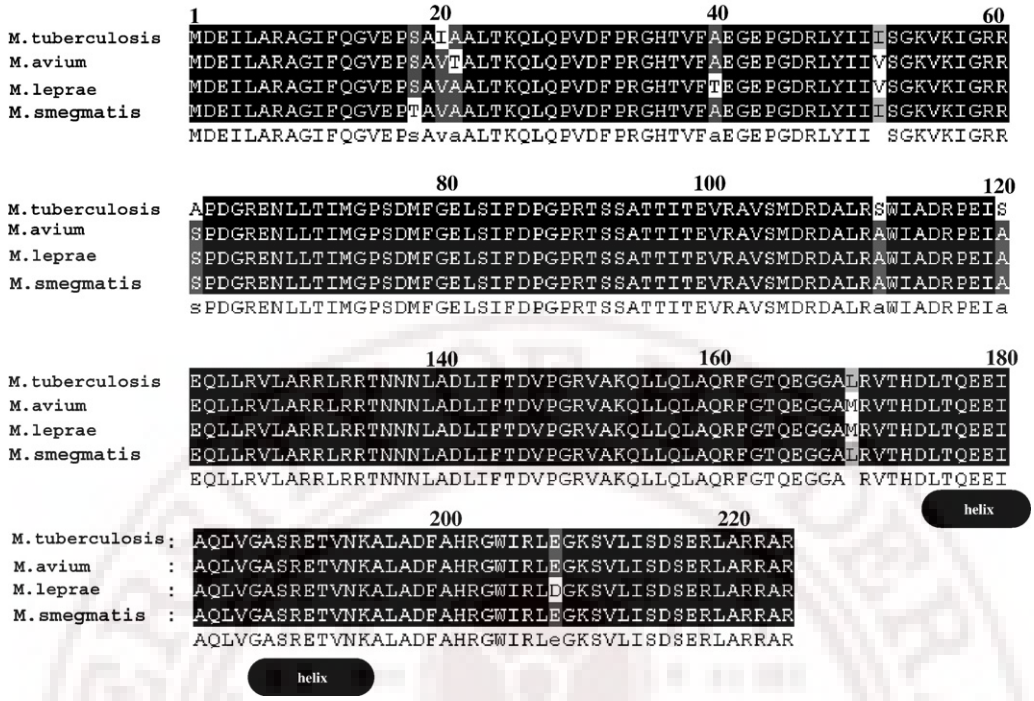


Fig. 2. Alignment of CRP orthologues from different species of mycobacteria reveals a highly conserved DNA binding domain. *M.tb*-CRP sequence is used as reference sequence and compared with other orthologues. Arial black shadow shows identity and the gray show similarity. Two helices (labeled as helix) are part of helix turn helix that assists in CRP box recognition.

tigrscripts/CMR/CmrHomePage.cgi). Sequences corresponding to *M.tb*-CRP-binding sites identified previously (Bai et al., 2005; Rickman et al., 2005) were also obtained. A local archive of genome sequences was assembled, curated and stacked. This resource was queried for known and putative homologies and to build consensus and alignments at a genome wide interface.

2.2. Prediction of CRP-binding sites

CRP-binding site recognition profile was calculated by positional Shannon relative entropy method as described earlier (Yellaboina et al., 2004; Prakash et al., 2005, Yellaboina et al., 2006). Sequences corresponding to *M.tb*-CRP-binding sites identified previously (Bai et al., 2005; Rickman et al., 2005) were used to generate input profile. The binding site profile thus generated was used to scan upstream sequences of all the genes of each of the mycobacterial genomes. The score of each site was calculated as the sum of the respective positional Shannon relative entropy of each of the four possible bases. A maximally scoring site was selected from the upstream sequence of each of the gene we looked at. The lowest score among the input binding sites was considered as cut-off score. The sites scoring higher than the cut-off value were predicted as potential binding sites that conform to the consensus sequence. For each of predicted regulon (*M.tb, M. avium, M. leprae* and *M. smegmatis*) sequence logo was generated by using Weblogo server (Crooks et al., 2004). The height of each stack of letters represented the degree of sequence conservation measured in bits. The height of each letter within a stack is proportional to its frequency at that position in the binding site. The letters are sorted with the most frequent on top (Fig. 1).

2.3. Prediction of operons and function annotation

Genes transcribed co-directionally and downstream to the predicted binding sites in *M. leprae, M. avium* subsp. *paratuberculosis* and *M. smegmatis* were chosen as potentially co-regulated genes (operons). These were used according to one or more of the following criteria (Yellaboina et al., 2004; Tundup et al., 2006; Yellaboina et al., 2006) [1] co-directionally transcribed orthologous gene pairs, conserved in at least 4 genomes; [2] genes belonging to the same cluster of orthologous gene function category whose intergenic distance is less than 200 base pairs; [3] if the first three letters in the gene names are identical (gene names for putative genes were assigned from COG database); and [4] if the intergenic distance is less than 90 base pairs.

We predicted the function of regulated genes if not done so previously. RPS-BLAST search against conserved domain database (CD search) at NCBI server (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) was used to infer the annotation. In principle, RPS-BLAST is a tool opposite of PSI-BLAST which searches the profiles against a database of sequences, hence the 'Reverse'. RPS-BLAST uses a BLAST-like algorithm, finding single-or double-word hits and then performing an ungapped extension on these candidate matches. If a sufficiently high-scoring ungapped alignment is produced, a gapped extension is performed and those (gapped) alignments with sufficiently low expect value are reported (Marchler-Bauer et al., 2002).

3. Results and discussion

3.1. CRP regulators from mycobacteria have conserved DNA binding domains

We aimed at exploiting the use of CRP binding sites in *M.tb* for the prediction of regulons in mycobacteria. DNA binding helix-turn-helix (HTH) motif was predicted at EMBOSS server (http://bioweb.pasteur.fr/seqanal/interfaces/helixturnhelix-simple.html). The sequence (176-LTQEEIAQLVGASRETVNKALA-196) from *M.tb*-CRP was identified as the likely HTH motif involved in DNA binding as it elicited maximum score in our analyses (data not shown). This HTH motif from *M.tb*-CRP was selected and compared with the orthologues. Our bioinformatics analysis of CRP regulators from different mycobacterial species was validated through CLUSTAL W based alignments, which revealed high-level identity in their DNA binding domains (Fig. 2). In the comparative amino acid sequence analyses with *M.tb*-CRP, we found that CRP orthologues from *M. avium, M. leprae* and *M. smegmatis* to be 96%, 96% and 97% identical respectively.

3.2. Novel CRP binding sites in M.tb genome

In earlier computational predictions of CRP-regulon, 73 binding elements in *M.tb* genome were observed (Bai et al., 2005). Of the top 44 binding sites identified by us 25 were reported previously (Bai et al., 2005). Our analyses thus

Table 1
Predicted CRP binding sites in *M. tuberculosis* genome

Score	Position	Binding site	Gene	Synonym	Product
3.96721	−287	AATGTGATCTAGGTCACGTG	frdA	Rv1552	Fumarate reductase
3.92739	−176	ATTGTGAGTTGGATCACGTT	sucC	Rv0951	Succinyl-CoA synthetase subunit beta
3.92564	−156	GCCGTGAGATTCGTCACGTC	–	Rv1810	Hypothetical protein
3.92181	−11	CGTGTGAACGATGTCACGCC*	galU	Rv0993	UTP-glucose-1-phosphateU-transferase
3.90978	−4	ACTGTGACGCCCGTCACAAC*	–	Rv0104	cAMP binding protein
3.90294	−182	GCTGTGAGCCGAATCACGAC*	–	Rv3645	Membrane linked adenylate cyclase
3.90045	−378	ATCGTGAAGCCGTTCACGCT*	glnD	Rv2918c	PII uridylyl-transferase
3.89854	−160	GTTGTGACGGGCGTCACAGT*	ctpB	Rv0103c	Cation-transporter p-type ATPaseB
3.89686	−41	CGCGTGACATGTGTCACATG	PE_PGRS44	Rv2591	PE-PGRS
3.89517	−141	AGTGTGATTTACATCACATA	–	Rv2700	Secreted alanine rich protein
3.89328	−44	GTCGTGATACGACTCACGCG	echA6	Rv0905	Enoyl-CoA hydratase
3.89246	−169	ACCGTGACGCCCCTCACGGC*	fadD21	Rv1185c	Acyl-CoA synthase
3.89159	−136	AGTGTGAACAAGCTCACATG	–	Rv0885	Hypothetical protein
3.88671	−92	CATGTGAGCTTGTTCACACT	serC	Rv0884c	Phosphoserine aminotransferase
3.88342	−110	GGCGTGACATCGTTCACACG*	–	Rv0992c	5-formyltetrahydrofolatecyclo-ligase
3.88095	−120	AACGTCACATGCGTCACGGC	–	Rv1581c	Probable phiRv1 phage protein
3.87653	−154	AACGTGATCCAACTCACAAT	–	Rv0950c	Hypothetical protein
3.87318	−117	TATGTGATGTAAATCACACT	–	Rv2699c	Hypothetical protein
3.86965	−3	CGCGTGAGTCGTATCACGAC*	accD3	Rv0904c	Acetyl-CoAcarboxylaseCo-transferase
3.86178	−90	GCTGTCAAATCCGTCACGAA	–	Rv2336	Hypothetical protein
3.86153	−80	GATGTGACTCAAGTGACACG	–	Rv1159	Conseved transmembrane protein
3.85583	−306	TGCGTGAGGAGCCTCACGGC*	–	Rv2650c	PhiRv2 prophage protein
3.85125	−70	CGTGTCACTTGAGTCACATC	–	Rv1158c	Hypothetical ALA,PRO-rich protein
3.84869	−70	ATCGTGACTTTGCTGACGTG	–	Rv0019c	Hypothetical protein
3.84804	—216	GCGGTGATCGGCGTCACGCC	PE24	Rv2408	PE
3.84665	−177	GACGTCAACCAGTTCACGCT*	mmaA3	Rv0643c	Methoxy mycolic acid synthase
3.83995	−80	ATGGTGATCTAGTTCACGAA	–	Rv1230c	Membrane protein
3.83989	−236	CGCGTGACTGAAATCACAAC	–	Rv1566c	Possible inv protein
3.83976	−229	CGTGTGACAGCTGTGACGGT	wag22	Rv1759c	PE-PGRS
3.82802	−367	ACCGTCACAGCTGTCACACG	–	Rv1760	Hypothetical protein
3.82654	−86	GATGTGATGCACTTGACATC	PE33	Rv3650	PE
3.82347	−245	GTGGTGAGCTGGTTCACACC*	–	Rv2407	Ribonuclease Z
3.81909	−35	GGTGTGAACCAGCTCACCAC	–	Rv2406c	Hypothetical protein
3.81879	−341	TTCGTGAGGCGTGTGACGAA*	–	Rv3113	Phosphatase
3.81542	−7	GACGTCATGGATTTCACGAC*	fadE27	Rv3505	Acyl-CoA dehydrogenase
3.81481	−50	ACCGTGACATCGATGACAGC*	–	Rv3031	Glycoside hydrolase
3.81167	−168	ACGGTGACAGCGCTCACGGT*	moaX	Rv3323c	MOAD-MOAE fusion protein
3.81104	−272	TAGGTGACCAAACTCACGCT	PPE11	Rv0453	PPE
3.80557	−318	CCTGTGACCGGTGTCACTGC*	ephA	Rv3617	Probable epoxide hydrolase
3.80304	−136	CGTGTGACCAAACTCACCGC	PE15	Rv1386	PE
3.79969	−175	ATCGTGACACCGGTAACGGC*	–	Rv0520	Methyltransferase/methylase
3.79829	−371	GATGTGACCGTGGTAACGTA*	pdhC	Rv2495c	Dihydrolipoamide acetyltransferase
3.79647	−69	AGTGTAACGCATATCACGTG	–	Rv0452	Transcriptional regulatory protein
3.7954	−288	ATAGTGACGGCCGTCACAGC*	–	Rv3690	Conserved membrane protein

*New identified sites

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

Score	Position	Binding site	Gene	Synonym	Product
4.04838	−353	AGTGTGACCTCCGTCACATC	–	MAP3737	Hypothetical protein
4.02131	−137	AGTGTGATCTAGGTGACGTG	–	MAP3267c	Hypothetical protein
4.00972	−82	GGTGTGATTTACCTCACACC	–	MAP2817	Hypothetical protein
4.00494	−116	GGTGTGAGGTAAATCACACC	–	MAP2816c	Hypothetical protein
3.99554	−383	GACGTGACGAGGTTCACGGC	fadD29	MAP3284c	FadD29
3.99383	−56	ACTGTGAGATTGCTCACAGT	–	MAP3671	Hypothetical protein
3.98389	−45	GCCGTGATACGACTCACGAG	echA6	MAP0840	Enoyl-CoA hydratase
3.98388	−62	GTTGTGAGCCCGCTCACAGG	–	MAP0174	Hypothetical protein
3.98224	−37	GCCGTGATTCAGTTCACACC	–	MAP0227c	Hypothetical protein
3.9796	−25	CGTGTGACCAACCTCACATT	–	MAP2149c	Hypothetical protein
3.97915	−111	TTTGTGAGCCGCTTCACACC	–	MAP2220	Ribonuclease Z
3.97805	−101	AACGTCACAAACGTCACGGT	lpqR	MAP0670	LpqR
3.97759	−373	GGCGTCACCGAGGTCACGCT	–	MAP0727	Hypothetical protein
3.97627	−3	CTCGTGAGTCGTATCACGGC	accD3	MAP0839c	AccD3
3.96913	−37	GGTGTGAAGCGGCTCACAAA	–	MAP2219c	Hypothetical protein
3.96489	−41	AACGTGAACGCCGTGACGGC	pepC	MAP0632	Putative aminopeptidase 2
3.96428	−69	GACGTCAACCAGTTCACGCT	recC	MAP4094c	RecC
3.96198	−293	TTTGTGATTGATCTCACGGA	–	MAP1418c	Hypothetical protein
3.96178	−399	AACGTCACCCAACTCACGAG	–	MAP1601	Hypothetical protein
3.95747	−40	AACGTGACGGGTGTGACGGA	pknD	MAP3387c	PknD
3.95283	−187	TCCGTCAACCGCGTCACGTC	echA21	MAP0249c	Enoyl-CoA hydratase
3.94986	−4	ATCGTGATAGCGGTGACGAT	–	MAP2875	Hypothetical protein
3.9487	−87	CGGGTGACCCCGGTCACGCT	ephD	MAP1955c	Short chain dehydrogenase
3.94584	−157	AGCGTGACCGGGGTCACCCG	dlaT	MAP1956	Dihydrolipoamide acyltransferase
3.94403	−127	GGTGTCAGCATGGTCACATA	–	MAP2018c	Hypothetical protein
3.94196	−101	GACGTGACCTCGATGACACG	–	MAP0097c	Hypothetical protein
3.93609	−119	ATTGTGATTTGGATCACCTA	sucC	MAP0896	Succinyl-CoA synthetase subunit beta
3.93595	−35	CACGTCACCTAGATCACACT	hsp18_3	MAP3268	hsp18_3
3.93244	−62	ATCGTCACCGCTATCACGAT	fadD13	MAP2874c	FadD13
3.93233	−1	ATGGTGAACAAATTCACGAC	lpqL_1	MAP3906	LpqL_1
3.92976	−111	AACGTGATTGGCATGACGAG	–	MAP1597	Hypothetical protein
3.92919	−96	CGGGTGAAGCGGGTCACGAC	–	MAP0683	Hypothetical protein
3.92773	−32	ACCGTCATGGAGATCACGGG	fadA3	MAP2407c	Acetyl-CoA acetyltransferase
3.92509	−163	GACGTCACGGCTTTCACGGC	–	MAP1333	Hypothetical protein
3.92501	−285	GCGGTGATCTACGTGACGCC	–	MAP1644	Hypothetical protein
3.92295	−78	CGCGTGAGTAGGGTGACATC	–	MAP2860	Hypothetical protein
3.92295	−392	CGCGTGAGTAGGGTGACATC	–	MAP2861	Short chain dehydrogenase
3.92038	−3	GCCGTGATGGACCTGACGCA	–	MAP2928c	Short chain dehydrogenase
3.91804	−248	TCGGTGACCCGTTTCACGCC	–	MAP1475	Hypothetical protein
3.9169	−65	CGGGTGACCCGGCTCACGGT	valS	MAP2271c	valyl-tRNA synthetase
3.90643	−244	ATTGTGACCGGCCTCACTGC	–	MAP0948	Hypothetical protein
3.90494	−269	ACTGTTAGTTACATCACACC	–	MAP1693c	Hypothetical protein
3.90214	−89	TGGGTGACCCTTGTCACAGC	–	MAP0725	Hypothetical protein
3.90027	−318	GGCGTGAACAACCTCACCGG	–	MAP2531	Hypothetical protein
3.89953	−189	ACCGTTATCGTTGTCACGCA	–	MAP2500	Hypothetical protein
3.89821	−366	ATCGTAAAGCCGTTCACGCT	glnD	MAP2986c	PII uridylyl-transferase
3.89757	−325	AGCGTGACCTCGCTTACACC	–	MAP1558c	Hypothetical protein
3.89655	−93	ACTGTGAATTAGTTAACAAG	fadE24	MAP3188	FadE24
3.89528	−104	AAGGTCAAGACCGTCACGTC	fadE9	MAP4214c	FadE9
3.89286	−56	TACGTCACCGGGGTGACGTT	–	MAP2780	Hypothetical protein
3.89245	−294	TGCGTGATGCCCTTGACGAA	fadD2	MAP3714	acyl-CoA synthase
3.89228	−43	AGTGTGAGGTGTATTACACA	–	MAP3952	Hypothetical protein
3.89158	−96	GGTGTGACGAGTTTCACTAC	fbpC1	MAP0217	FbpC1
3.8888	−365	TTTGTGACTCACCTCACTTG	sodA	MAP0187c	SodA
3.8888	−25	TTTGTGACTCACCTCACTTG	–	MAP0188c	Hypothetical protein
3.88826	−369	GCGGTGATCTGGCTGACGTG	embR_1	MAP0230	EmbR_1
3.888	−3	GAGGTGACGCAATTGACGCC	atsG	MAP3791c	AtsG
3.88752	−117	ATTGTTAGCGCGGTCACAGA	–	MAP2060c	Hypothetical protein
3.88613	−208	GGCGTCACCCTGCTGACGGT	–	MAP0053c	Hypothetical protein
3.8858	−8	TATGTGATTTGTATAACGCA	–	MAP3944c	Hypothetical protein
3.88484	−119	GATGTCAGGGTGGTGACATG	parB	MAP4344c	ParB
3.88483	−19	GCAGTGAGGCCGGTCACAAT	–	MAP0947c	Hypothetical protein
3.884	−397	CAAGTGAGGTGAGTCACAAA	–	MAP0189	Hypothetical protein

Table 2 (*continued*)

Score	Position	Binding site	Gene	Synonym	Product
3.88344	6	GATGTGACCATTGTGACCAG	smc	MAP2990c	Smc
3.88298	−157	GCCGTCACCGACGTCACCCC	mbtH_3	MAP2169c	MbtH_3
3.88268	−281	GGTGTGATGTAACTAACAGT	papA2	MAP1694	PapA2
3.88184	−385	TCTGTGATCAAGATTACGCC	–	MAP3628c	Hypothetical protein
3.88184	−3	TCTGTGATCAAGATTACGCC	–	MAP3629c	Hypothetical protein
3.87809	−388	AATGTGACCAGCTTGACCAC	–	MAP1708	Hypothetical protein
3.87579	−147	GGGGTGACGTCGCTGACGGC	lipA	MAP1959	Lipoyl synthase
3.87578	−103	AGTGTCATACAAATCACCTC	fadA6_4	MAP3337	FadA6_4
3.87541	−93	TAGGTGATCCAAATCACAAT	–	MAP0895c	Hypothetical protein

highlighted 19 new *M.tb*-CRP binding sites upstream of various operons in *M.tb* genome (Table 1). The CRP-binding element from fumarate reductase (*Rv1552*) received the highest score (Table 1). While this DNA element was also reported as a potential CRP-binding site in earlier reports (Bai et al., 2005; Rickman et al., 2005, Spreadbury et al., 2005), we recently provided experimental evidence for specific DNA: protein interaction between this DNA element and recombinant purified *M.tb*-CRP (Akhter et al., 2007). We also showed that recombinant *M.tb*-CRP is not able to compete for binding if mutant oligodeoxynucleotides were used in electrophoretic mobility shift assays (Akhter et al., 2007). Also, Spreadbury et al (2005) proposed some potential genes as members of the CRP-regulon. There are obvious overlaps between the results from previous studies and the present one, but we were nonetheless able to locate some novel members of the CRP-regulon in *M.tb* genome. While previous studies (Bai et al., 2005, Spreadbury et al., 2005) utilized information from *E. coli* CRP-regulon, we in our study used only the available information from *M.tb* CRP-regulon.

Interestingly, most of the novel operons are functionally critical for *M.tb* and several of them are conserved (see below) within the pathogenic mycobacteria. For example *Rv0904* (AccD3) and *Rv0993* (galU) were observed among them and interestingly these enzymes were found as key factors during cell wall biogenesis. Another gene *Rv3645* (membrane linked adenylyl cyclase) was also found conserved in comparative regulon analyses. It is tempting to speculate that this adenylyl cyclase could be acting as a switch during cAMP based signaling in cellular stress and perhaps responsible for maintaining cAMP homeostasis in bacterial cells. *Rv2918* (gln D) and *Rv992* (putative 5-formyltetrahydrofolate cyclo-ligase) were also among the novel predicted boxes that were observed to be conserved in these regulons. These enzymes are also related to vital metabolic pathways of these pathogens.

3.2.1. Boxes associated with cell wall biogenesis

One of the novel-binding sites was located upstream of Rv0993 (gal U), which actually is a UTP: alpha-D-glucose-1-phosphate uridylyltransferase. It converts glucose-1-phosphate to UDP-glucose, that is of central importance in the synthesis of the components of the cell envelope of *E. coli* and in both galactose and trehalose metabolism (Weissborn et al., 1994). In case of *M.tb*, galactose is essential for the linking of the peptidoglycan and mycolic acid cell wall layers and is therefore essential for survival of mycobacteria (Weston et al., 1997). Another interesting box observed was the one associated with regulation of Rv3031, a conserved hypothetical protein actually belonging to glycoside hydrolase family of proteins. We further observed that Rv3032 is present in the same operon downstream to Rv3031, which was previously annotated as glycosyl transferase. Members of this family of enzymes transfer activated sugars (UDP, ADP, GDP or CMP linked sugars) to a variety of substrates, including glycogen, fructose-6-phosphate and lipopolysaccharides. These enzymes are directly involved in cell wall biogenesis. Next we found novel site upstream of Rv0643c, a methoxy mycolic acid synthase, involved in mycolic acid biosynthesis which is a key component of the mycobacterial cell wall (Kremer et al., 2000). Yet another novel CRP binding another site was located within Rv0904c (AccD3). This gene codes for putative acetyl CoA carboxylase carboxyl transferase, which catalyses the initial steps of fatty acid biosynthesis and was reported as a key enzyme in mycolic acid biosynthesis (Gande et al., 2004), an exclusive component of mycobacterial cell wall biogenesis.

3.2.2. Putative regulatory elements controlling 5′-3′ Cyclic Adenosine Monophosphate (cAMP) signaling

A CRP binding box upstream of Rv0104, a conserved hypothetical protein, was also identified. Upon sequence analysis, it was found to have a cyclic AMP binding domain at its C-terminal. This domain is very similar to the effector domain of CRP family proteins and cAMP binding domain (regulatory domain) of cAMP dependent protein kinases and therefore suggesting a role for this protein in cAMP mediated signaling in M.tb. We further observed that Rv0103c (probable cation-transporter) and Rv0104 share common CRP binding sites and probably represent a case of divergent gene expression.

We also located a CRP binding box upstream to the gene encoding membrane linked adenylyl cyclase (Rv3645). Adenylyl cyclases catalyze the production of cAMP, which further acts as a secondary signal. This protein, which was previously speculated to be involved in cAMP mediated signaling in mycobacteria (Linder and Schultz, 2003), is anchored to the membrane in M.tb. Gene regulation of adenylyl cyclase could be a case of feed back type of regulation for activation of CRP.

3.2.3. Other potential boxes

Novel binding site was also present in Rv0520, a methyl transferase believed to be involved in ubiquinone pathway. We

also observed a CRP binding site in Rv2699, as reported earlier (Bai et al., 2005). Sequence analysis of Rv2699 revealed similarity to methyl transferase proteins involved in ubiquinone pathway suggesting that the two enzymes with related function may be expressed under the control of same regulator.

Another important element observed within Rv2918c (GlnD-uridyl transferase) regulates the catalytic activity of glutamine synthetase (Garcia and Rhee, 1983). Rv3113 (phosphatase) has a new CRP binding site regulating the gene Rv3114 (nucleoside deaminase) involved in salvage pathways of nucleotides. Rv3505 (fadE27) and Rv3617 (ephA) also carried new CRP binding sites, possibly involved in regulating probable acyl-CoA dehydrogenase and putatative epoxide hydrolase, respectively.

3.3. CRP regulon with operon context in *M. leprae, M. avium subsp. paratuberculosis and M. smegmatis*

Given the observation that the CRP proteins from all mycobacteria have identical DNA binding domains, we extended the same profile matrix constructed for *M.tb* to predict CRP binding sites in the genomes of *M. leprae, M. avium* subsp. *paratuberculosis and M. smegmatis* (Tables 2, 3 and 4). This is the first attempt to interrogate these genomes for CRP regulon signatures. We further described the operon context of these regulons across the genomes (Supplementary Tables 1 to 4).

3.4. Conserved orthologues of CRP regulated genes across mycobacteria

A comparative analysis of CRP target genes in various mycobacteria enabled us to identify the common CRP regulated genes across mycobacteria and at least 18 genes were found to be common (Table 5). Conservation of these genes in the predicted CRP regulons suggests an important role of their cognate gene products in the life cycle of mycobacteria.

Most of the genome decay in *M. leprae* is via deletion and pseudogenization of its genes. Not all genes however were rendered superfluous by this mechanism of evolution. It has been still a quite successful pathogen in terms of virulence because it retained much of its core genome. Regulation of virulence therefore is still a key issue in *M. leprae* that makes it an interesting pathogen whenever comparative genomics of slow growing mycobacteria is addressed. Homology of most of its pseudogenized genes to 'live' genes of the present day *M.tb* makes the case even more interesting. Pseudogenized chunks still exist within the genome and homology searches therefore are capable of pointing out those CRP-regulon like regions when we attempt to do

Table 3
Predicted CRP binding sites in *M. leprae* genome

Score	Position	Binding site	Gene	Synonym	Product
3.71657	−390	ACTGTGAACCAAGTCACTAC	–	ML0201	Hypothetical protein
3.70052	−139	CGTGTGACTGATGTGACACG	–	ML0185	Hypothetical protein
3.69878	−11	CGTGTGAACGATGTCACTCC	galU	ML0182	UTP-glucose-1-phosphate uridylyltransferase
3.68332	−168	ATTGTGATTTGTATCACTGT	sucC	ML0155	Succinyl-CoA synthetase subunit beta
3.68277	−362	GTTGTGACCCATCTCACTGT	sodA	ML0072	Superoxide dismutase
3.67869	−240	TGCGTGATCTGCTTGACGAT	–	ML0298	Sulfur carrier protein ThiS
3.67586	−181	CTGGTGATAGCCCTCACGCA	–	ML0141	Hypothetical protein
3.67469	−87	GGAGTGACATCGTTCACACG	–	ML0181	Hypothetical protein
3.65712	−149	ACAGTGATACAAATCACAAT	–	ML0154	Hypothetical protein
3.62614	−351	ACCGTGACTAGGGTGACCAA	–	ML0333	Hypothetical protein
3.62494	−261	CGGGTGATAAGAGTCACCAG	–	ML0410	Putative PE-family protein
3.59242	−124	GACGTGAGGGCCATTACGCA	–	ML0229	Hypothetical protein
3.59072	−174	GGCGTGATTTCCCTTACATC	–	ML0240	Hypothetical protein
3.57863	−201	GCTGTGGCTAGTGTCACGTC	rpmF	ML0173	50S ribosomal protein L32
3.52958	−4	GTTGTGAGCAAGTTTACCGA	–	ML0243	acyl-CoA synthase
3.52512	−122	CCAGTGACCGAAGTGACCGA	–	ML0336	Putative ABC-tranporter ATP-binding protein
3.52352	−249	ATTGTCAGAGGCTTTACACG	–	ML0107	Hypothetical protein
3.50566	−338	CGGGTCAACGGGATCACCGA	–	ML0279	Hypothetical protein
3.48974	−189	GCTGTTACCCTAGTGACCCT	pabA	ML0015	Para-aminobenzoate synthase component II
3.48673	−240	GTCGTCGGTTGGGTCACGGT	purN	ML0160	Phosphoribosylglycinamide formyltransferase
3.45768	−81	AATGTCGAGCAGATCACGGA	–	ML0023	Hypothetical protein
3.45074	−329	ATTGTGCGCCGTATCACGGG	rplY	ML0245	50S ribosomal protein L25
3.45049	−248	ACGGTAAGTGGGCTGACGAA	–	ML0383	Hypothetical protein
3.45045	−12	AGGGTCAAAACCATGACCTC	pgi	ML0150	Glucose-6-phosphate isomerase
3.43921	−398	AGTGTCGGCGAAATCACATT	mas	ML0139	Putative mycocerosic synthase
3.43748	−86	GCAGTGGAATTTATCACGAT	–	ML0065	Putative monooxygenase
3.43732	−26	ATGGTCAGTGCATTAACACG	–	ML0162	Hypothetical protein
3.43312	−135	TTTGTCACACCCCTTACCGG	–	ML0314	Putative esterase
3.42825	−91	AATGTGATTTCGCCGACACT	fadD28	ML0138	acyl-CoA synthase
3.42801	−396	TACGTCATCGACGCCACGGA	rpsI	ML0365	30S ribosomal protein S9
3.42317	−94	CGTGTGGTGTTATTCACTAC	fbpC	ML0098	Antigen 85C, mycolyltransferase

Table 4
Predicted CRP binding sites in *M. smegmatis* genome

Score	Position	Binding site	Gene	Product
3.98115	−42	AATGTGAGGATCGTCACGCG	MSMEG0397	Conserved hypothetical protein
3.9723	−126	GATGTGATCGTCGTCACGTG	MSMEG0161	Oxalate/formate antiporter
3.9668	−207	ATCGTGATCTGCCTCACGTT	MSMEG4634	Conserved hypothetical protein
3.96558	−144	ATTGTGATGTGTATCACGGT	MSMEG5503	Succinyl-CoA synthetase subunit beta like protein
3.96441	−135	ACTGTGACGCGCATCACGTT	MSMEG6785	Conserved hypothetical protein
3.96317	−45	CTTGTGATGCACGTCACGAC	MSMEG3787	Hypothetical protein
3.96203	−385	GTCGTGATGAATGTCACGTC	MSMEG0953	Hypothetical protein
3.95591	−61	AGTGTGATTTACATCACACC	MSMEG2762	Conserved hypothetical protein
3.95501	−195	AACGTGACGCGCATCACGTC	MSMEG0413	Carboxyphosphonoenolpyruvate phosphonomutase-like protein
3.95431	−327	GGCGTGATGCCGGTCACGGG	MSMEG4836	Oxidoreductase
3.94687	−18	GGTGTGAGCTGTCTCACATG	MSMEG0739	Carbon monoxide dehydrogenase
3.94579	−79	GCTGTGAATCCAGTCACAGC	MSMEG3635	Hypothetical protein
3.94366	−147	ATCGTGATCTGCCTCACACT	MSMEG4561	Aldo/keto reductase
3.94267	−157	CGCGTGACGTGGCTCACGCG	MSMEG4632	Conserved hypothetical protein
3.94199	−125	GGTGTGATGTAAATCACACT	MSMEG2763	Conserved hypothetical protein
3.93621	−177	AGCGTCACCTGCGTCACGGT	MSMEG3816	Universal stress protein family domain protein
3.936	−107	GGCGTGACCCCGATCACGAG	MSMEG5035	Multidrug resistance transporter
3.9321	−233	CTCGTGATGTCGCTCACGCC	MSMEG1818	Phosphoribosylaminoimidazole carboxylase
3.92604	−221	TTCGTGATGGCGGTCACGCT	MSMEG0576	STAS domain protein
3.92588	−240	GCCGTGACATGCGTGACGTC	MSMEG2221	Substrate-CoA ligase
3.92112	−236	GGTGTCACCGAGGTCACGGG	MSMEG3365	Conserved hypothetical protein
3.92019	−42	AGTGTGAACTGTGTCACCTC	MSMEG0349	Ribonucleoside-diphosphate reductase
3.91823	−167	GCTGTGACTGGATTCACAGC	MSMEG3636	Transcription regulator
3.91242	−136	ACCGTGATACACATCACAAT	MSMEG5504	Lipoprotein NlpD
3.91112	−68	GACGTGAGCACCCTCACACG	MSMEG0950	Conserved hypothetical protein
3.91111	−183	GGCGTGAGGGAGCTCACGAA	MSMEG4148	Transcriptional regulator tetR family
3.91015	−54	GCCGTGATGGCAGTCACAAC	MSMEG6453	Hypothetical protein
3.90522	−207	GACGTGACACCCGTGACGGT	MSMEG2272	Conserved hypothetical protein
3.9047	−64	AGAGTGACCTCGGTCACGCT	MSMEG3737	trp-G type glutamine amidotransferase/dipeptidase
3.89664	−158	GGCGTGATCGTCGTGACGCT	MSMEG6098	Hypothetical protein
3.89398	−109	GATGTGACACCTGTGACAGT	MSMEG5447	Conserved hypothetical protein
3.89266	−137	GTGGTGATCTAGATCACGCT	MSMEG0195	Hypothetical protein
3.89196	2	ACTGTGAAGCGAATGACGGT	MSMEG1555	Hypothetical protein
3.89164	−41	AACGTCACGCCCATCACGCC	MSMEG2055	nuoJ
3.88848	−65	TCTGTCACATATCTCACGTT	MSMEG6232	Bacterial NAD-glutamate dehydrogenase superfamily
3.88613	−216	CGCGTGACGATCCTCACATT	MSMEG0396	fadE5
3.88278	−123	CCGGTGACCACGGTCACGCC	MSMEG1565	Glucosamine-fructose-6-phosphate aminotransferase
3.88066	−230	GTCGTAACCTGCGTCACGCG	MSMEG5242	Channel protein
3.87833	−56	AACGTGACCCAGGTCACTTA	MSMEG0158	Oxalate/formate antiporter
3.8772	−109	ACCGTAATCTGCGTCACGTG	MSMEG5382	Dehydrogenase DhgA
3.87575	−371	TGCGTGAAGGCGTTCACACC	MSMEG1123	Probable metalloprotease zinc transmembrane protein
3.87495	−60	CCTGTGAGCCGGGTCACCAC	MSMEG2582	Acetyltransferase
3.87448	−106	GCTGTGACCGCCGTCACCAG	MSMEG1048	Conserved hypothetical protein
3.87374	−47	AATGTGAGCTGCGTAACACC	MSMEG0894	Aldehyde dehydrogenase family protein
3.87358	−63	TCAGTGACCTGGGTCACGTG	MSMEG4216	Uncharacterized BCR
3.87319	−78	GACGTCACCGGGCTCACGAT	MSMEG6192	Acyltransferase
3.87309	−108	GCTGTGATGGAACTGACGGG	MSMEG0722	Hypothetical protein
3.87203	−142	GGCGTCATGCAGCTCACGAT	MSMEG2104	N5IS1096
3.86893	−220	AGCGTGATCTAGATCACCAC	MSMEG0194	Transcriptional regulator
3.86792	−25	TATGTGATCTACGTCACTGG	MSMEG0142	Probable transcription regulator protein
3.86576	−230	TCGGTGAAGCCTGTCACGCG	MSMEG2530	Hypothetical protein agmatinase
3.86496	−16	GGCGTGAAGTTCATGACGAA	MSMEG1056	Conserved hypothetical protein
3.86432	−11	TCGGTGAGGCCCGTCACGTT	MSMEG6429	Domain of unknown function (DUF427) superfamily
3.86424	6	AGTGTGACCGCCATGACACA	MSMEG0299	Aldehyde dehydrogenase
3.86323	−271	GTCGTCATAGCCTTCACGTT	MSMEG2143	Fic protein family
3.86273	−76	GGCGTGACCGTGGTCACCGG	MSMEG1564	Conserved hypothetical protein
3.86087	−15	ACGGTGATCGTGCTCACGTT	MSMEG2733	Conserved hypothetical protein
3.86075	−62	AACGTGACTTGCCTCACTTC	MSMEG5285	dctA
3.86056	−151	GACGTCACGGCGATCACGCA	MSMEG2946	glycosyl transferase 2-hydroxy-3-oxopropionate reductase
3.86051	−237	TTCGTCACGTCGATCACGGC	MSMEG5848	Conserved hypothetical protein
3.86001	−74	AACGTGAAGGCTATGACGAC	MSMEG2144	gp36
3.86	−232	GGCGTCAGGGTGCTCACGCG	MSMEG5860	Domain of unknown function DUF140 superfamily pyruvate dehydrogenase complex
3.85957	−316	ACCGTGACGGTTCTGACGAT	MSMEG4183	Hypothetical protein

Table 5
Distribution of conserved orthologues of CRP regulated genes across mycobacterial genomes

Gene	Product	Mtb	Mavi	Mlep	Msme
–	Hypothetical protein (possible secreted alanine rich protein)	Rv2700	MAP2817		MSMEG2762
–	Hypothetical protein	Rv2699c	MAP2816c		
echA6	enoyl-CoA hydratase	Rv0905	MAP0840		
–	Hypothetical protein (Beta lactamase type Zn dependent hydrolase)	*Rv0906	*MAP0841		
–	Hypothetical protein (AmpC, Beta-lactamase class C)	*Rv0907	*MAP0842		
ctpE	CtpE	*Rv0908	*MAP0843		
accD3	AccD3	Rv0904c	MAP0839c		
–	Hypothetical protein (signaling protein with CBS and cAMP binding domain)	Rv2406c	MAP2219c		
sucC	Succinyl-CoA synthetase subunit beta	Rv0951	MAP0896	ML0155	MSMEG5503
sucD	Succinyl-CoA synthetase alpha subunit	*Rv0952	*MAP0897	*ML0156	
glnD	PII uridylyl-transferase	Rv2918c	MAP2986c		
fadE9	FadE9	Rv0752c	MAP4214c		
mmsB	MmsB	*Rv0751c	*MAP4213c		
fbpC1	FbpC1 (Ag85C)	Rv3803c	MAP0217	ML0098	
–	Membrane linked adenylate cyclase	Rv3645		ML0201	
galU	Putative UTP-glucose-1-phosphate uridylyltransferase	Rv0993		ML0182	
–	Hypothetical protein (5-formyltetrahydrofolate cyclo-ligase)	Rv0992c		ML0181	
–	Hypothetical protein (similar to peptidase)	Rv0950c	MAP0895c	ML0154	*MSMEG5504

*Upstream of the corresponding gene do not have CRP box, gene lies with in CRP regulated operon.
Blank space means orthologues are not present.

comparative genomic exercises. This gives a lot of insight to the organization and arrangement of these genes and about the rate of substitutions and clock rate at which the genes such as CRP like regulons are pseudogenized.

3.4.1. Related to cAMP signaling in mycobacterium

CRP binding sites from *Rv3645* (membrane linked adenylate cyclase) and *Rv2406c* (cAMP binding protein) were found to be conserved in pathogenic mycobacteria (Table 5). The membrane-linked adenylate cyclase catalyzes the production of cAMP and, as discussed earlier may be involved in signaling in mycobacteria. *Rv2406c* encodes a hypothetical protein, sequence analyses of which reveals the presence of a CBS domain and a cAMP-binding domain. This protein might also be a part of cAMP regulated signal network in mycobacteria. *Rv2406c* also has CRP-binding element, which is conserved in pathogenic mycobacteria (Table 5).

3.4.2. Antibiotic resistance operon

Rv0905 (echA6), *Rv0906*, *Rv0907* and *Rv0908* (ctpE) constitute an operon which, in our comparative analyses, appeared to be conserved across mycobacterial genomes in terms of the CRP binding site (Table 5). *Rv0906* and *Rv0907* code for two hypothetical proteins and in our sequence analyses and conserved domain search these were found to belong to beta lactamase family of proteins. Rv0906 is a putative beta lactamase and belongs to a class of Zn dependent hydrolase. Rv0907 belongs to Amp C class of beta lactamase. The beta-lactam class of antibiotics has not been used in the treatment of *M.tb* or other mycobacterial infections as mycobacteria are resistant to these antibiotics and produce beta-lactamases (Kasik, 1979; Kwon et al., 1995). However, the effectiveness of beta-lactam/beta-lactamase inhibitor combinations has been shown in-vitro for *M.tb* (Chambers et al., 1995), and also in clinical settings of TB (Chambers et al., 1998; Cynamon et al., 1983; Segura et al., 1998; Sorg and Cynamon, 1987), *M. avium* (Casal et al., 1987), and *M. smegmatis* (Yabu et al., 1985). *Rv0908* (ctpE), which codes for putative cation transporter, is also present in the same operon. Metallo beta lactamase family enzymes are metal dependent and require Zn for their activity (Schilling et al., 2005). Both a putative Zn dependent beta lactamase and probable Zn transporter are part of this operon under same control and seem to have related function.

3.4.3. Cell wall components

Another conserved CRP binding site observed in our analyses was present in *Rv3803* (Ag85c), *MAP0217* and *ML0098* orthologues from *M.tb*, *M. avium* subsp. *paratuberculosis* and *M. leprae,* respectively. The Ag85 complex, family of three proteins of 30 to 32kDa (Ag85A, Ag85B, and Ag85C), forms a major fraction of the secreted proteins in *M.tb* culture filtrate. All of this complex possesses a mycolyl transferase enzyme activity essential for the final stages of mycobacterial cell wall assembly (Belisle et al., 1997) and the same is also involved in the host immune surveillance and defense system (Takayama et al., 2005). These three proteins are coded by three paralogous genes located in distinct regions of the bacterial genome (Content et al., 1991). These genes do not show any resemblance in their 5′ upstream region, and are probably regulated independently at the transcriptional level (Content et al., 1991). Interestingly, we found a CRP box upstream of only *Rv3803* (Ag85c) and this DNA element was conserved across all pathogenic mycobacteria and interestingly, was absent in *M. smegmatis* (Table 5). Deletion of the gene encoding antigen 85C protein alters the bacterial cell wall and its permeability, but does not kill the cells (Jackson et al., 1999).

CRP binding site in *Rv0993* (gal U), which encodes UTP: alpha-D-glucose-1-phosphate uridylyltransferase, is also conserved

in its *M. leprae* orthologue. As discussed in earlier section the cognate gene product is important in cell envelope biogenesis of mycobacteria (Weston et al., 1997).

Rv0904c (AacD3) also has a CRP binding site, which is conserved in pathogenic mycobacteria *M.tb,* and *M. avium* subsp. *paratuberculosis. Rv0905* encodes putative Enoyl-CoA Hydratase which catalyses the elongation of unsaturated fatty acid chain. *Rv0904c* (AccD3), which codes for putative beta sub unit of acetyl-coenzyme A carboxylase carboxyl transferase is an enzyme of the mycolic acid biosynthetic pathway (Gande et al., 2004) and is critical for cell wall formation. *Rv0904* shares CRP binding sites with *Rv0905*, an example of divergent regulation. It is therefore, interesting to document divergent regulation of operons amidst a convergently evolving genome.

3.4.4. Metabolic enzymes

The *sucCD* operon comprising of *Rv0951* (sucC) *and Rv0952* (sucD) which encode succinyl coA synthetase beta and alpha respectively, has a CRP binding site that is conserved among mycobacteria (Table 5). Succinyl-CoA synthetase is responsible for carrying out two unrelated but vital metabolic functions. One, it catalyzes the substrate-level phosphorylation step of the citric acid cycle (Kaufman et al., 1953), and two, it replenishes succinyl-CoA for ketone body catabolism(Ottaway et al., 1981) and for porphyrin synthesis (Labbe et al., 1965). We observed the presence in the same locus of an unrelated ORF on opposite strand *Rv0950c* that shares a common CRP binding site with *sucCD* operon and encoding a hypothetical protein. It could be a case of divergent transcription regulation. *Rv0752c* (fadE9) and *Rv0753c* (MmsB) constitute an operon and have a CRP binding box, which is conserved in both *M.tb* and *M. avium* subsp. *paratuberculosis* orthologues. *Rv0752c* encodes a putative Acyl CoA dehydrogenase, which is a flavoprotein catalyzing desaturation of acyl-CoA esters and plays an important role in the oxidation of fatty acyl-CoA esters. The ORF *Rv0753c* encodes a putative 3-Hydroxy isobutyrate dehydrogenase catalyzed oxidation of 3-Hydroxyisobutyrate to methylmalonate semialdehyde.

We broadened our search of the CRP-binding sites in intergenic regions of annotated as well as non-annotated open reading frames. This was then extended to other mycobacterial genomes to pin-point the common CRP regulated genes across the genus.

4. Conclusion

Rv3676 (*M.tb*-CRP) was earlier reported to be essential for the survival of mycobacteria inside macrophages and in animal models (Rickman et al., 2005). Further, in our analyses, we found high conservation of these CRP regulated genes among pathogenic mycobacteria than in non-pathogenic mycobacteria. This strengthens the notion that *M.tb*-CRP and its regulated genes are important in pathogenesis of mycobacteria and that these might have co-evolved with the pathogenic branch as a result of genome optimization aimed at devising new survival strategies. That many of these predicted target proteins are critical for the survival of the mycobacterium in the hostile environment of the macrophage adds a new dimension to our understanding of the regulatory complexity in *M.tb*. This regulatory paradigm however, awaits experimental validation *in vivo* of the computer-based predictions of complex networks of *M.tb* genes and cAMP as a candidate effector could play a critical role.

Acknowledgement

Research in SEH laboratory was supported by grants from the Department of Biotechnology and Council of Scientific & Industrial Research (CSIR), Govt. of India. YA is a recipient of Senior Research Fellowship from the CSIR. Thanks are also due to Ali Imran Jehangiri and Manjulatha Devi for their technical support.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.gene.2007.10.017.

References

Akhter, Y., Tundup, S., Hasnain, S.E., 2007. Novel biochemical properties of a CRP/FNR family transcription factor from *Mycobacterium tuberculosis*. Int. J. Med. Microbiol. 297, 451–457.

Akif, M., Akhter, Y., Hasnain, S.E., Mande, S.C., 2006. Crystallization and preliminary X-ray crystallographic studies of *Mycobacterium tuberculosis* CRP/FNR family transcriptional regulator. Acta Crystallogr. F 62, 873–875.

Bai, G., McCue, L.A., McDonough, K.A., 2005. Characterization of *Mycobacterium tuberculosis* Rv3676 (CRPMt), a cyclic AMP receptor protein-like DNA binding protein. J. Bacteriol. 187, 7795–8004.

Barnard, A.M., Green, J., Busby, S.J., 2003. Transcription regulation by tandem-bound FNR at *Escherichia coli* promoters. J. Bacteriol. 185, 5993–6004.

Belisle, J.T., Vissa, V.D., Sievert, T., Takayama, K., Brennan, P.J., Besra, G.S., 1997. Role of the major antigen of *Mycobacterium tuberculosis* in cell wall biogenesis. Science 276, 1420–1422.

Berg, O.G., von Hippel, P.H., 1988. Selection of DNA binding sites by regulatory proteins. II. The binding specificity of cyclic AMP receptor protein to recognition sites. J Mol Biol. 200, 709–723.

Casal, M.J., Rodriguez, F.C., Luna, M.D., Benavente, M.C., 1987. *In vitro* susceptibility of *Mycobacterium tuberculosis, Mycobacterium africanum*, *Mycobacterium bovis*, *Mycobacterium avium*, *Mycobacterium fortuitum*, and *Mycobacterium chelonae* to ticarcillin in combination with clavulanic acid. Antimicrob. Agents Chemother. 31, 132–133.

Chambers, H.F., et al., 1995. Can penicillins and other beta-lactam antibiotics be used to treat tuberculosis? Antimicrob. Agents Chemother. 39, 2620–2624.

Chambers, H.F., Kocagoz, T., Sipit, T., Turner, J., Hopewell, P.C., 1998. Activity of amoxicillin/clavulanate in patients with tuberculosis. Clin. Infect. Dis. 26, 874–877.

Cole, S.T., et al., 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 393, 537–544.

Content, J., de la Cuvellerie, A., De Wit, L., Vincent-Levy-Frebault, V., Ooms, J., De Bruyn, J., 1991. The genes coding for the antigen 85 complexes of *Mycobacterium tuberculosis* and *Mycobacterium bovis* BCG are members of a gene family: cloning, sequence determination, and genomic organization of the gene coding for antigen Ag85C of *M. tuberculosis*. Infect. Immun. 59, 3205–3212.

Crooks, G.E., Hon, G., Chandonia, J.M., Brenner, S.E., 2004. WebLogo: a sequence logo generator. Genome Res. 14, 1188–1190.

Cynamon, M.H., Palmer, G.S., 1983. *In vitro* activity of amoxicillin in combination with clavulanic acid against *Mycobacterium tuberculosis*. Antimicrob. Agents Chemother. 24, 429–431.

Gande, R., et al., 2004. Acyl-CoA carboxylases (accD2 and accD3), together with a unique polyketide synthase (Cg-pks), are key to mycolic acid biosynthesis in Corynebacterianeae such as *Corynebacterium glutamicum* and *Mycobacterium tuberculosis*. J. Biol. Chem. 279, 44847–44857.

Garcia, E., Rhee, S.G., 1983. Cascade control of *Escherichia coli* glutamine synthetase. Purification and properties of PII uridylyltransferase and uridylyl-removing enzyme. J. Biol. Chem. 258, 2246–2253.

Green, J., Scott, C., Guest, J., 2001. Functional versatility in the CRP-FNR superfamily of transcription factors: FNR and FLP. Adv. Microb. Physiol. 44, 1–34.

Jackson, M., et al., 1999. Inactivation of the antigen 85C gene profoundly affects the mycolate content and alters the permeability of the *Mycobacterium tuberculosis* cell envelope. Mol. Microbiol. 31, 1573–1587.

Kasik, J.E., 1979. Mycobacterial beta-lactamases. In: Hamilton-Miller, J.M.T., Smith, J.T. (Eds.), beta-Lactamases. Academic Press, New York, pp. 339–350.

Kaufman, S., Gilvarg, C., Cori, O., Ochoa, S., 1953. Enzymatic oxidation of alpha-ketoglutarate and coupled phosphorylation. J. Biol. Chem. 203, 869–888.

Korner, H., Sofia, H.J., 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.

Kremer, L., Baulard, A.R., Besra, G.S., 2000. Molecular Genetics of Mycobacteria; Eleventh chapter: Genetics of Mycolic Acid Biosynthesis. ASM Press, pp. 173–190. 1-55581-191-4.

Kwon, H.H., Tomioka, H., Saito, H., 1995. Distribution and characterization of beta-lactamase of mycobacteria and related organisms. Tuber Lung Dis 76, 141–148.

Labbe, R.F., Kurumada, T., Onisawa, J., 1965. The role of succinyl-CoA synthetase in the control of heme biosynthesis. Biochim. Biophys. Acta. 111, 403–415.

Linder, J.U., Schultz, J.E., 2003. The class III adenylyl cyclases:multi-purpose signalling modules. Cell. Signal. 15, 1081–1089.

Marchler-Bauer, A., Panchenko, A.R., Shoemaker, B.A., Thiessen, P.A., Geer, L.Y., Bryant, S.H., 2002. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. Nucleic Acids Res. 30, 281–283.

Mattow, J., et al., 2001. Identification of proteins from *Mycobacterium tuberculosis* missing in attenuated *Mycobacterium bovis* BCG strains. Electrophoresis 22, 2936–2946.

Ottaway, J.H., McClellan, J.A., Saunderson, C.L., 1981. Succinic thiokinase and metabolic control. Int. J. Biochem. 13, 401–410.

Prakash, P., Yellaboina, S., Ranjan, A., Hasnain, S.E., 2005. Computational prediction and experimental verification of novel IdeR binding sites in the upstream sequences of *Mycobacterium tuberculosis* ORFs. Bioinformatics 21, 2161–2166.

Rickman, L., et al., 2005. A member of the cAMP receptor protein family of transcription regulators in *Mycobacterium tuberculosis* is required for virulence in mice and controls transcription of the rpfA gene coding for a resuscitation promoting factor. Mol. Microbiol. 56, 1274–1286.

Schilling, O., et al., 2005. Zinc-and iron-dependent cytosolic metallo-beta-lactamase domain proteins exhibit similar zinc-binding affinities, independent of an atypical glutamate at the metal-binding site. Biochem. J. 385, 145–153.

Schultz, S.C., Shields, G.C., Steitz, T.A., 1991. Crystal structure of a CAP–DNA complex: the DNA is bent by 90 degrees. Science 2539, 1001–1007.

Segura, C., Salvado, M., Collado, I., Chaves, J., Coira, A., 1998. Contribution of beta-lactamases to beta-lactam susceptibilities of susceptible and multidrug-resistant *Mycobacterium tuberculosis* clinical isolates. Antimicrob. Agents Chemother. 42, 1524–1526.

Sorg, T.B., Cynamon, M.H., 1987. Comparison of four beta-lactamase inhibitors in combination with ampicillin against *Mycobacterium tuberculosis*. J. Antimicrob. Chemother. 19, 59–64.

Spiro, S., 1994. The FNR family of transcription regulators. Antonie van Leiwenhoek 66 , 23–26.

Spreadbury, C.L., et al., 2005. Point mutations in the DNA and cNMP-binding domains of the homologue of the cAMP receptor protein (CRP) in *Mycobacterium bovis* BCG: implications for the inactivation of a global regulator and strain attenuation. Microbiology 151, 547–556.

Takayama, K., Wang, C., Besra, G.S., 2005. Pathway to synthesis and processing of mycolic acids in *Mycobacterium tuberculosis*. Clin. Microbiol. Rev. 18, 81–101.

Tundup, S., Akhter, Y., Thiagarajan, D., Hasnain, S.E., 2006. Clusters of PE and PPE genes of *Mycobacterium tuberculosis* are organized in operons: evidence that PE Rv2431c is co-transcribed with PPE Rv2430c and their gene products interact with each other. FEBS Lett. 580, 1285–1293.

Weissborn, A.C., Liu, Q., Rumley, M.K., Kennedy, E.P., 1994. UTP: alpha-D-glucose-1-phosphate uridylyltransferase of *Escherichia coli*: isolation and DNA sequence of the galU gene and purification of the enzyme. J. Bacteriol. 176, 2611–2618.

Weston, A., Stern, R.J., Lee, R.E., Nassau, P.M., Monsey, D., Martin, S.L., Scherman, M.S., Besra, G.S., Duncan, K., McNeil, M.R., 1997. Biosynthetic origin of mycobacterial cell wall galactofuranosyl residues. Tubercle and Lung Disease 78, 123–131.

Yabu, K., Kaneda, S., Ochiai, T., 1985. Relationship between beta-lactamase activity and resistance to beta-lactam antibiotics in *Mycobacterium smegmatis*. Microbiol. Immunol. 29, 803–809.

Yellaboina, S., Ranjan, S., Chakhaiyar, P., Hasnain, S.E., 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.

Yellaboina, S., Ranjan, S., Vindal, V., Ranjan, A., 2006. Comparative analysis of iron regulated genes in mycobacteria. FEBS Lett. 580, 2567–2576.

Available online at www.sciencedirect.com
ScienceDirect
International Journal of Medical Microbiology 297 (2007) 451–457
IJMM
www.elsevier.de/ijmm

Novel biochemical properties of a CRP/FNR family transcription factor from *Mycobacterium tuberculosis*

Yusuf Akhter[a], Smanla Tundup[a], Seyed E. Hasnain[a,b,c,d,*]

[a]*Laboratory of Molecular and Cellular Biology, CDFD, Hyderabad 500076, India*
[b]*Jawaharlal Nehru Centre for Advanced Scientific Research, Jakkur, Bangalore 560064, India*
[c]*University of Hyderabad, Hyderabad 500046, India*
[d]*Institute of Life Sciences, Hyderabad 500046, India*

Received 19 October 2006; received in revised form 17 April 2007; accepted 24 April 2007

Abstract

Cyclic AMP (cAMP) receptor protein (CRP)/fumarate nitrate reductase regulator (FNR) family proteins are actively associated with defense against low oxygen stress, starvation and extreme temperature conditions. They are DNA-binding proteins and regulate target genes carrying the regulatory CRP/FNR cognate nucleotide sequence elements. Recombinant protein encoded by the *Mycobacterium tuberculosis* ORF *Rv3676*, a putative CRP/FNR regulator, was purified from *Escherichia coli* and was found to exist as dimer, devoid of any metal cation cofactor. Purified rRv3676 exhibited cAMP binding in a concentration-dependent manner. At lower concentrations of cAMP (6–10 μM) rRv3676 shows positive cooperativity; at 10 μM cAMP the protein exists in the most open conformation. rRv3676 could bind specifically to the putative CRP/FNR nucleotide sequence elements as evident from electrophoretic mobility shift assay.

Keywords: Hypoxia; Starvation; Cyclic AMP; *Rv3676*; Transcription factor; Gene regulation

Introduction

Mycobacterium tuberculosis (*M.tb*), the causative agent of tuberculosis, is responsible for ~3 million deaths every year (World Health Organization, 2006). *M.tb* can exist in latent phase (Ulrichs and Kaufmann, 2006), where it can survive in a very hostile environment, which includes low nutrition and hypoxia, for long periods. CRP/FNR (cAMP receptor protein/fumarate and nitrate reductase regulator) is one of the members of a family of transcriptional regulators. With over 370 family members, these DNA-binding proteins predominantly function as positive transcriptional regulators and are known to be associated with defense against oxygen stress and starvation, and at the same time respond to other environmental signals. The distinctive features of CRP/FNR superfamily proteins include the presence of a nucleotide-binding domain and a helix-turn-helix DNA-binding motif at the N- and C-terminal, respectively. The prototype cAMP-binding domain (Schultz et al., 1991) is a versatile structure that has evolved to accommodate different functional specificities in response to a broad range of signals (Green et al., 2001; Korner et al., 2003).

M.tb H37Rv ORF *Rv3676* codes for a putative CRP/FNR protein (Cole et al., 1998), which is required for

*Corresponding author. University of Hyderabad, Gachibowli, Hyderabad 500046, India. Tel.: +91 40 2301 0121; fax: +91 40 2301 1090.
E-mail address: vc@uohyd.ernet.in (S.E. Hasnain).

1438-4221/$ - see front matter © 2007 Elsevier GmbH. All rights reserved.
doi:10.1016/j.ijmm.2007.04.009

virulence in mice and controls transcription of specific genes (Rickman et al., 2005). Recently, Bai et al. (2005) reported the characterization of Rv3676 protein using computational and experimental methods (Bai et al., 2005). We recently described the crystallization and preliminary X-ray diffraction data for this CRP/FNR regulator (Akif et al., 2006). Phylogenetically, *M.tb Rv3676* is nearest to the CooA branch represented by the CO sensor protein of *Rhodospirillum rubrum* (Korner et al., 2003). However, the relative positions of the regulatory and the DNA-binding domain are strikingly different in that the recognition helix of CooA is rotated 180° away from the position occupied in CRP-cAMP. Further differences between CooA and CRP include an extended N-terminus providing a ligand to the heme of the opposite subunit, an 11-amino-acid extension in the regulatory domain (positions 72–82) to accommodate the heme and a different composition of the hinge region toward the C-terminus, which causes the displacement of the DNA-binding domain (Lanzilotta et al., 2000). Sequence alignment suggests that these 11 residues are not fully conserved in *Rv3676*.

Escherichia coli CRP and FNR regulate transcription globally in response to glucose starvation and anaerobic conditions, respectively (Kolb et al., 1993). *E. coli* FNR is structurally related to CRP except for the presence of four conserved cysteine residues at the N-terminal extension, which form part of an iron–sulfur cluster and a redox-sensing domain of FNR. This iron–sulfur cluster is absent in *M.tb* Rv3676 similar to other members of the same family from other systems such as e.g. *Pseudomonas stutzeri* (Vollack et al., 1999) and *Bradyrhizobium japonicum* (Mesa et al., 2003). Although these proteins do not have an iron–sulfur cluster, they are the regulators of oxygen tension sensu stricto.

The earlier report by Bai et al. (2005) focused on CRP regulon prediction and the experimental validation of the same and provided the first direct evidence for cAMP binding to a transcription factor in *M.tb*, thereby suggesting a role for cAMP-mediated signal transduction in this bacterium. We now describe the purification and comprehensive characterization of a CRP/FNR regulator from *M.tb* in terms of oligomeric state, cAMP and DNA binding. Our results point to some new unusual properties of Rv3676 protein, which could have physiological relevance.

Materials and methods

Bacterial strains and plasmids

E. coli DH5α and *E. coli* BL21DE3 bacterial strains were used for cloning and expression purposes, respectively. DNA manipulations were carried out in pET23a plasmid vector using standard techniques. Integrity of the plasmid constructs was confirmed by DNA sequencing.

Cloning, expression and purification of recombinant *M.tb* Rv3676

M.tb ORF *Rv3676* was PCR amplified from *M.tb* H37Rv genomic DNA using forward (GGATAT**CATATG**GTGGACGAGATCCTGGCCAGGG) and reverse (CG**CTCGAG**CCTCGCTCGGCGGGCCAGTC) primers with restriction enzyme sites for cloning (shown in bold). The amplicon was cloned into the corresponding sites of pET23a, and recombinant Rv3676 protein was purified as a 6 × His-tagged fusion protein from *E. coli* BL21 (DE3) cells as described earlier (Akif et al., 2006). Protein concentration was estimated using the dye-binding method (Bradford, 1976). To determine suitable storage conditions, aliquots of recombinant Rv3676 (rRv3676) were dialyzed in different buffers, namely phosphate-buffered saline (PBS), 10 mM Tris and 10 mM HEPES. Storage temperature was also optimized, and the conditions under which rRv3676 was most stable were selected.

Analytical size exclusion chromatography

The oligomeric state of native recombinant protein was determined by analytical size exclusion chromatography using a Superose 6 fast protein liquid chromatographic column (BIORAD) at room temperature with PBS as running buffer. A standard curve was prepared according to the instruction manual using standard molecular weight markers. The void volume was determined using Blue Dextran 2000. The elution parameter K_{av} was calculated as follows: $K_{av} = (V_e - V_0)/V_s$, where V_e is the elution volume for the protein, V_0 the column void volume, and V_s the total stationary phase volume. K_{av} was plotted against log molecular weight.

Spectral analyses

To detect the presence, if any, of any associated cofactor, absorption was measured between 200 and 800 nm using a Perkin-Elmer spectrophotometer. Fluorescence spectrometric measurements and ligand-binding assays were carried out using a Perkin-Elmer LS50B luminescence spectrometer and a sample volume of 200 μl with 0.3 cm path length. Tryptophan fluorescence was measured at an excitation wavelength of 295 nm. The slit widths for excitation and emission were 10 and 20 nm, respectively. Emission spectra were recorded between 310 and 500 nm. All spectra measurements were corrected by subtracting the corresponding buffer

backgrounds. Increasing concentrations of urea (1–8 M) and a constant concentration (3 μM) of recombinant protein was used to study the denaturation kinetics of the protein. At 8 M urea the protein was fully unfolded, and the spectrum of fully unfolded protein was further compared with that of 6 μM free tryptophan. The circular dichroism (CD) spectra of recombinant native protein and liganded protein, incubated with different concentrations of cAMP (6–16 μM), were recorded using a JASCO CD spectrometer (Model J-810) between 200 and 250 nm in steps of 0.5 nm with four accumulations in each step. The spectral baseline was corrected by subtracting the respective blanks. Molar ellipticity, expressed in millidegrees, was plotted as a function of wavelength. The secondary structure content of the protein was calculated by using k2d software (www.embl-heidelberg.de/~andrade/k2d/). For CD and fluorimetric spectral analysis, 5 and 3 μM recombinant protein was used, respectively.

Electrophoretic mobility shift assay (EMSA)

Gel retardation assays were carried out as described earlier (Prakash et al., 2005). Complementary synthetic oligodeoxyribonucleotides corresponding to the CRP/FNR-binding site (AATGTGATCTAGGTCACGTG) present upstream of *Rv1552* (*frdA*) were end labeled with [γ-^{32}P]ATP using T4 polynucleotide kinase. One nanogram labeled oligonucleotide was incubated with 3 μg recombinant protein in binding buffer (10 mM Tris–HCl, 50 mM NaCl, 50 mM $MgCl_2$, 1 μg BSA, 1 μg poly dI:dC, 1 mM EDTA, 1 mM DTT, 1 mM PMSF and 10% glycerol) in 20 μl reaction volume, incubated for 30 min at room temperature and fractionated on a 5% polyacrylamide gel in TBE. After electrophoresis at 200 V at 4 °C, the gel was dried and analyzed by autoradiography. To check for the specificity of the complex, unlabeled homologous oligonucleotide or an oligonucleotide carrying a specific mutation (*mut*) critical for binding (AATTTGATCTAGGTCACGTG, shown as underlined) were used in competition assays.

Results

Purified rRv3676 exists in dimeric state

M.tb rRv3676 was purified as a 6 × His-tagged protein using affinity chromatography as described in Materials and methods. Purified rRv3676 was stable at 4 °C in PBS while at lower temperatures and in other buffers it formed insoluble aggregates. The purified recombinant protein was fractionated by electrophoresis on a 10% polyacrylamide gel and stained with Coomassie Brilliant Blue (Fig. 1a, inset). Gel filtration analysis was carried out to determine the oligomeric nature, if any, of the rRv3676 protein. rRv3676 exists as a pure dimer of ~53 kDa, as evident from analytical size exclusion chromatography (Fig. 1).

Purified rRv3676 has no associated cation cofactors

In most oxygen tension-sensing proteins belonging to the same family, transition metals like Fe or Ni are associated with the protein to sense the fluctuations of oxygen availability via redox mechanisms (Korner et al., 2003). We therefore scanned the absorption spectrum of purified rRv3676 to check for the presence, if any, of a metal ion cofactor. Spectral analysis revealed two peaks,

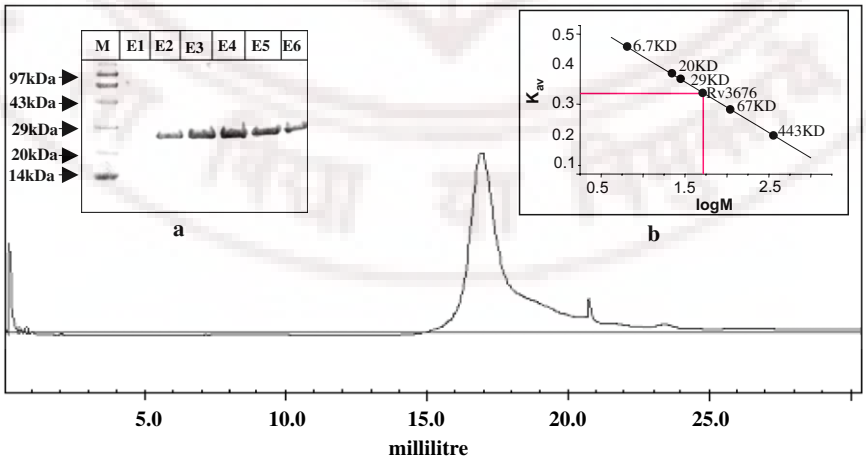


Fig. 1. Recombinant Rv3676 protein exists as a dimer: (a) Coomassie-stained polyacrylamide gel showing the ORF *Rv3676*-encoded protein of *Mycobacterium tuberculosis* purified from *E. coli*. M represents the protein molecular size marker (Medium range, Genei, India.), E1–E6 show successive TALON column eluted fractions of the recombinant protein. (b) The purified protein was pooled and fractionated on a Superose 6 FPLC column, resulting in a single peak. The calculated molecular mass of the recombinant protein was ~53 kDa corresponding to a dimeric state.

one at 295 nm and the other at 280 nm (Fig. 2). The first peak corresponds to tryptophan, while the other peak is due to phenylalanine and tyrosine. The fact that we could not see any other peak clearly indicates that rRv3676 does not have any other associated metal ion cofactor. This rather unexpected finding suggests that *M.tb* Rv3676 apparently uses some other mechanism(s) to sense effector signals.

cAMP binds to purified rRv3676 in a concentration-dependent manner

Results of in-silico (data not shown) analyses revealed the presence of a putative cAMP-binding domain at the N-terminal end of rRv3676 protein, thereby raising a strong probability that cAMP may be acting as an effector of Rv3676 protein. We therefore subjected purified rRv3676 to CD analysis in the presence and absence of cAMP as ligand. The change in secondary structure was calculated using k2d software (http://www.bork.embl-heidelberg.de/~andrade/k2d) based upon a method developed earlier (Yang et al., 1986). A comparison of CD spectra of cAMP-free and cAMP-bound rRv3676 provides evidence of binding (change in secondary structure of purified rRv3676 upon interaction with cAMP). This change in secondary structure clearly appears to be a function of increasing concentration of cAMP (Fig. 3). That cAMP indeed causes concentration-dependent conformational alterations within rRv3676 was further confirmed by tryptophan fluorescence spectrometry.

The two tryptophan residues (Trp112 and Trp203) present in Rv3676 protein were used as probe to study conformational changes in the protein in solution upon urea-induced denaturation. Purified rRv3676 unfolds completely in the presence of 8 M urea without any further increase in fluorescence (Fig. 4), indicating the presence of fully unfolded protein molecules. The maximum wavelength of absorbance of denatured rRv3676 is approximately 360 nm, which is equal to the maximum wavelength of absorbance of 6 μM free tryptophan (data not shown). As protein unfolds (relaxed) tryptophan residues are exposed to the solvent, resulting in an increase in relative fluorescence. We therefore used the fluorescence method to assay whether

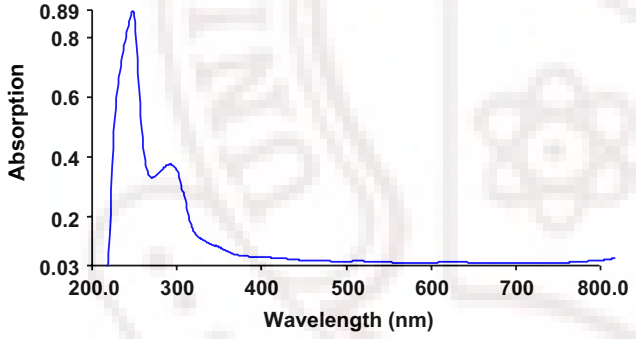


Fig. 2. Absorption spectrum of purified rRv3676 indicating the absence of any metal ion cofactor. The spectrum shows two prominent peaks, one at 295 nm (corresponds to tryptophan) and the other at 280 nm (corresponds to tyrosine and phenylalanine).

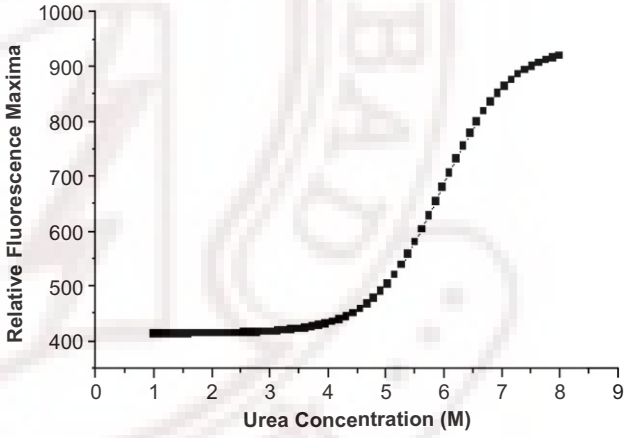


Fig. 4. Denaturation of recombinant Rv3676 in the presence of urea. Recombinant Rv3676 is completely denatured in the presence of 8 M urea as evident from maximum fluorescence.

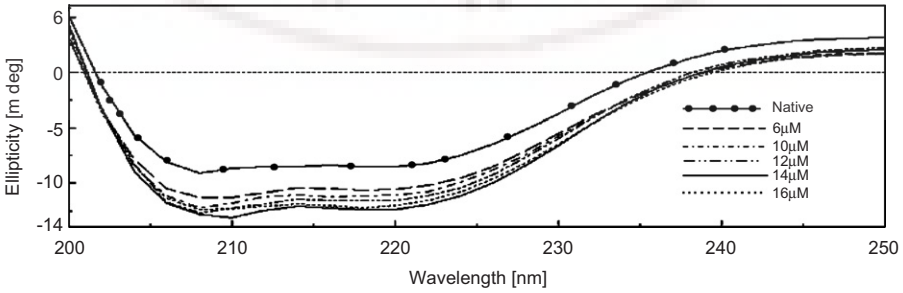


Fig. 3. Purified rRv3676 protein displays cAMP-binding activity as evident from circular dichroism (CD) spectral analysis (see Materials and methods). Binding of cAMP to rRv3676 is evident from a change in secondary structure of the native protein upon interaction with cAMP.

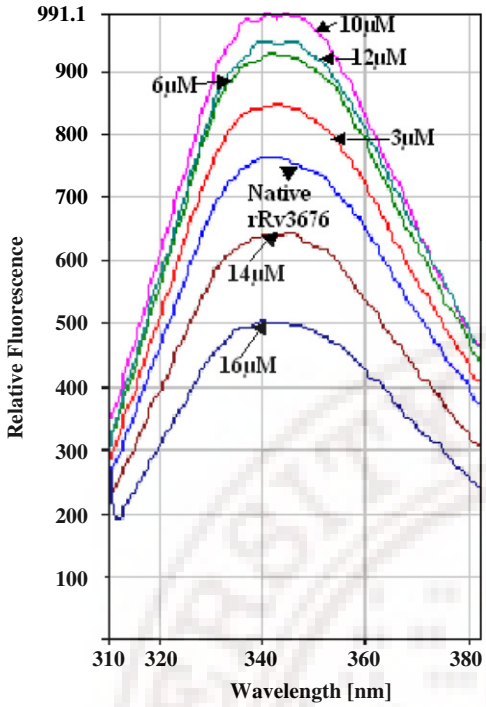


Fig. 5. Fluorescence emission spectra of rRv3676 as a function of cAMP concentration (6–16 μM). The fluorescence maximum of the protein increases steadily up to 10 μM cAMP and later drops as a function of increasing cAMP concentration (12–16 μM).

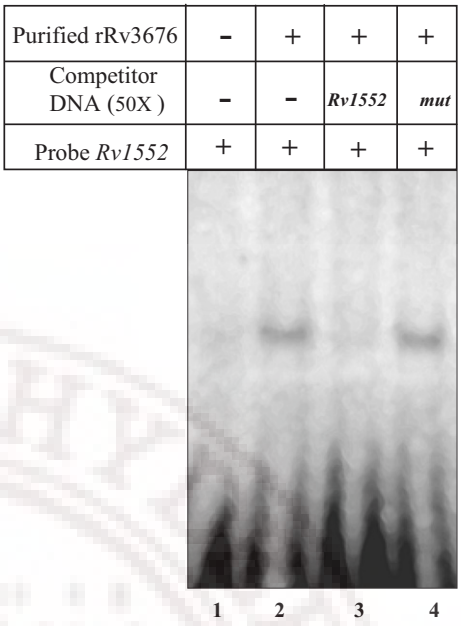


Fig. 6. Recombinant Rv3676 binds to the CRP/FNR-binding element present upstream of *Rv1552* (*frdA*). Specificity of binding was confirmed by competition with a 50-fold excess of unlabeled ligand (lane 3). The complex is unaffected when a mutant oligonucleotide (*mut*) carrying an alteration in the binding sequence is used in the competition assay (lane 4).

increasing concentrations of cAMP have any effect on the unfolding of rRv3676 protein. PBS was used as solvent, which has a physiological pH and an ionic strength similar to the intracellular milieu of the bacilli. Physiological cAMP levels are in the range of 0–<10 μM. At lower concentrations (6–10 μM) the binding shows positive cooperativity, and at 10 μM cAMP the protein is in the most open conformation. This is evident from the increase in tryptophan fluorescence (Fig. 5). With further increase of cAMP (12–16 μM), the relative tryptophan fluorescence decreases, suggesting that the protein is getting compacted. This protein compaction could be a reflection of a feedback regulation.

Purified rRv3676 binds in vitro to the CRP/FNR cognate nucleotide sequence motif present upstream of *Rv1552*

Having shown that Rv3676 is a likely member of the CRP/FNR family of DNA-binding proteins, we tested whether purified rRv3676 indeed displays such an activity. We selected *M.tb* ORF *Rv1552*, which is putatively regulated by the CRP/FNR family of transcriptional regulators. Synthetic oligodeoxyribonucleotide corresponding to the CRP/FNR cognate DNA-binding element present in ORF *Rv1552* (*frd*A) was radiolabeled and used as probe in EMSA using purified rRv3676. the results clearly show a shift in the mobility of the CRP/FNR probe upon incubation with rRv3676 protein (Fig. 6, lane 2). The specificity of the binding is evident from the disappearance of the complex in a competition assay using a 50-fold excess of homologous unlabeled CRP/FNR oligonucleotides (Fig. 6, lane 3). The specificity of this DNA–protein interaction is further emphasized by the absence of any competition when a mutated version of the oligonucleotide (*mut* in Fig. 6, lane 4) is used. These results demonstrate that rRv3676 indeed specifically interacts with the CRP/FNR cognate DNA sequence motifs.

Discussion

M.tb harbors a single member of the CRP/FNR superfamily, i.e. *Rv3676*. That this gene is important is evident from knockout studies. An Rv3676 knock-out strain is impaired in growth under in-vitro conditions, in bone marrow-derived macrophages and also in an animal model (Rickman et al., 2005). We therefore selected ORF *Rv3676* for further analyses of its protein product. Purified rRv3676 exists in a single oligomeric state as a homodimer (Fig. 1) and is active in terms of

DNA binding. It is interesting to note that, despite an only weak DNA-binding activity, the interaction is very specific as could be seen from the inability of the mutant oligonucleotide to compete for binding.

Most other proteins of this family are active as dimers (Korner et al., 2003) but, unlike Rv3676, contain metal cations such as iron and nickel as cofactors. Interestingly, Rv3676 does not carry a metal-binding motif, and the absence of any metal cofactor is indeed evident from the spectral features of rRv3676. *M.tb* Rv3676 thus appears to be different from other oxygen-sensing proteins in terms of non-requirement of a metal cationic cofactor.

To investigate the ability of rRv3676 to bind to its cognate DNA motif, we carried out EMSA using purified rRv3676 and a radiolabeled oligonucleotide carrying the CRP/FNR-binding site present upstream of the *frd* (*Rv1552*) gene encoding the fumarate reductase enzyme. This binding site was identified as a putative binding site in recent reports (Bai et al., 2005; Spreadbury et al., 2005). In our in-silico regulon prediction studies, this motif elicited the highest score (Akhter et al., unpublished data), and we therefore selected this for EMSA. It has been reported that Rv3676 senses oxygen (Bai et al., 2005; Spreadbury et al., 2005) indirectly by controlling the expression of genes such as *frd*. Fumarate serves as an alternative electron acceptor in the absence of oxygen, and this is mediated by a membrane-linked fumarate reductase enzyme complex (Lambden and Guest, 1976). The putative CRP/FNR-binding site, present upstream of the *frd* operon, was recognized by rRv3676 protein as evident from EMSA.

The *M.tb* genome encodes as many as 15 adenylate cyclases, suggesting that cAMP may have an important role in mycobacteria. It has indeed been reported that cAMP can alter the gene expression profile of *M.tb* during anaerobic conditions (Gazdik and McDonough, 2005). The predicted cAMP-binding site in rRv3676 indeed shows binding to cAMP leading to conformational changes in the protein as evident from spectral analyses. The extent of change in secondary structure is maximal in the presence of 10 μM cAMP. The effect of cAMP binding on the DNA-binding efficiency of Rv3676 has already been reported earlier (Bai et al., 2005). cAMP, acting as effector, is known to modulate the regulation of a large number of target genes, and it is likely that Rv3676 is involved in this process. While these in vitro findings point to the importance of cAMP, it remains to be experimentally demonstrated whether cAMP is actually involved in regulating gene expression by recruiting Rv3676 protein.

While the biophysical features of purified *M.tb* Rv3676 protein described here are physiologically relevant, experimental validation in vivo will be required to dissect the complete network of *M.tb* genes regulated by Rv3676 and cAMP.

Acknowledgments

Research in S.E. Hasnain's laboratory was supported by grants from the Department of Biotechnology and Council of Scientific & Industrial Research (CSIR), Government of India. Y. Akhter and S. Tundup are recipients of Junior Research Fellowship from the CSIR. We acknowledge Dr. Nasreen, Z. Ehtesham and Ms. Krishnaveni Mohareer for critical discussions during the course of this study.

References

Akif, M., Akhter, Y., Hasnain, S.E., Mande, S.C., 2006. Crystallization and preliminary X-ray crystallographic studies of *Mycobacterium tuberculosis* CRP/FNR family transcriptional regulator. Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun. 62, 873–875.

Bai, G., McCue, L.A., McDonough, K.A., 2005. Characterization of *Mycobacterium tuberculosis* Rv3676 (CRPMt), a cyclic AMP receptor protein-like DNA binding protein. J. Bacteriol. 187, 7795–7804.

Bradford, M.M., 1976. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal. Biochem. 72, 248–254.

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

Gazdik, M.A., McDonough, K.A., 2005. Identification of cyclic AMP-regulated genes in *Mycobacterium tuberculosis* complex bacteria under low-oxygen conditions. J. Bacteriol. 187, 2681–2692.

Green, J., Scott, C., Guest, J., 2001. Functional versatility in the CRP–FNR superfamily of transcription factors: FNR and FLP. Adv. Microb. Physiol. 44, 1–34.

Kolb, A., Busby, S., Garges, S., Adhya, S., 1993. Transcriptional regulation by cAMP and its receptor protein. Annu. Rev. Biochem. 62, 749–795.

Korner, H., Sofia, H.J., 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.

Lambden, P.R., Guest, J.R., 1976. Mutants of *Escherichia coli* K12 unable to use fumarate as an anerobic electron acceptor. J. Gen. Microbiol. 97, 145–160.

Lanzilotta, W.N., Schuller, D.J., Thorsteinsson, M.V., Kerby, R.L., Roberts, G.P., Poulos, T.L., 2000. Structure of the

CO sensing transcription activator CooA. Nat. Struct. Biol. 7, 876–880.

Mesa, S., Bedmar, E.J., Chanfon, A., Hauke, H., Fischer, H., 2003. *Bradyrhizobium japonicum* NnrR, a denitrification regulator, expands the FixLJ-FixK2 regulatory cascade. J. Bacteriol. 185, 3978–3982.

Prakash, P., Yellaboina, S., Ranjan, A., Hasnain, S.E., 2005. Computational prediction and experimental verification of novel IdeR binding sites in the upstream sequences of *Mycobacterium tuberculosis* ORFs. Bioinformatics 21, 2161–2166.

Rickman, L., Scott, C., Hunt, D.M., Hutchinson, T., Menéndez, M.C., Whalan, R., Hinds, J., Colston, M.J., Green, J., Buxton, R.S., 2005. A member of the cAMP receptor protein family of transcription regulators in *Mycobacterium tuberculosis* is required for virulence in mice and controls transcription of the rpfA gene coding for a resuscitation promoting factor. Mol. Microbiol. 56, 1274–1286.

Schultz, S.C., Shields, G.C., Steitz, T.A., 1991. Crystal structure of a CAP–DNA complex: the DNA is bent by 90 degrees. Science 253, 1001–1007.

Spreadbury, C.L., Pallen, M.J., Overton, T., Behr, M.A., Mostowy, S., Spiro, S., Busby, S.J., Cole, J.A., 2005. Point mutations in the DNA- and cNMP-binding domains of the homologue of the cAMP receptor protein (CRP) in *Mycobacterium bovis* BCG: implications for the inactivation of a global regulator and strain attenuation. Microbiology 151, 547–556.

Ulrichs, T., Kaufmann, S.H., 2006. New insights into the function of granulomas in human tuberculosis. J. Pathol. 208, 261–269.

Vollack, K., Hartig, E., Korner, H., Zumft, W.G., 1999. Multiple transcription factors of the FNR family in denitrifying *Pseudomonas stutzeri*: characterization of four *fnr*-like genes, regulatory responses and cognate metabolic processes. Mol. Microbiol. 31, 1681–1694.

World Health Organization, 2006. Tuberculosis Fact Sheet No. 104 – Global and Regional Incidence. World Health Organization, Geneva.

Yang, J.T., Wu, C.S., Martinez, H.M., 1986. Calculation of protein conformation from circular dichroism. Methods Enzymol. 130, 208–269.

Acta Crystallographica Section F
Structural Biology and Crystallization Communications
ISSN 1744-3091

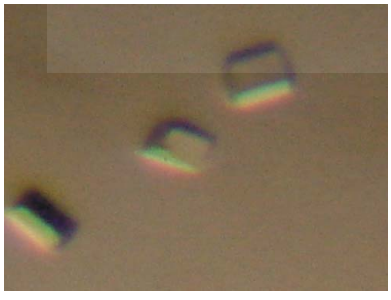
Crystallization and preliminary X-ray crystallographic studies of *Mycobacterium tuberculosis* CRP/FNR family transcription regulator

Mohd. Akif,[a] Yusuf Akhter,[a] Seyed E. Hasnain[a,b,c] and Shekhar C. Mande[a]*

[a]Centre for DNA Fingerprinting and Diagnostics, ECIL Road, Nacharam, Hyderabad 500076, India, [b]University of Hyderabad, Hyderabad 500046, India, and [c]Jawaharlal Nehru Centre for Advanced Scientific Research, Jakkur, Bangalore 560064, India

Correspondence e-mail: shekhar@cdfd.org.in

Received 26 June 2006
Accepted 19 July 2006



CRP/FNR family members are transcription factors that regulate the transcription of many genes in *Escherichia coli* and other organisms. *Mycobacterium tuberculosis* H37Rv contains a probable CRP/FNR homologue encoded by the open reading frame Rv3676. The deletion of this gene is known to cause growth defects in cell culture, in bone marrow-derived macrophages and in a mouse model of tuberculosis. The mycobacterial gene Rv3676 shares ∼32% sequence identity with prototype *E. coli* CRP. The structure of the protein might provide insight into transcriptional regulation in the pathogen by this protein. The *M. tuberculosis* CRP/FNR transcription regulator was crystallized in space group $P2_12_12_1$, with unit-cell parameters $a = 54.1$, $b = 84.6$, $c = 101.2$ Å. The crystal diffracted to a resolution of 2.9 Å. Matthews coefficient and self-rotation function calculations reveal the presence of two monomers in the asymmetric unit.

1. Introduction

Mycobacterium tuberculosis, the causative agent of tuberculosis, is one of the most dreaded human pathogens and causes ∼3 million deaths every year (World Health Organization, 2006). The characteristic feature of *M. tuberculosis* is reactivation from a latent phase, which causes the disease (Ulrichs & Kaufmann, 2006). *M. tuberculosis* possesses intricate mechanisms for survival inside the hostile environment of the host. It can persist for a long period of time in a dormant or non-replicating state, which may become active in replicating bacilli after several years when the host becomes immunocompromized (Ulrichs & Kaufmann, 2006). The outcome of *M. tuberculosis* infection solely depends upon its interactions with the environment provided by the host (Kaufmann, 2001). The reactivation may be caused by a variety of environmental signals and is mediated through several transcriptional regulators. A family of transcriptional regulators belonging to the CRP/FNR class is actively associated with low oxygen stress and starvation, including perception of various other environments (Korner *et al.*, 2003). This regulator may therefore play an important role in reactivating the dormant bacilli.

The cyclic AMP (cAMP) receptor protein (CRP) is a well known transcription regulator that regulates the transcription of many genes in *Escherichia coli* (Schultz *et al.*, 1991). It is one of the best studied transcription regulators and is also referred to as CAP (catabolite activator protein). CRP is a 45 kDa dimeric protein that has both cAMP- and DNA-binding domains (Aiba *et al.*, 1982). cAMP is required for the regulatory and DNA-binding activity of CRP (Passner *et al.*, 2000). The DNA-binding domain has a well conserved helix–turn–helix motif. The archetypical cAMP-binding domain has evolved to accommodate different functional specificities in signal detection, DNA binding and interaction with RNA polymerase to allow different family members to respond to a wide range of signals (Green *et al.*, 2001).

M. tuberculosis H37Rv contains a probable CRP/FNR homologue encoded by the open reading frame Rv3676 (Cole *et al.*, 1998). It is reported as a specific transcription factor, deletion of which is known to cause growth defects in laboratory medium, in bone marrow-

derived macrophages and in a mouse model of tuberculosis (Rickman *et al.*, 2005). Although Rv3676 shares 32% sequence identity with *E. coli* CRP, it exhibits wide divergence at the N-terminal region. The lack of conserved residues in this region might suggest different interactions of Rv3676 with RNA polymerase. Moreover, only four of the six residues that are involved in cAMP binding in *E. coli* CRP are conserved in Rv3676. It has been reported that the CRP/FNR homologue is closer to the COOA branch represented by the CO sensor protein from *Rhodosprillium rubrum* (Korner *et al.*, 2003). Considering the fact that Rv3676 shares heterogeneity in both the DNA-binding and cAMP-binding sequences compared with other prototypes, its structural properties should be interesting to address. To understand the molecular mechanism of transcription regulation of the CRP/FNR family regulator in *M. tuberculosis*, Rv3676 has been crystallized for structure determination. This paper describes the preliminary crystallographic characterizations of the *M. tuberculosis* CRP/FNR transcription regulator.

2. Material and methods

2.1. Expression and purification

E. coli BL21 (DE3) cells harbouring the expression vector pET28a with the gene Rv3676, cloned in *Nde*I and *Hin*dIII sites with a six-His tag at the C-terminus, were grown in 200 ml LB supplemented with 30 µg ml^{-1} kanamycin. The culture was induced at an OD_{600} of 0.4 with 0.4 m*M* IPTG at 300 K and 200 rev min^{-1} to allow protein expression. The cells were harvested by centrifugation and resuspended in lysis buffer (PBS; 50 m*M* phosphate buffer pH 8 and 155 m*M* NaCl) supplemented with 0.1 m*M* PMSF. After sonication, the supernatant was applied onto a Talon cobalt-affinity resin column (Clontech, USA) pre-equilibrated with lysis buffer, followed by washing with five bed volumes with lysis buffer supplemented with 10 m*M* imidazole. The recombinant protein was eluted with lysis buffer supplemented with 250 m*M* imidazole.

2.2. Crystallization

The purified protein was concentrated to 9 mg ml^{-1} and dialyzed against 10 m*M* Tris pH 8 containing 20 m*M* NaCl using a Centricon concentrator (Amicon; 3 kDa molecular-weight cutoff). The purified protein precipitated when stored at 253 K. Thus, freshly purified protein was used for crystallization trials with a variety of random conditions using a Magic 96 matrix. The hanging-drop vapour-diffusion technique was used for random screening of crystallization conditions. Crystals were obtained when 2 µl protein solution was mixed with 2 µl well solution and allowed to equilibrate against 500 µl well solution at 277 K. 0.2 *M* Li_2SO_4 and 15% ethanol in 0.1 *M* sodium citrate buffer pH 5.5 were used in the well solution.

2.3. Data collection

Crystals of CRP/FNR were soaked in artificial mother liquor (0.1 *M* sodium citrate buffer pH 5.5, 0.2 *M* Li_2SO_4 and 21% ethanol) supplemented with 20% ethylene glycol as the cryoprotectant. The diffraction data were collected from a single crystal at 100 K at the XRD1 beamline at the ELETTRA synchrotron facility, Trieste, Italy using a MAR CCD 165 detector. The crystal-to-detector distance was maintained at 200 mm with oscillations of 1°, covering up to 180° in order to obtain complete data. Determination of unit-cell parameters and integration of reflection data was performed by *DENZO*/*SCALEPACK* (Otwinowski & Minor, 1997). The intensities were then converted to structure-factor amplitudes by the *TRUNCATE* program from the *CCP*4 suite (Collaborative Computational Project, Number 4, 1994). The self-rotation function was calculated using *POLARRFN* as available in the *CCP*4 suite to identify the noncrystallographic twofold symmetry.

Structure solution was attempted by molecular replacement using *AMoRe* (Navaza, 1994) as well as *Phaser* (McCoy *et al.*, 2005) from the *CCP*4 suite (Collaborative Computational Project, Number 4, 1994). The homologous *E. coli* CRP complex with cAMP (PDB code 1i5z; Weber & Steitz, 1987) having 32% sequence identity was taken as a search model for molecular replacement. Molecular replacement was also attempted with the reduced CO-sensing protein from *R. rubrum* as the search model (PDB code 1ft9; Lanzilotta *et al.*, 2000).

3. Results and discussion

M. tuberculosis CRP/FNR (Rv3676) is a protein of 224 amino acids, which was purified to homogeneity from an *E. coli* heterologous expression system. The use of His_6 from the pET28a vector led to the introduction of ten additional amino acids in addition to the six histidines into the protein. The protein purity was observed to be better then 99% on SDS–PAGE. The yield of protein was 45 mg from 1 l culture. Upon gel filtration, the protein eluted at a Stokes radius consistent with a dimer, which is in keeping with the quaternary structures observed for other transcription factors of this family.

The protein was stable at room temperature, but was found to precipitate at 253 K. The freshly purified protein was used for crystallization trials and crystals were obtained after one week when the crystallization plate was incubated at a constant temperature of 277 K. The crystals were very unstable at room temperature and often dissolved even while being inspected under an optical microscope. Surprisingly, when the plates were re-incubated at 277 K, good diffraction-quality crystals reappeared in the plates (Fig. 1). These crystals were therefore quickly mounted in cryoloops, frozen and used for data collection.

The completeness of the data was found to be 99.5% (Table 1). The 2.9 Å data were processed using the *HKL*-2000 program suite and the crystal was found to belong to space group $P2_12_12_1$. The data-collection statistics are shown in Table 1. A twinning test (http://nihserver.mbi.ucla.edu/Twinning/) showed that the data were not twinned. Assuming the presence of two monomers in the asymmetric unit, a value of the Matthews coefficient of 2.58 Å^3 Da^{-1} was

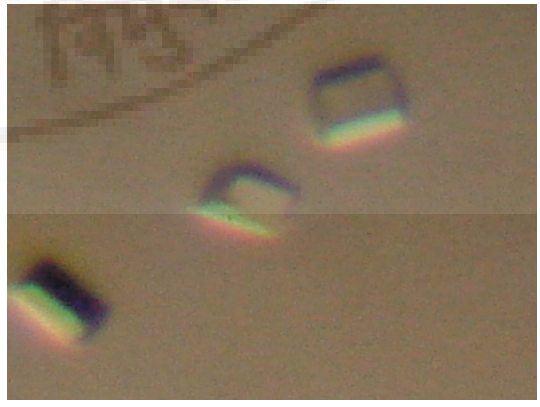


Figure 1
Crystals of CRP/FNR transcription factor grown using 15% ethanol, 0.2 *M* Li_2SO_4 and 0.1 *M* sodium phosphate citrate pH 5.5. The size of each crystal is approximately 0.1 × 0.15 × 0.1 mm.

Table 1
Diffraction data statistics.

Values in parentheses are for the last resolution shell (3.00–2.90 Å).

Wavelength (Å)	1.0
Resolution (Å)	50–2.9
Space group	$P2_12_12_1$
Unit-cell parameters (Å)	a = 54.1, b = 84.6, c = 101.2
Completeness (%)	99.5 (99.4)
$R_{\rm merge}$† (%)	8 (27.3)
$I/\sigma(I)$	16.2 (5.1)
Unique reflections	10769 (1056)
Total reflections	168417
Redundancy	15.6
Matthews coefficient (Å^3 Da^{-1})	2.58
No. of molecules in ASU	2

† $R_{\rm merge} = \sum |I_{hkl} - \langle I_{hkl}\rangle| / \sum |I_{hkl}|$, where I_{hkl} are the intensities of symmetry-redundant reflections and $\langle I_{hkl}\rangle$ is the average over all reflections.

obtained, which corresponds to a solvent content of 52.4% (Matthews, 1968). Because of the presence of two molecules in the asymmetric unit, the self-rotation function was used to determine noncrystallographic symmetry. The two peaks at $\varphi = 90^\circ$ and $\omega = 30$ and 60° from the c^* axis in a self-rotation function plot show the expected twofold noncrystallographic symmetry. However, a satisfactory solution using molecular replacement could not be obtained with either *E. coli* CRP or the CO-sensing protein from *R. rubrum*.

The structural details of interactions between transcription factors and a specific DNA sequence is well established for the cAMP receptor (CRP) family of transcription factors. As in the catabolic gene activator CAP (Busby & Ebright, 1999) and the CO-sensing protein, the binding of cAMP switches the protein from an off-state conformation (refractile to DNA binding) to an on-state conformation (allowing DNA binding). In CAP, the binding of cAMP alters the DNA-binding domain. The alteration of the conformation of the DNA-binding domain, which is more than 20 Å away, is not well understood, primarily because the available structures of CAP are either in the presence of cAMP alone or as a cAMP and DNA complex (Schultz *et al.*, 1991; Weber & Steitz, 1987). Therefore, the nature of conformational changes that take place upon cAMP binding will be better understood if the structure of CAP is known in the absence of the activator cAMP.

In conclusion, we have crystallized the CRP/FNR family transcription factor from *M. tuberculosis*. However, the low sequence homology with other known CRPs and the non-availability of the uncomplexed CRP structure meant that we could not obtain molecular-replacement solutions. This suggests the need to obtain a structure solution using experimental phasing techniques such as multi-wavelength anomalous dispersion (MAD) or multiple isomorphous replacement (MIR).

We thank the staff of the XRD1 beamline at the ELETTRA synchrotron, Trieste, Italy for assistance during data collection and the Indo–Italian POC-DST Action for financial assistance in performing experiments at ELETTRA. This work is supported by grants from the Department of Biotechnology and the Council of Scientific and Industrial Research (CSIR), New Delhi, India. MA and YA gratefully acknowledge financial support from the CSIR for a Senior Research Fellowship and Junior Research Fellowship, respectively; SCM is an International Senior Research Fellow of the Welcome Trust, UK.

References

Aiba, H., Fujimoto, S. & Ozaki, N. (1982). *Nucleic Acids Res.* **10**, 1345–1361.
Busby, S. & Ebright, R. H. (1999). *J. Mol. Biol.* **293**, 199–213.
Cole, S. T. *et al.* (1998). *Nature (London)*, **393**, 537–544.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Green, J., Scott, C. & Guest, J. (2001). *Adv. Microb. Physiol.* **44**, 1–34.
Kaufmann, S. H. (2001). *Nature Rev. Immunol.* **1**, 20–30.
Korner, H., Soab, H. J. & Zumft, W. G. (2003). *FEMS Microbiol. Rev.* **27**, 559–592.
Lanzilotta, W. N., Schuller, D. J., Thorsteinsson, M. V., Kerby, R. L., Roberts, G. P. & Poulos, T. L, (2000). *Nature Struct. Biol.* **7**, 876–880.
McCoy, A. J., Grosse-Kunstleve, R. W., Storoni, L. C. & Read, R. J. (2005). *Acta Cryst.* D**61**, 458–464.
Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.
Navaza, J. (1994). *Acta Cryst.* A**50**, 157–163.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Passner, J. M., Schultz, S. C. & Steitz, T. A. (2000). *J. Mol. Biol.* **304**, 847–859.
Rickman, L., Scott, C., Hunt, D. M., Hutchinson, T., Menendez, M. C., Whalan, R., Hinds, J., Colston, M. J., Green, J. & Buxton, R. S. (2005). *Mol. Microbiol.* **56**, 1274–1286.
Schultz, S. C., Shields, G. C. & Steitz, T. A. (1991). *Science*, **253**, 1001–1007.
Ulrichs, T. & Kaufmann, S. H. (2006). *J. Pathol.* **208**, 261–269.
Weber, I. T. & Steitz, T. A. (1987). *J. Mol. Biol.* **198**, 311–326.
World Health Organization (2006). *Tuberculosis Fact Sheet No. 104: Global and Regional Incidence.* Geneva: World Health Organization.

FEBS Letters 580 (2006) 1285–1293

Clusters of PE and PPE genes of *Mycobacterium tuberculosis* are organized in operons: Evidence that PE Rv2431c is co-transcribed with PPE Rv2430c and their gene products interact with each other

Smanla Tundup[a], Yusuf Akhter[a], Dorairajan Thiagarajan[c], Seyed E. Hasnain[a,b,d,*]

[a] *Laboratory of Molecular and Cellular Biology, Centre For DNA Fingerprinting and Diagnostics, Hyderabad, AP 500 076, India*
[b] *Jawaharlal Nehru Centre for Advanced Scientific Research, Jakkur, Bangalore 560 064, India*
[c] *Indian Immunologicals Ltd., Hyderabad 500 019, India*
[d] *University of Hyderabad, Hyderabad, 500 046, India*

Received 9 November 2005; revised 29 December 2005; accepted 14 January 2006

Available online 23 January 2006

Edited by Judit Ovádi

Abstract **About 10% of the coding capacity of the *Mycobacterium tuberculosis* (*M. tb*) genome is devoted to the PE/PPE family of genes scattered throughout the genome. We have identified 28 PE/PPE operons which are organized within the *M. tb* genome in such a way that most PE members are upstream to PPE members. One example of such a gene arrangement is the PPE gene Rv2430c, earlier shown by us to code for a highly antigenic protein eliciting strong B-cell responses in TB patients [Choudhary, R.K., Mukhopadhyay, S., Chakhaiyar, P., Sharma, N., Murthy, K.J.R., Katoch V.M. and Hasnain, S.E. (2003) PPE antigen Rv2430c of *Mycobacterium tuberculosis* induces a strong B cell response. Infect. Immun. 71, 6338–6343], situated downstream to PE gene Rv2431c. Rv2431c and Rv2430c are transcribed as an operon. Expression of either rRv2431c or rRv2430c alone in *E. coli* limited their localization to the inclusion bodies. However, when they were co-expressed, both the proteins appeared in the soluble fraction. These two proteins interact with each other and form oligomers when alone, however, when present together they exist as heteromer.**

Keywords: PE/PPE; Rv2431c; Rv2430c; *Mycobacterium tuberculosis*; Protein–protein interaction

1. Introduction

One of the major findings from the genome sequence of *Mycobacterium tuberculosis* (*M. tb*) was the presence of the PE/PPE family of genes representing about 10% of the coding capacity [1,2]. These families of genes encode proteins carrying Proline-Glutamic acid (PE) or Proline-Proline-Glutamic acid (PPE) motifs found near the N-terminus and hence the name PE/PPE. The conserved N-terminal and variable C-terminal sequences of the proteins coded by members of this family are thought to confer antigenic variation among different strains of *M. tb* [1]. The PE family has been classified into several subfamilies. The largest of these is the polymorphic repetitive sequence class (PGRS), characterized by high glycine content, also referred to as fibronectin binding proteins [3,4]. Based on the amino acid sequence at conserved positions, the PPE protein family has been classified into three categories. Proteins with major polymorphic tandem repeats (PPE-MPTR) is the major subfamily with other subfamily having conserved GxxSVPxxW motif and the last one being a subfamily of unrelated proteins [1,5]. Some of these proteins are shown to exhibit surface localization [6–8]. With very little understanding of their functional, immunological and other role, these families of proteins have been nonetheless shown to exhibit varying degrees of polymorphism between different clinical isolates of *M. tb* (H37Rv) and CDC1551 [9]. Mutation in the two of PE_PGRS genes of *Mycobacterium marinum* homologues of *M. tb* Rv3812 and Rv1651c, rendered *M. marinum* strains incapable of replication in macrophages and also resulted in decreased persistence in granulomas [10]. The individual role of PE and PGRS domains of a PE_PGRS family member Rv1818c was investigated. It was shown that the DNA construct as well as the protein corresponding only to the PGRS domain showed humoral responses when immunized in mice. However, when mice were immunized with PE domain, more interferon gamma was secreted from the splenocytes as compared to immunization with PGRS domain. These results suggest that PE based vaccine can elicit an effective cellular immune response, which is influenced by the Gly-Ala rich PGRS domain [11]. We earlier showed that a PPE family protein, Rv2608, a member of the major polymorphic tandem repeat (MPTR) subfamily, elicits high humoral response and low T-cell response in TB patients [12]. Since many of the PE/PPE proteins are known to elicit strong immune response [6,11–14, Hasnain et. al, (unpublished)], their functional importance however, in terms of host pathogen interaction has not been addressed adequately. The PPE gene Rv0915c, separated from Rv0916c by 14 bp intergenic region, was found to develop both CD4 and CD8 specific T cell responses and could provide protection against *M. tb* comparable to *M. bovis* BCG vaccination when immunized in C57BL/6 mice [13]. More recently a *M. marinum* homologue of *M. tb* PPE gene Rv1787 was found to be associated with the ability of *M. tb* to grow in macrophages and also virulence in mice [15]. PPE ORF Rv1787 is separated from a PE ORF Rv1788 by 78 bp, which is followed by ORF Rv1789 with 13 bp intergenic region.

The PE/PPE family of genes is organized such that they are scattered throughout the genome. We now describe the observation that a significant number of the PE/PPE genes appear

*Corresponding author. Fax: +91 40 23011090.
E-mail address: seh@uohyd.ernet.in (S.E. Hasnain).

0014-5793/$32.00
doi:10.1016/j.febslet.2006.01.042

to be arranged in a definite pattern of operonic arrangement, where a PE is followed by a PPE gene, separated by less than 90 bp. The immunodominant Rv2430c [14] is a typical example of an arrangement where a PE gene Rv2431c, which codes for another immunodominant protein [unpublished] and shown to be upregulated in an *IdeR* mutant in microarray experiments, precedes Rv2430c [16] by 46 bp intergenic region. We show for the first time that PE (rRv2431c) and PPE (rRv2430c) proteins interact with and solubilize each other only when co-expressed. When expressed alone, both the proteins rRv2430c (PPE) and rRv2431c (PE) go into inclusion bodies. We further demonstrate the ability of the rRv2431c and rRv2430c to hetero-oligomerize. Based on this, we propose that these PE/PPE gene pairs are coregulated, co-expressed, interact functionally, exist in dynamic form and help each other either in secretion or surface localization or in some other as yet unknown cellular functions.

2. Materials and methods

2.1. Bacterial strains and growth media

E. coli strains (DH-5α, XL-1 blue and BL-21 cells) were grown in Luria Bertani, LB medium, obtained from Hi-Media, India. Antibiotics were used at the following concentrations (μg/ml): ampicillin, 100; kanamycin, 50; tetracycline, 15.

2.2. Primers and plasmids

All the primers used in this study (Table 1) were procured from Sigma, USA. The plasmids pBluescript II KS (Stratagene, USA), pQE30 (Qiagen, Inc., USA), pETDuet (Invitrogen, USA) and pET23a (Novagen, EMD Biosciences, Inc., Germany) were used to construct recombinant plasmids described in this study (Table 2). The clones were confirmed by DNA sequencing.

2.3. Molecular cloning and purification of recombinant proteins

The ORF Rv2431c was amplified from the *Mycobacterium tuberculosis* genome using the primers FP2431c (P1) and RP2431c (P2) bearing the restriction sites *Nde*I and *Hin*dIII containing 6×His tag at the 5′ of the reverse primer. The 300 bp amplicon was cloned in pBSK(+) linearized with *Sma*I and further subcloned as a *Nde*I *Hin*dIII fragment in the corresponding sites of pET23a. For co-expression study Rv2431c::pET23a was digested with *Nde*I and *Xho*I and the resultant 300 bp fragment was cloned in the corresponding sites within MCSII of pETDuet expression vector. The *Bam*HI and *Hin*dIII fragment of ORF Rv2430c from pQE30 expression vector was subcloned into MCSI of pETDuet expression vector in which ORF Rv2431c was already cloned at MCSII. The clone pETDuet:30:31 was digested with *Bam*HI, end filled with Klenow Polymerase and religated in order to bring the gene in frame. The plasmids were transformed and expressed in *E. coli* BL-21 strain and the recombinant proteins were purified using Cobalt affinity chromatography. The individual proteins rRv2430c and rRv2431c were purified by on-column refolding method using 8–0 M urea gradient [17]. The co-expressed proteins were purified with 200 mM Imidazole in 1× PBS from soluble fraction by Cobalt affinity chromatography.

Table 1
List of primers used in this study with their sequence

Primer ID	Sequence (5′–3′)
P1-FP2431c	TCCATATGTCTTTTGTGATCACAAAT
P2-RP2431c	TAAAGCTTAGTGGTGGT GGTGGTGGTGA CTAA AGGTCTTGACGTTGTC
P3-RP2430c	AAGCTTCTAAGTGTCTGTACGCGATGA

2.4. Reverse transcriptase polymerase chain reaction (RT-PCR)

Total *M. tb* RNA was a kind gift from AstraZeneca, Bangalore. First-strand synthesis was carried out using Moloney Murine Leukemia Virus Reverse Transcriptase (M-MLV RT) from Invitrogen using Reverse primer RP2430c (Table 1) specific to ORF Rv2430c following the conditions specified by the manufacturer. Subsequent second-strand synthesis was performed using *AccuTaq* Polymerase (Sigma) using primers FP2431c (P1), RP2431c (P2) and RP2430c (P3). The total RNA of *M. tb* was used to carry out RT-PCR. The annealing temperatures used were 41 °C for Rv2431c and 48 °Cfor Rv2431c + Rv2430c amplification. The PCR product was analyzed on 1.5% agarose gel.

2.5. UV and chemical crosslinking analysis

The purified proteins rRv2430c, rRv2431c and co-purified rRv2430c and rRv2431c were exposed at room temperature to UV at 254 nm wave length and energy 1200×100 μJ/CM^2 for 30, 60, 120, 180 and 300 s, respectively. The buffer used was 1× PBS and it was the same buffer in which protein was purified. For chemical crosslinking, the proteins were incubated with 0.1%, 0.2%, 0.3%, 0.5%, 1% and 2% glutaraldehyde for 20 min at room temperature. The samples were fractionated using Tris–Tricene SDS–polyacrylamide gel electrophoresis (PAGE) for crosslinking analysis.

2.6. Western blot analysis

For Western blot analysis, equal amount of purified proteins were separated using Tris–Tricine SDS–PAGE. After electrophoresis, the protein samples were transferred to nitrocellulose membrane using a Trans-Blot electrophoresis transfer cell (Pharmacia Biotech, USA). Western blot analysis was conducted using antisera to rRv2431c (1:1000) and rRv2430c (1:300) generated in mice followed by incubation with horseradish peroxidase-conjugated secondary antibody (1:10 000). To quantify the protein, a chemiluminescent signal was developed using detection reagents from ECL Plus kit (Amersham Pharmacia Biotech, England) and the signal was recorded on X-ray film (Hyperfilm Amersham, Pharmacia Biotech, UK).

2.7. Gel filtration analysis

Gel filtration (BioRad, USA) was used to analyze the purified proteins. About 250 μg of each recombinant protein was loaded on Superose 6 column and compared with protein molecular size standards. The corresponding peaks were collected and analyzed on Tris–Tricene SDS–PAGE.

Table 2
Plasmids constructs used in this study

Plasmids	Specification	Reference
pBSK::31	PCR amplified 300 bp ORF Rv2431c cloned in PBSK(+) vector at *Smal* site	This study
pET23a::31	*Nde*I/*Hin*dIII Rv2431c fragment derived from PBSK::31 cloned in *Nde*I/*Hin*dIII site of pET23a vector	This study
pQERv2430c	*Bam*HI/*Hin*dIII generated 585 bp fragment from pGEMT-easy:2430 vector cloned *Bam*HI/*Hin*dIII site of pQE30 vector	[14,17]
pETDuet:30	*Bam*HI/*Hin*dIII generated 585 bp fragment from pQERv2430c cloned in *Bam*HI/*Hin*dIII site of MCSI of pETDuet	This study
pETDuet:30:31	*Nde*I/*Xho*I generated 300 bp fragment from pET23a::31 cloned in *Nde*I/*Xho*I site of pETDuet:30 vector	This study

2.8. Spectrometric analyses

Fluorescence spectra of purified recombinant proteins were recorded on a Perkin–Elmer LS-3B Spectrofluoremeter. Fluorescence studies were carried out in presence of increasing concentration of NaCl (140–2000 mM) for both rRv2430c and rRv2430c-rRv2431c complex. The samples were incubated for 5 min prior to excitation. In each measurement the protein was excited at 295 nm and emission spectra were recorded from 300 to 480 nm. Each fluorescence spectrum was obtained by averaging three different recordings. The slits of excitation and emission were kept at 5 nm and scan speed at 50 nm/s. Throughout the experiment, the temperature was maintained at 25 °C. 1 μM of protein in 1× PBS was used for each reading.

3. Results

3.1. Significantly large number of the PE/PPE genes are organized as an operon in a definite pattern

The organization, within the *M. tb* genome, of the PE/PPE family of genes which code for highly antigenic proteins, was analyzed. Intergenic region less than 90 bp was used to predict a probable operon. The maximum number of base pairs in a given sequence to contain any ρ-dependent termination is at least 100 bp. If two given genes are separated by less than 90–100 bp, the possibility for the two genes to be in operon would be true if the intergenic region does not have any ρ-independent termination sequence. Only 3 ρ-independent termination sequences were previously predicted in the entire *M. tb* (H37Rv) genome using software TRANSTERM available at TIGR (USA) website, http://www.tigr.org/software/TransTermResults/ntmt02.html. None of the ρ-independent termination falls into intergenic sequence of the predicted operons. A total of 41 operons were short-listed in this analysis. Of these 41 operons, 28 operons contain only the PE/PPE genes. The remaining are clusters of either PE or PPE genes with non-PE/PPE members. Some of the genes such as Rv0915c, Rv2430c, Rv3872 and a homologue of Rv1787 in *M. marinum* [13–15,18], belonging to the PE/PPE family which plays a significant role in the host pathogen interaction, are also predicted in our analysis to be in operon with the other PE/PPE genes. There was a definite pattern of arrangement of these gene pairs where, in most cases, a bi-cistronic organization could be seen with a PE gene followed by a PPE gene in the operon (Table 3). Some others involve either PE or PPE with the non-PE/PPE members, including ESAT-6 like secretory proteins.

3.2. PPE ORF Rv2430c and PE ORF Rv2431c are co-transcribed at the mRNA level

Having predicted and short-listed ORFs Rv2431c and Rv2430c as belonging to one operon, RT-PCR was carried out with total RNA isolated from *M. tb* as described in Section 2. After the first strand synthesis with primer P3, the reverse transcribed template was used to carry out two different sets of PCR for checking the whole operonic amplification and also for Rv2431c. PCR amplification was carried out with primer pairs P1 and P2, corresponding to Rv2431c, which showed an amplification of 300 bp (Fig. 1B, lane 3) and P1 (FP2431c, specific to ORF Rv2431c) and P3 (RP2430c, specific to Rv2430c), gave an amplification of 931 bp corresponding to the anticipated size of both Rv2431c and Rv2430c (Fig. 1B, lane 4) including the intergenic sequence of 46 bp. The specific amplification of the 931 and 300 bp fragments

Table 3
Probable operonic organization of PE/PPE genes based on intergenic distance

Genes	Intergenic region (bp)	Function
Rv0096	18	PPE
Rv0097	Overlapping by 4[a]	Possible oxidoreductase
Rv0098	4	Hypothetical Protein
Rv0099	4	FadD10
Rv0100	Overlapping by 19[a]	Hypothetical protein
Rv0101		nrp
Rv0152c	9	PE
Rv01S1c		PE
Rv0256c	18	PPE
Rv0255c	24	cobQ
Rv0254c		cobU
Rv0280	23	PPE
Rv0281		Hypothetical protein
Rv0282	Overlapping by 4[a]	Hypothetical protein
Rv0283	Overlapping by 4[a]	Conserved membrane protein
Rv0284	Overlapping by 4[a]	FtsK
Rv0285	2	PE
Rv0286	48	PPE
Rv0287	29	esxG
Rv0288	10	esxH
Rv0289	46	Conserved membrane protein
Rv0290	Overlapping by 4[a]	Conserved membrane protein
Rv0291	Overlapping by 4[a]	mycP3
Rv0292	Overlapping by 14[a]	Conserved membrane protein
Rv0293		Hypothetical protein
Rv0304	55	PPE
Rv0305		PPE
Rv0354	82	PPE
Rv0355		PPE
Rv0388c	52	PPE
Rv0387c		Hypothetical protein
Rv0442c	26	PPE
Rv0441c		Hypothetical protein
Rv0745	19	Hypothetical protein
Rv0746		Hypothetical protein
Rv0832	Overlapping by 4[a]	PE_PGRS
Rv0833		PE_PGRS
Rv0916c	14	PE
Rv0915c		PPE
Rv0980c	81	PE_PGRS
Rv0979c		Hypothetical protein
Rv1040c	76	PE
Rv1039c		PPE
Rv1088	Overlapping by 179[a]	PE
Rv1089		PE
Rv1168	17	PE
Rv1169		PPE
Rv1195	46	PE
RV1196		PPE
Rv1386	Overlapping by 4[a]	PE
Rv1387		PPE

(*continued on next page*)

Table 3 (*continued*)

Genes	Intergenic region (bp)	Function
Rv1646	77	PE
Rv1647	6	Hypothetical protein
Rv1648		Hypothetical protein
Rv1706c	39	PPE
Rv1705c	40	PPE
Rv1704c	64	cycA
Rv1703c		Predicted *o*-methyltransferase
Rv1787	78	PPE
Rv1788	13	PE
Rv1789		PPE
Rv1806	14	PE
Rv1807		PPE
Rv2099c	Overlapping by 4[a]	PE
Rv2098c	51	PE_PGRS
Rv2097c	8	Hypothetical protein
Rv2096c	Overlapping by 4[a]	Transcription regulator
Rv2095c	67	Predicted transcription
Rv2094c	16	Regulator
Rv2093c	48	Hypothetical protein
Rv2092c	41	tatc
Rv2091c		helY
		Hypothetical protein
Rv2107	55	PE
Rv2108		PPE
Rv2431c	**46**	**PE**
Rv2430c		**PPE**
Rv2489c	Overlapping by 35[a]	Hypothetical protein
Rv2488c	81	Predicted ATPase
Rv2487c		PE_PGRS
Rv2769c	79	PE
Rv2768c		PPE
Rv2774c	10	Hypothetical protein
Rv2773c	11	dapB
Rv2772c	85	Hypothetical protein
Rv2771c	63	Hypothetical protein
Rv2770c		PPE
Rv2853	36	PE_PGRS
Rv2854	12	Hypothetical protein
Rv2855		gorA
Rv3018c	86	PPE
Rv3017c		esxQ(jag)
Rv3019	33	esxR
Rv3020	46	PF
Rv3021	Overlapping by 16[a]	PPE
Rv3022		PPE
Rv3135	61	PPE
Rv3136		PPE
Rv3345c	Overlapping by 281[a]	PE_PGRS
Rv3344c	48	PE_PGRS
Rv3343c		PPE
Rv3477	36	PE
Rv3478	50	PPE
Rv3479		Possible Transmembrane Protein
Rv3511	Overlapping by 38[a]	PE_PGRS
Rv3512		PE_PGRS

Table 3 (*continued*)

Genes	Intergenic region (bp)	Function
Rv3622c	11	PE
Rv3621c	56	PPE
Rv3620c	26	esxW
Rv3619c		esxV
Rv3652	Overlapping by 4[a]	PE_PGRS
Rv2653		PE_PGRS
Rv3738	51	PPE
Rv3739		PPE
Rv3746c	72	PPE
Rv3745c		Hypothetical protein
Rv3872	30	PE
Rv3873		PPE
Rv3893c	78	PE
Rv3892c		PPE

[a]Overlap between the two genes resulting in absence of any spacer.

confirms the existence of the two ORFs, Rv2431c and Rv2430c, as part of a single operon. No amplification was seen when the PCR was carried out with RNA sample alone (without making cDNA) using primer pairs P1 and P3. This rules out DNA contamination during the process of total RNA preparation (Fig. 1B, lane 2). Assuming that the genes separated by short intergenic sequences tend to have related function and interact physically, we performed further experiments to investigate interaction between these two proteins, if any.

3.3. *Recombinant proteins rRv2431c and rRv2430c form inclusion bodies when over-expressed alone, but appear in soluble fraction when co-expressed together in* E. coli

Having shown that the existence of ORF Rv2431c and Rv2430c as one operon separated by a 46 bp intergenic sequence, we designed experiments to investigate their interaction at the protein level. The PE/PPE proteins are mostly known to form inclusion bodies when over-expressed in *E. coli* [12,14,17]. The recombinant PPE protein Rv2430c (23 kDa) was earlier shown to remain in the inclusion body when expressed alone [14,17]. Similarly, the recombinant PE protein Rv2431c (11 kDa) when over-expressed alone in *E. coli* also remains within inclusion bodies (Fig. 2A). In order to study whether the two proteins are solubilized when co-expressed in *E. coli*, both the genes were cloned into pETDuet expression vector to construct expression plasmid pETDuet:30:31. When pETDuet:30:31 was used to transform *E. coli* BL-21 cells, it could be seen (Fig. 2B) that the supernatant contains much (50–60%) of the recombinant 23 and 11 kDa proteins. These results indicate that expression of these two genes together results in the corresponding proteins getting into the soluble fraction. Both the proteins contain histidine tags either at their N-terminal (rRv2430c) or the C-terminal (rRv2431c). This property was exploited to affinity purify the coexpressed proteins from the supernatant to 90–95% homogeneity for further studies (Fig. 2C).

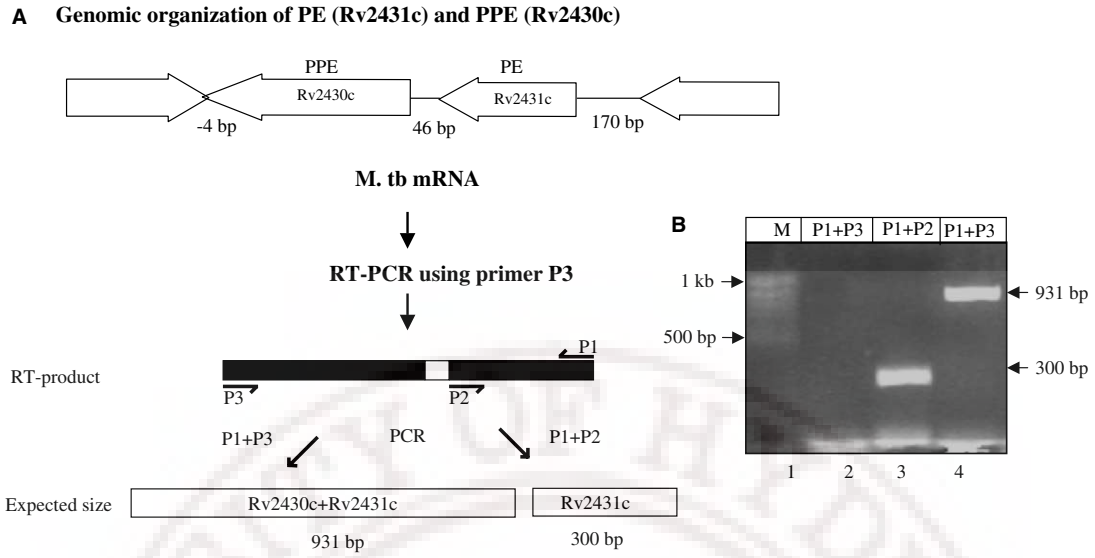


Fig. 1. RT-PCR analysis demonstrates that Rv2430c and Rv2431c are co-transcribed as an operon. (A) Schematic representation of the organization of Rv2431c and Rv2430c ORFs in *Mycobacterium tuberculosis* genome. The intergenic sequence between the Rv2431c and Rv2430c is 46 bp. The position of the various primers and the corresponding predicted amplicons are shown. (B) PCR amplification generates specific amplicons pointing to the organizational arrangement of Rv2430c and Rv2431c within an operon. The product of the reverse transcription reaction, using *M. tb* RNA and P3 primer, was used as template for the subsequent PCR amplification. Lane 2: PCR amplification, using forward primer P1 specific to Rv2431c and reverse primer P3 specific to Rv2430c, using *M. tb* total RNA as template did not generate any amplicon thereby ruling out possible DNA contamination during RNA preparation. RT-product amplified by using the primer P3 was used as template to carry out further PCR. Lane 3: the 300 bp PCR amplification of Rv2431c generated by using primer P1 and P2. Lane 4: the 931 bp amplicon corresponding to Rv2431c and Rv2430c, using primer pair P1 and P3. Lane 1: 50 bp DNA size marker.

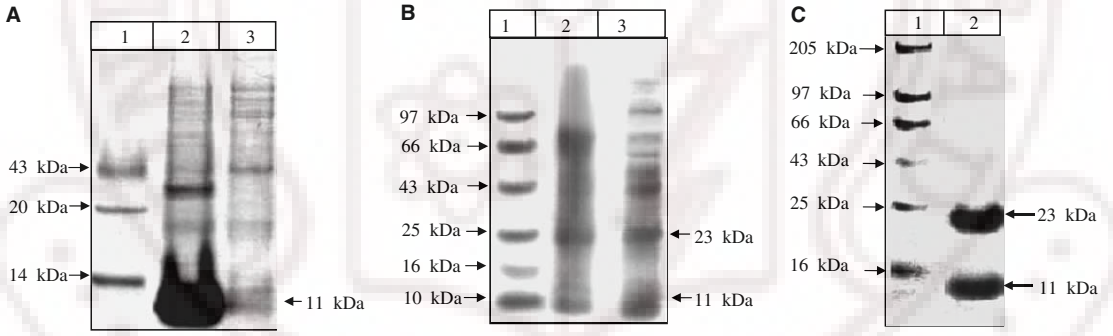


Fig. 2. SDS–PAGE analysis shows that PE (rRv2431c) and PPE (rRv2430c) proteins solubilize each other. (A) Expression of rRv2431c in *E. coli* BL-21 strain. Lane 2 shows that the recombinant Rv2431c protein forms inclusion bodies and is mostly in pellet and very little, if any, in supernatant (lane 3). (B) Co-expression of rRv2431c and rRv2430c. Equal amounts of supernatant and pellet were fractionated on 10% Tris–Tricene SDS–PAGE. Lane 2 shows the two over-expressed bands, corresponding to the co-expressed rRv2431c and rRv2430c proteins, in soluble form and lane 3 shows the two proteins in inclusion bodies. (C) Co-purification of rRv2431c and rRv2430c using COBALT affinity column. Lane 2 shows the purification of rRv2431c and rRv2430c. Both the proteins are Histidine tagged at C-terminal and N-terminal, respectively. Protein molecular size marker is shown in lane 1. Coomassie Blue G250 was used to stain the SDS–polyacrylamide gel.

3.4. rRv2431c and rRv2430c proteins interact with each other as evident from protein crosslinking and Western blot analysis

Having shown that the two proteins expressed together are in the soluble form, the physical interaction between them, if any, was further checked using UV (Fig. 3A) or glutaraldehyde crosslinking (Fig. 3B). Exposure to UV causes crosslinking of interacting proteins (Smanla et al., manuscript in preparation). We therefore used this simple UV crosslinking to determine direct interaction between rRv2431c and rRv2430c. The co-purified rRv2430c and rRv2431c protein fraction was exposed to UV for different times, 30 s to 5 min. The individual proteins were also exposed to UV for 2 min in the presence of an unrelated protein rRv2626c (16 kDa), which served as negative control. SDS–PAGE analysis of UV cross-linked co-purified protein reveals a band corresponding to 34 kDa (Fig. 3A, lanes 7–11), which is not seen when rRv2430c (Fig. 3A, lanes 2, 3) or rRv2431c (lanes 4 and 5) was mixed with unrelated recombinant protein rRv2626, whether untreated with UV (Fig. 3A, lanes 2 and 4) or after UV treatment (Fig. 3A, lanes 3 and 5). The presence of a 34 kDa protein band corresponding to 23 and 11 kDa protein suggests an interaction between the proteins rRv2430c and rRv2431c, generating a complex as a consequence of UV crosslinking. In order to confirm the data from UV crosslinking, glutaraldehyde crosslinking in a

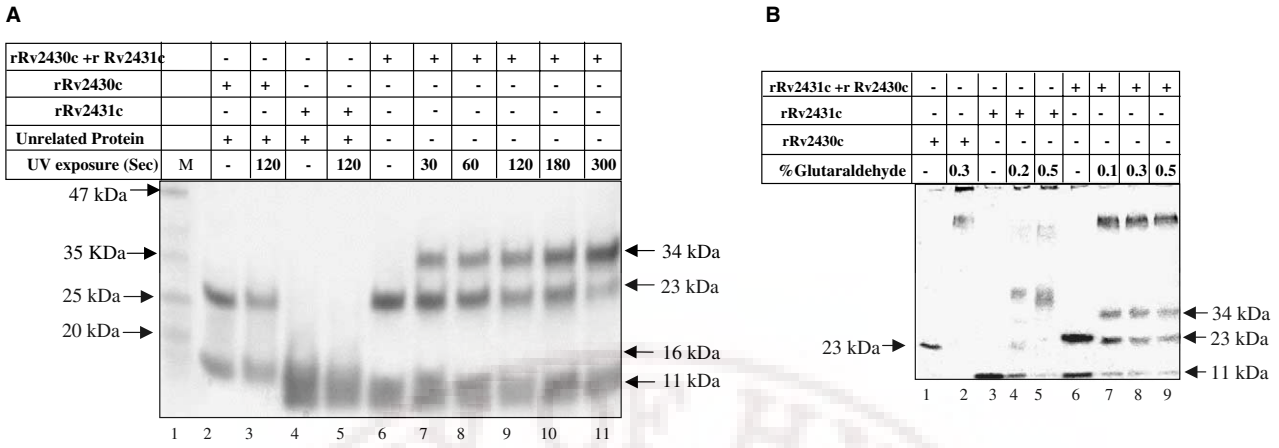


Fig. 3. Crosslinking experiments show PE (rRv2431c) and PPE (rRv2430c) proteins interact with each other. (A) UV crosslinking of rRv2431c and rRv2430c proteins. The recombinant proteins were fractionated on SDS–PAGE with or without UV exposure. Lane 1 represents protein molecular size marker. The other lanes show recombinant proteins without UV exposure (lanes 2, 4, and 6) or exposure to UV for different times (lanes 3, 5, and 7–11). Note that rRv2430c does not interact with an unrelated protein rRv2626c (lane 2) even after UV exposure (lane 3). Similarly rRv2431c fails to interact with the unrelated control protein rRv2626c (lane 4) even after UV exposure (lane 5). The co-expressed rRv2430c and rRv2431c in the absence of UV crosslinking do not form complex (lane 6). Lanes 7–11 show rRv2431crRv2430c complex corresponding to 34 kDa generated as a function of time of UV exposure. (B) Glutaraldehyde crosslinking shows rRv2431c and rRv2430c interact and form oligomer individually. Proteins were incubated for 20 min in the presence or absence of different concentration of glutaraldehyde and fractionated on SDS–PAGE. Lanes 2, 4 and 5 show oligomer formation of proteins at 0.3%, 0.2% and 0.5% glutaraldehyde concentration. Lanes 7–9 show the oligomer band of rRv2430c, the interacting band corresponds to 34 kDa and monomeric forms of individual species. No interacting or oligomer band is seen in proteins without glutaraldehyde (lanes 1 and 3). Coomassie Blue G250 was used to stain the polyacrylamide gel.

concentration ranging from 0.1% to 0.5% was carried out. The individual proteins reveal oligomeric states upon glutaraldehyde crosslinking (Fig. 3B, lanes 2, 4 and 5) but not in the absence of glutaraldehyde (Fig. 3B, lanes 1 and 3). While rRv2430c protein shows two states of oligomerization of very high molecular sizes (Fig. 3B, lane 2), rRv2431c shows more than two oligomeric states corresponding to approximately 22 and 55 kDa with 0.2% glutaraldehyde (Fig. 3B, lane 4). Similar to results obtained from UV crosslinking, glutaraldehyde crosslinking of co-purified rRv2430c and rRv2431c also shows a band of 34 kDa (Fig. 3B, lanes 7–9) corresponding to size of the complex obtained with rRv2430c and rRv2431c. While UV crosslinking shows only the rRv2431c and rRv2430c complex, but not their oligomeric state, glutaraldehyde crosslinking shows both the homomeric and heteromeric forms. The band corresponding to 34 kDa represents the PE/PPE complex rather than the trimeric rRv2431c alone. To confirm that the 34 kDa band is indeed a complex of PE and PPE proteins, Western blot analysis was carried out using anti-rRv2431c and anti-rRv2430c sera. As could be seen in Fig. 4A, lane 2, when the co-purified protein was UV crosslinked and then subjected to Western blot analysis using anti-rRv2431c serum, only the bands corresponding to rRv2431c protein and the 34 kDa PE/PPE complex can be seen, but not the band corresponding to 23 kDa (rRv2430c) protein. Similarly using anti-rRv2430c antibody, bands corre-

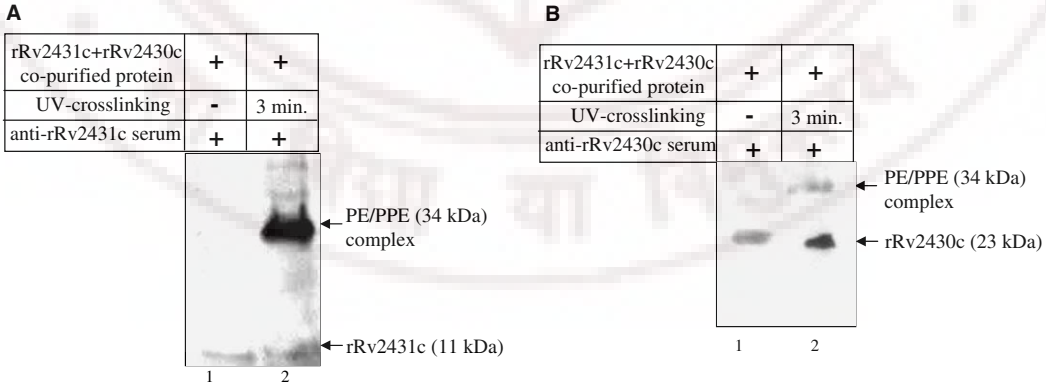


Fig. 4. Western blot analysis confirms interaction between rRv2431c and rRv2430c. Western blotting was carried out using anti-rRv2431c/anti-rRv2430c serum with copurified rRv2431c + rRv2430c protein separated on Tris–Tricine SDS–PAGE with or without exposure to UV. (A) Lane 1: only the 11 kDa band corresponding to rRv2431c could be seen using anti-rRv2431c serum when the co-purified protein was not exposed to UV. Lane 2: the interacting rRv2431c and rRv2430c complex corresponding to 34 kDa as well as the 11 kDa band corresponding to PE rRv2431c could be seen using anti-rRv2431c serum when the sample was exposed to UV for 3 min. (B) Lane 1: 23 kDa band corresponding to rRv2430c is seen when Western blotting was carried out using anti-rRv2430c serum and co-purified rRv2431c + rRv2430c proteins, but without exposure to UV. Lane 2: upon UV exposure bands corresponding to rRv2430c (23 kDa) and the interacting band (34 kDa) could also be seen. The absence of 23 kDa band corresponding to rRv2430c protein using anti-rRv2431c serum (Fig. 4A) and 11 kDa band corresponding to rRv2431c using anti-rRv2430c serum rules out cross-reactivity of the antibodies generated against these two proteins.

sponding only to 23 kDa (rRv2430c) and the interacting rRv2431c:rRv2430c complex (34 kDa) could be seen (Fig. 4B, lane 2), but not the band corresponding to 11 kDa (rRv2431c) protein. A similar pattern was observed when Western blotting was performed after the proteins were subjected to glutaraldehyde crosslinking (data not shown). These results demonstrate that the anti-rRv2431c and anti-rRv2430c antibodies do not cross react with rRv2430c and rRv2431c proteins.

3.5. Gel filtration shows PE protein rRv2431c and PPE protein rRv2430c complex exists in different forms

The tendency of the two proteins to form oligomer when alone and in hetero-oligomeric state is further confirmed by using size exclusion chromatography of individual and the co-purified proteins. Individual rRv2430c shows a wide peak corresponding to more than 600 kDa (Fig. 5, blue curve) while rRv2431c also showed two sharp oligomeric peaks corresponding to approximately 230 and 162 kDa (Fig. 5, purple curve). The gel filtration pattern of the peaks of co-purified rRv2431c and rRv2430c (green curve) is significantly different from the peaks of the individual proteins. The size of the co-purified proteins is much lower, as the V_e of the co-purified protein is higher, than that of the individual proteins. The co-purified proteins show two peaks, the first wider one (V_e = 13.2) and a second sharper peak (V_e = 16.2). SDS–PAGE analysis of

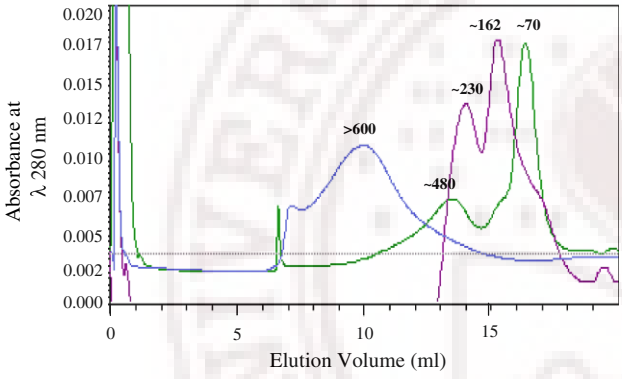


Fig. 5. Size exclusion chromatography of rRv2431c, rRv2430c and rRv2431c– rRv2430c complex proteins shows various states of existence. About 400 μg/ml of each protein was loaded on Superose 6 column. Protein molecular sizes in kDa are also indicated on top of each peak. The different lines are: blue, rRv2430c; purple, rRv2431c; green, co-purified rRv2431c and rRv2430c.

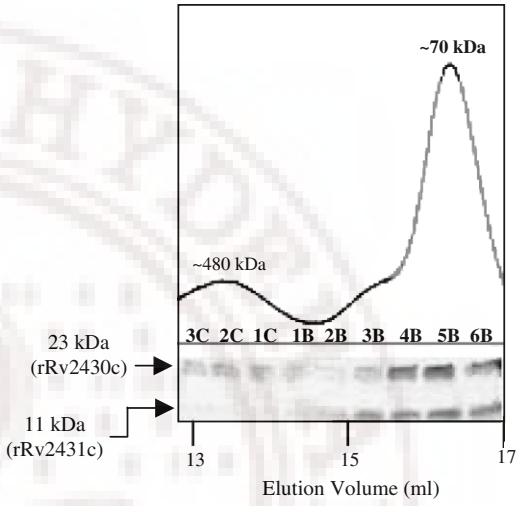


Fig. 6. SDS–PAGE analysis of gel filtration collected fraction shows that heteromer of rRv2431c and rRv2430c exists in higher concentration than individual oligomer. The curve represents rRv2431c + rRv2430c co-purified protein (Fig. 5, green curve), collected fractions of which were analyzed on SDS–PAGE. The elution fractions 3C–1C contain more of rRv2430c species. It can be seen that there is an increase in the concentration of both rRv2431c and rRv2430c fractions in 3B–6B. Fractions 4B–5B correspond to the sharpest peak containing equal amounts of very high concentrations of both the proteins. Elution volume corresponds to each fraction is given in milliliter. Coomassie Blue G250 was used to stain the SDS–polyacrylamide gel.

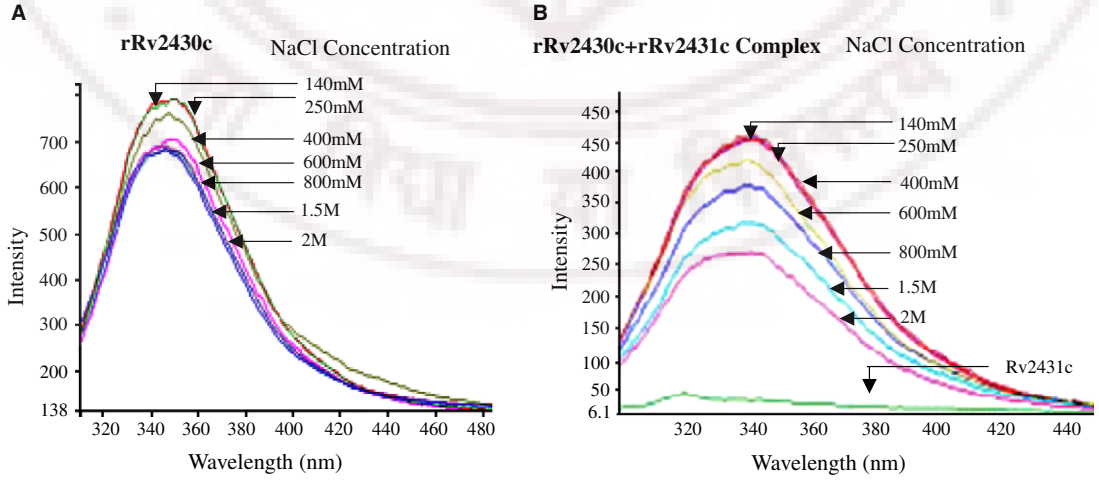


Fig. 7. rRv2431c induces changes in conformation of rRv2430c. The recombinant proteins rRv2430c and co-purified rRv2431c-rRv2430c were excited at 295 nm as a function of increasing salt concentration (140 mM to 2 M) and emission spectra were recorded between 300 and 480 nm. The arrows with respective concentration of salt show the decrease in intensity of fluorescence with increase in salt concentration. The last line in the graph shows the emission spectrum of rRv2431c excited at 295 nm (B). No such change in intensity of fluorescence emitted by rRv2430c was observed with similar increase in salt concentration (A).

each of the fraction collected (Fig. 6) reveals that the first peak contains more species of rRv2430c (480 kDa) in an oligomeric state while the second sharp peak (~70 Da) corresponds to the complex PE/PPE proteins (rRv2431c and rRv2430c) of higher concentration than the rRv2430c oligomer species. The presence of a high molecular size (~70 kDa) of the heteromer, unlike in the crosslinking experiments where a heterodimer of 34 kDa was observed, suggests that heteromer also exists in tetrameric state involving both rRv2430c and rRv2431c. The co-purified proteins mostly exist as heteromer rather than oligomer of one species, suggesting that the interaction between the PE and PPE proteins is stronger leading to heteromeric state rather than interacting individually. As can be seen in Fig. 5, both rRv2431c and rRv2430c exist as oligomer of higher molecular weight than the complex rRv2430c and rRv2431c protein.

3.6. Salt induced changes in conformation of the protein rRv2430c in rRv2430c: rRv2431c complex

The interaction between these two proteins is also supported by fluorescence studies. rRv2431c does not carry any tryptophan residue, therefore, when excited at 295 nm, the fluorescence emission will only be due to rRv2430c if complexed with rRv2431c. The fluorescence emission spectrum was recorded by excitation at 295 nm. It can be clearly seen (Fig. 7A) that increase in concentration of salt in oligomeric rRv2430c did not affect the aromatic residues environment in the protein thereby making it refractile to significant change in the fluorescence intensity at salt concentration as high as 2 M. The rRv2431c: rRv2430c complex however displayed a significant decrease in the fluorescence intensity as a function of increasing salt concentration (Fig. 7B). The decrease in fluorescence intensity brought about by the increase in salt concentration in rRv2431c:rRv2430c co-purified protein (Fig. 7B) is due to the change in conformation of rRv2430c with respect to rRv2431c. rRv2431c expectedly alone does not have an emission spectrum when excited at 295 nm specific to tryptophan residue (Fig. 7B), but otherwise does generate a spectrum when the protein is excited at 285 nm (data not shown).

4. Discussion

The completion of many bacterial genomes has allowed the analysis of gene clusters, leading to interesting conclusions about the tendencies of the genes with related functions to remain together across several genomes [19], particularly in the case of genes whose proteins products physically interact [20]. The organization of genes in the operons is believed to provide the advantage of coordinated regulation and production of functionally and temporally related genes. The PE/PPE genes in *M. tb*, though scattered throughout the genome, are not randomly organized. We investigated the presence, if any, of a defined pattern of genomic organization of the PE/PPE genes. As shown, most of the PE/PPE genes fall into operons either with PE/PPE genes only or with many other genes belong to ESAT-6 like secretory protein. One such typical example of an organization where a PE gene is followed by a PPE family member was selected to understand the effect of such genetic organizational proximity. Two ORFs, PE Rv2431c and PPE Rv2430c, among the other predicted operons was selected on the basis of: (a) their highly antigenic property [14], (b) 30–40% identity with other ORFs forming operon, and (c) micro array data showing upregulation of both the ORFs by 1.6-fold in *IdeR* mutant *M. tb* [16]. RT-PCR analysis of this operon indeed confirmed that these two genes are co-transcribed. The presence of a loose ribosome binding site (ACGGAA), within the 46 bp spacer separating the two ORFs, points to a possibility of the large mRNA generating two translation products corresponding to ORF Rv2431c (11 kDa) and ORF Rv2430c (23 kDa).

We then investigated whether the two proteins, which are the gene products of a single operon, physically associate with each other. The properties of these proteins to interact with each other, was clearly indicated when the two proteins were co-expressed in *E. coli*, which led to localization of both the proteins in soluble form. When over-expressed alone, the protein could only be recovered from inclusion bodies and could be solubilized by in vitro methods [14,18]. The pattern of organization is likely to have a functional significance in facilitating the secretion of the immunodominant Rv2430c. The argument that these proteins may be involved in secretion and membrane localization is further strengthened by organization of some of the PE/PPE genes in operon with those encoding putative transmembrane proteins, and ESAT-6 like secretory proteins (esxG, esxH, esxQ, esxR, esxW and esxV) (Table 3). The predicted PE35 and PPE68 operon falls within the *M. tb* RD1 region and is possibly thought to be responsible for the primary attenuation of *M. bovis* to *M. bovis* BCG. Recombinant secretory protein ESAT-6 was found to induce immune response only when at least 11 genes (which include PE35/PPE68 predicted operon) flanking *esxA* located in RD-1 region was introduced in *M. bovis* BCG, pointing to their role in modulating the secretory apparatus for ESAT-6 secretion [18]. RD1 encompasses the well-known early secreted antigenic target ESAT-6 (*esxA*) and *esxB* [21] which are found to be in operon and interact with each other [22,23]. Recently Brodin et al., have further identified the residues involved in the secretion, complex-formation, virulence and immunogenicity of these two secretory proteins [24]. The PE/PPE member Rv0915 and a *M. marinum* gene (homologue of Rv1787) are predicted to be in operon with upstream PE genes (Table 3) shown to have crucial roles in host–pathogen interaction [13,15]. Irrespective of whether these are surface localized or secreted, they perhaps help the bacterium to survive under various stresses exerted by the host. This also suggests the existence of other secretion systems to deliver the effector proteins to host cells [25–28].

Results of UV as well as glutaraldehyde crosslinking experiments demonstrated the ability of rRv2431c and rRv2430c to physically interact with each other. This was further confirmed by Western blot analysis (Fig. 4A and B). UV crosslinking (Fig. 3A) could capture only the interacting band unlike glutaraldehyde crosslinking, which showed both oligomeric and interacting states (Fig. 3B). This, therefore, indicates that the interaction between the two proteins is stronger than individual oligomer formation. The existence of the PE/PPE mainly as hetero-tetramer (~70 kDa) rather than as oligomer was evident when the gel filtration fractions were analyzed on SDS–PAGE (Fig. 6). Fluorescence spectra of the co-purified protein showed that rRv2431c could bring about a change in conformation of rRv2430c at as high as 2 M concentration of salt when excited at 295 nm. rRv2431c protein does not contain

any tryptophan residues and thus did not give excitation at the particular wavelength. CD analysis of on-column refolded rRv2431c (data not shown) and rRv2430c [17] proteins indicated proper folding, with high content of α-helix, which is in agreement with in silico prediction. However, it failed to show interaction when the two separately purified recombinant proteins were mixed at equimolar concentration and subjected to crosslinking experiments (data not shown), though both the individual proteins tend to oligomerize when subjected to chemical crosslinking (Fig. 3A and B). The co-purified recombinant proteins are highly stable in vitro at room temperature (compared to individual proteins). Likewise when the proteins are individually expressed they go into the inclusion bodies but are solubilized when co-expressed. These results point to the significance of the interaction between the two proteins during the translational processes leading to the appropriate folding in order to remain in native states to perform biological functions. It would be tempting to speculate on the impact of this protein–protein co-operativity in modulating the host immune function given the fact that these two members of the PE/PPE family have immunodominant B-cell function. That such co-operativity between in vivo co-expressed PE/PPE family of proteins is a prelude to a likely immune sensing (quorum sensing) role for these proteins is currently being investigated.

Acknowledgments: This work is partly supported by research grant from DBT (to S.E.H.) and the award of CSIR Junior Research Fellowships (to S.T and Y.A). We wish to thank Krishnaveni for her help during this study.

References

[1] Cole, S.T. et al. (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 393, 537–544.
[2] Chakhaiyar, P. and Hasnain, S.E. (2004) Defining the mandate of tuberculosis research in a postgenomic era. Med. Princ. Pract. 13, 177–184.
[3] Abou-Zeid, C., Garbe, T., Lathigra, R., Wiker, H.G., Harboe, M., Rook, G.A. and Young, D.B. (1991) Genetic and immunological analysis of *Mycobacterium tuberculosis* fibronectin-binding proteins. Infect. Immun. 59 (8), 2712–2718.
[4] Espitia, C., Laclette, J.P., Mondragon-Palomino, M., Amador, A., Campuzano, J., Martens, A., Singh, M., Cicero, R., Zhang, Y. and Moreno, C. (1999) The PE-PGRS glycine-rich proteins of *Mycobacterium tuberculosis*: a new family of fibronectin-binding proteins? Microbiology 145, 3487–3495.
[5] Adindla, S. and Guruprasad, L. (2003) Sequence analysis corresponding to the PPE and PE proteins in *Mycobacterium tuberculosis* and other genomes. J. Biosci. 28, 169–179.
[6] Banu, S., Honore, N.B., Saint-Joanis, B., Philpott, D., Prevost, M.C. and Cole, S.T. (2002) Are the PE-PGRS proteins of *Mycobacterium tuberculosis* variable surface antigens? Mol. Microbiol. 44, 9–19.
[7] Brennan, M.J., Delogu, G., Chen, Y., Bardarov, S., Kriakov, J., Alavi, M. and Jacobs Jr., W.R. (2001) Evidence that mycobacterial PE_PGRS proteins are cell surface constituents that influence interactions with other cells. Infect. Immun. 69, 7326–7333.
[8] Sampson, S.L., Lukey, P., Warren, R.M., van Helden, P.D., Richardson, M. and Everett, M.J. (2001) Expression, characterization and sub cellular localization of the *Mycobacterium tuberculosis* PPE gene Rv1917c. Tuberculosis 81, 305–317.
[9] Fleischmann, R. et al. (2002) Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. J. Bacteriol. 184, 5479–5490.
[10] Ramakrishnan, L., Federspiel, N.A. and Falkow, S. (2000) Granulomaspecific expression of *Mycobacterium* virulence proteins from the glycine-rich PE-PGRS family. Science 288, 1436–1439.
[11] Delogu, G. and Brennan, M.J. (2001) Comparative immune response to PE and PE_PGRS antigens of *Mycobacterium tuberculosis*. Infect. Immun. 69, 5606–5611.
[12] Chakhaiyar, P., Nagalakshmi, Y., Aruna, B., Murthy, K.J., Katoch, V.M. and Hasnain, S.E. (2004) Regions of high antigenicity within the hypothetical PPE major polymorphic tandem repeat open-reading frame, Rv2608, show a differential humoral response and a low T cell response in various categories of patients with tuberculosis. J. Infect. Dis. 190, 1237–1244.
[13] Skeiky, Y.A., Ovendale, P.J., Jen, S., Alderson, M.R., Dillon, D.C., Smith, S., Wilson, C.B., Orme, I.M., Reed, S.G. and Campos-Neto, A. (2000) T cell expression cloning of a *Mycobacterium tuberculosis* gene encoding a protective antigen associated with the early control of infection. J. Immunol. 165, 7140–7149.
[14] Choudhary, R.K., Mukhopadhyay, S., Chakhaiyar, P., Sharma, N., Murthy, K.J.R., Katoch, V.M. and Hasnain, S.E. (2003) PPE antigen Rv2430c of *Mycobacterium tuberculosis* induces a strong B cell response. Infect. Immun. 71, 6338–6343.
[15] Li, Y., Miltner, E., Wu, M., Petrofsky, M. and Bermudez, L.E. (2005) *Mycobacterium avium* PPE gene is associated with the ability of the bacterium to grow in macrophages and virulence in mice. Cell Microbiol. 7, 539–548.
[16] 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.
[17] Choudhary, R.K., Pullakhandam, R., Ehtesham, N.Z. and Hasnain, S.E. (2004) Expression and characterization of Rv2430c, a novel immunodominant antigen of *Mycobacterium tuberculosis*. Protein Expres. Purif. 36, 249–253.
[18] Pym, A.S., Brodin, P., Majlessi, L., Brosch, R., Demangel, C., Williams, A., Griffiths, K.E., Marchal, G., Leclerc, C. and Cole, S.T. (2003) Recombinant BCG exporting ESAT-6 confers enhanced protection against tuberculosis. Nat. Med. 9, 533–539.
[19] Overbeek, R.M., Fonstein, S., D'Souza, M., Pusch, G.D. and Maltsev, N. (1999) The use of gene cluster to infer functional coupling. Proc. Natl. Acad. Sci. USA 96, 2896–2901.
[20] Dandekar, T., Snel, B., Huynen, M. and Bork, P. (1998) Conservation of gene order: a fingerprinting of proteins that physically interact. Trends Biochem. Sci. 23, 324–328.
[21] Dillon, D.C., Alderson, M.R., Day, C.H., Bement, T., Campos-Neto, A., Skeiky, Y.A., Vedvick, T., Badaro, R., Reed, S.G. and Houghton, R.L. (2000) Molecular and immunological characterization of *Mycobacterium tuberculosis* CFP-10, an immunodiagnostic antigen missing in *Mycobacterium bovis* BCG. J. Clin. Microbiol. 38, 3285–3290.
[22] Berthet, F.X., Rasmussen, P.B., Rosenkrands, I., Andersen, P. and Gicquel, B. (1998) A *Mycobacterium tuberculosis* operon encoding ESAT-6 and a novel low-molecular mass culture filtrate protein (CFP-10). Microbiology 144, 3195–3203.
[23] Limei, M.O. and Peter, A. (1998) Protein–protein interactions of proteins from the ESAT-6 Family of *Mycobacterium tuberculosis*. J. Bacteriol. 186, 2487–2491.
[24] Brodin, P., de Jonge, M.I., Majlessi, L., Leclerc, C., Nilges, M., Cole, S.T. and Brosch, R. (2005) Functional analysis of esat-6, the dominant T-cell antigen of *mycobacterium tuberculosis*, reveals key residues involved in secretion, complex-formation, virulence and immunogenicity. J. Biol. Chem. 280, 33953–33959.
[25] Finlay, B.B. and Falkow, S. (1997) Common themes in microbial pathogenicity revisited. Microbiol. Mol. Biol. Rev. 61, 136–169.
[26] Gey Van Pittius, N.C., Gamieldien, J., Hide, W., Brown, G.D., Siezen, R.J. and Beyers, A.D. (2001) The ESAT-6 gene cluster of *Mycobacterium tuberculosis* and other high G + C gram-positive bacteria. Genome Biol. 2, research 0044.1–0044.18.
[27] Pallen, M.J. (2002) The ESAT-6/WXG100 super family and a new gram positive secretion system? Trends Microbiol. 10, 209–212.
[28] Tekaia, F., Gordon, S.V., Garnier, T., Brosch, R., Barrell, B.G. and Cole, S.T. (1999) Analysis of the proteome of *Mycobacterium tuberculosis* in silico. Tuber. Lung Dis. 79, 329–342.

Infectious Agents and Cancer

BioMed Central

Review

Open Access

The co-evolved *Helicobacter pylori* and gastric cancer: trinity of bacterial virulence, host susceptibility and lifestyle

Yusuf Akhter[1], Irshad Ahmed[2], S Manjulata Devi[1] and Niyaz Ahmed*[1]

Address: [1]Pathogen Evolution Group, Laboratory of Molecular and Cellular Biology, Centre for DNA Fingerprinting and Diagnostics (CDFD), Hyderabad, India and [2]Department of Microbiology, Shri Shivaji College of Arts, Commerce and Science, Akola, India

Email: Yusuf Akhter - yusuf.akhter@gmail.com; Irshad Ahmed - sirfirshadahmed@gmail.com; S Manjulata Devi - manju@cdfd.org.in; Niyaz Ahmed* - niyaz@cdfd.org.in

* Corresponding author

Published: 04 January 2007

Infectious Agents and Cancer 2007, **2**:2 doi:10.1186/1750-9378-2-2

Received: 29 November 2006
Accepted: 04 January 2007

This article is available from: http://www.infectagentscancer.com/content/2/1/2

Abstract

Helicobacter pylori is an important yet unproven etiological agent of gastric cancer. *H. pylori* infection is more prevalent in developing Asian countries like India and it is usually acquired at an early age. It has been two decades since Marshall and Warren (1984) first described curved bacilli in the stomach of ulcer and gastritis patients. This discovery has won them the Nobel Prize recently, but the debate whether *H. pylori* is a pathogen or a commensal organism is still hot. Associations with disease-specific factors remain illusive years after the genome sequences were made available. Cytotoxin-associated antigen A (Cag*A*) and the so-called plasticity region cluster genes are implicated in pathogenesis of the carcinoma of stomach. Another virulence factor VacA whose role is still debatable, has recently been projected in pathology of gastric cancer. Studies of the evolution through genetic variation in *H. pylori* populations have provided a window into the history of human population migrations and a possible co-evolution of this pathogen with its human host. Possible symbiotic relationships were seriously debated since the discovery of this pathogen. The debate has been further intensified as some studies proposed *H. pylori* infection to be beneficial in some humans. In this commentary, we attempt to briefly discuss about *H. pylori* as a human pathogen, and some of the important issues linked to its pathophysiology in different hosts.

'We dance around in a ring and suppose, the secret sits in the middle and knows' – Robert Frost

Background

Barry J. Marshall and Robin Warren, two Australian researchers who discovered the bacterium *Helicobacter pylori* in 1982 have been awarded Nobel Prize of 2005 in Physiology or Medicine. This 'old fashioned medical detective work' impressed the Nobel Assembly of the Karolinska Institute, to move away from basic research [1,2] and to reward the research that proposes a much controversial bacterial organism as a dangerous pathogen. It was a long-standing dogma in the medical science that stress and lifestyle factors lead to gastritis and peptic ulcer disease. Warren and Marshal rebutted that dogma and made it clear that curved bacilli called *Campylobacter pyloridis* (later named as *Helicobacter pylori*) were the main cause of peptic ulceration, distal gastric adenocarcinoma, and gastric lymphoma [3]. Soon after this, *H. pylori* colonization model became one of the best-studied examples of pathogen evolution and its role in infection biology. This marked also the beginning of a contest on how long *H. pylori* had been colonizing human stomach, promoting

the analogy of a symbiotic organism coevolved with its human host.

H. pylori *as a marker of human peopling and migration: example of co-evolution*

H. pylori is presumably co-evolved with its host and therefore, origins and expansion of multiple populations and sub populations of *H. pylori* mirror ancient human migrations. Ancient origins of *H. pylori* in the world and in India are not clear and debatable. It is not clear how different waves of human migrations in different continents contributed to the evolution of strain diversity of *H. pylori*. Our group has recently attempted to address these issues through mapping genetic origins of *H. pylori* of native Peruvians (of Amerindian ancestry) [4] and Indians (Devi *et al.*, unpublished data) and their genomic comparison with hundreds of isolates from different geographic regions. For this purpose, genetic identity of strains was dissected by fluorescent amplified fragment length polymorphism (FAFLP) analysis, multilocus sequence typing (MLST) of the housekeeping genes and the sequence analyses of the *bab*B adhesin and *oip*A genes. The whole *cag* pathogenicity-island (*cag*PAI) from these strains was also analyzed using PCR and gene sequencing. In case of South American *H. pylori* populations, it was observed that while European genotype (*hp*-Europe) predominated in native Peruvian strains, approximately 20% of these strains represented a sub-population with an Amerindian ancestry (*hsp-Amerind*). However, all of these strains were shown to harbor a complete, 'western' type *cag*PAI, irrespective of their ancestral affiliation and the motifs surrounding it. This indicated a possible acquisition of *cag*PAI by the *hsp-Amerind* strains from the European strains, during decades of co-colonization. These observations, therefore, suggested presence of ancestral *H. pylori* (*hsp-Amerind*) in Peruvian Amerindians, which possibly managed to survive and compete against the Spanish strains that arrived to the New World about 500 years ago. It was suggested that this might have happened after native Peruvian *H. pylori* strains acquired *cag*PAI sequences, either by new acquisition in *cag*-negative strains or by recombination in *cag* positive Amerindian strains. In case of Indian strains, almost all the isolates analyzed revealed a European ancestry and belonged to MLST genogroup hp-Europe. The *cag*PAI harbored by Indian strains also revealed European features upon PCR based analysis and whole PAI sequencing. These observations suggest that *H. pylori* in India have ancient origins in Europe (Devi *et al*, unpublished data). These results are expected to strengthen speculations related to large-scale replacement of the ancient indigenous people of India by Indo-Aryan nomads, bringing first Neolithic practices and languages from the Fertile Crescent.

H. pylori *in gastric diseases*

H. pylori causes peptic ulceration, gastric adenocarcinoma, and gastric lymphoma. Gastric adenocarcinoma is the second highest cause of cancer deaths worldwide mainly due to high incidence, aggressive disease course, and lack of effective treatment options leading to a death toll of one million per annum worldwide [3]. *H. pylori* is implicated in distal gastric adenocarcinoma, which is more common than the proximal one. *H. pylori* also causes B cell mucosa-associated lymphoid tissue (MALT) lymphoma of the stomach [3] but at the same time negatively associated with more severe forms of reflux esophagitis and its sequelae – Barrett's esophagus and esophageal adenocarcinoma [5,6]. This negative correlation is the main reason that makes *H. pylori* a lesser evil. There has been a recent interest to look if *H. pylori* causes or facilitates human diseases of the gut other than the upper gastrointestinal tract or syndromes like idiopathic thrombocytopenic purpura [5,6], skin diseases, liver diseases, and cardiovascular and cerebrovascular disease. But many of these have been associated more commonly with Helicobacters other than *H. pylori* [7,8].

Bacterial encoded proinflammatory and carcinogenic factors

Studies reveal that the risk for developing gastric carcinoma was much greater with the *H. pylori* infection [9]. The *cag*A gene of *H. pylori* is the main virulence factor that leads to the development of gastric adenocarcinoma through derangement of cellular architecture and signaling. Presence of a functional *cag*A gene determines the *H. pylori* strain type to be aggressive or mild. The *cag*A-positive strains cause much intense ulceration of stomach or duodenum and are more damaging than the *cag*A-negative ones [10] leading to atrophic gastritis and gastric carcinoma [11,12]. CagA, the effector protein product of *cag*A, is tyrosine phosphorylated by SRC kinases after its secretion on the intestinal mucosal surface [13]. EPIYA motifs in the CagA protein sequence play a critical role in tyrosine phosphorylation, which in turn activates a SHP2 phosphatase to act as a oncoprotein. As SHP2 helps in cell growth and motility, its deregulation by CagA is an important oncogenic mechanism encoded by *H. pylori*. CagA based on sequence variation at the SHP2 binding site, is sub-classified into two main epidemiological types – East-Asian CagA (with stronger SHP2 binding and greater biological activity) and Western CagA (diminished SHP2 binding and milder ulcerative potentials). Strains with multiple CagA tyrosine phosphorylation motifs are more commonly associated with gastric cancer than those with fewer C type motifs [14-16].

Incidence of infection with *H. pylori* carrying biologically more active CagA might explain the high occurrence of gastric carcinoma in some countries such as Japan and

Korea. However, other populations with extremely high infection rates, such as Indians have almost negligible incidence of gastric carcinoma [17]. Possible reasons for such strange differences of disease outcome might be explained in the light of differences in genetic susceptibility among host populations, environmental factors such as dietary habits, and strain differences of *H. pylori*.

H. pylori has a single copy of the *vac*A gene encoding VacA protein, a secreted 95 kDa peptide. The *vac*A gene varies in the signal sequence (alleles s1a, s1b, s1c, s2) and/or its middle region (alleles m1, m2) among different *H. pylori* populations. The different allotypes of s and m regions determine the extent of cytotoxicity of VacA. Strains with *vac*A genotype s1/m1 are more commonly associated with gastric cancer than the other types [18]. Among other functions, VacA has been shown to induce apoptosis in epithelial cells. Recently, VacA has been proposed to be a potent immunomodulatory toxin, targeting the adapted immune system to suppress local immune responses to prolong the outcome of infection and thus prevent clearance by the host immune system [19]. The VacA has been the subject of intense biochemistry but lacked solid evidences that it is indeed involved in pathogenesis. A recent study argues that VacA has a miniscule role as a virulence factor during cell evasion by *H. pylori*. They showed that the *vac*A null mutant of *H. pylori* was able to evade specific cell lines, as did its wild type [20]. Therefore, the VacA involvement is still part of a debate on its being a true virulence factor and awaits further investigation.

Apart from the cardinal virulence factors CagA and VacA, several other proteins of the *cag*PAI, outer membrane envelope proteins, flagellins, adhesins, neutrophil activating protein (NAP), porins, LPS, urease and some members of the so called plasticity region cluster possibly playing an important role in inflammatory processes.

Microevolution during colonization: can it be linked to virulence optimization?

It has long been assumed that i) the *H. pylori* virulence factors are stable characteristic amid an otherwise fast evolving and recombining genome and ii) that these factors can be linked to disease progression or outcome, at any time. However, several reports present data against these assumptions. Two subclones of a *H. pylori* strain co-colonized a single patient with variations in *vac*A mid region, rendering one of the two sub-clones non-toxic [21]. The reason for this was clearly the microevolution *via* recombination within the stomach. Our group has previously shown a large deletion in *vac*A gene occurring in one of the two isolates of a common progenitor strain in a French patient, obtained 9 years apart [22]. This was most probably a case of adaptation or evolution *in vivo*. Duplication or deletion of the *cag*A gene has been shown by Aras *et al.*, [23] in two isolates existing in one individual and recovered 7 years apart. Kersulyte *et al*, have shown complete deletion of *cag*PAI through recombination [24]. In addition, various genotyping methods applied to two or more *H. pylori* isolates obtained from the same patient revealed similar fingerprints, with minor differences [25,26]. This may be possible due to the fact that two or more isolates recovered from a patient may share an ancestral relationship with a founder strain but have undergone independent genomic alterations. This phenomenon has been termed as 'microevolution' [25,27]. However, sequence evidence is necessary to confirm the location and extent of microevolution and phenotypic confirmation [16] is required to ascertain if such microevolution leads to alteration or optimization of virulence in response to change in the gastric environment.

Host genetic factors in H. pylori *induced carcinoma*

Host factors also play an important role in predisposition to *H. pylori* induced diseases and susceptibilities towards severe pathological outcomes. The host factors relevant in *H. pylori* induced diseases mainly include components of gastric secretion system and the immune apparatus. Interestingly, the gastritis and ulcer disease that result from bacterial infection, have distinct clinical profiles and are inversely associated with a high degree of acid secretion, whereas, gastric cancers are associated with low acid secretion due to loss of parietal cell mass [28,29]. In a recent study involving an East Indian population, authors suggested an association between the IL1β gene polymorphisms and *H. pylori*-mediated duodenal ulcer risk. They further observed effects of specific IL1β genotypes on the expression of IL1β mRNA in the gastric mucosa. Their *in vivo* studies were further substantiated, for the first time, by *in vitro* experiments, which represent the opposite homozygous risk genotypes that were observed in duodenal ulcer patients [30]. So this might explain the fact that differences in carcinogenesis risk in people from different geographical areas might reflect differences in their genetic make up.

The developing country enigma: Indians, diet and predisposition to gastric cancer?

What is enigmatic about the gastric cancer scenario in India? The answer is not simple. This country has a high prevalence of *H. pylori* infections and a low risk of gastric cancer in contrast to some of the developed countries with a low *H. pylori* colonization rate like China and Japan. India is known for a very high incidence of *H. pylori* infection [31,32]. Biologically inactive CagA could be a contributory factor in low prevalence of gastric ulcers and cancer in India. However, phenotyping studies based on *in vitro* assessment of CagA function in Indian isolates have not been done. In our opinion it will be inappropriate to implicate CagA functionality alone. The spectrum

and outcome of pathology in *H. pylori* infection is intricately governed by all the three factors – virulence, host genetics and the environment. It appears that the environment of stomach (acidity, buffering and mucus content) governed by lifestyle factors (diet, food habits, alcoholism, oral hygiene, water hygiene, personal hygiene, proximity with farming communities and animals) and the genetic determinants of susceptibility are chief drivers of the pathological outcome. Although poverty-associated factors (overcrowding, poor sanitation, lower socioeconomic status, compromised water hygiene etc.) in countries like India facilitate high frequency of *H. pylori* colonization, rapid re-colonization post eradication and lower age of acquisition [33]; a surprising fact is that such areas are at lowest risk of developing gastric cancer [34]. Correlation between *H. pylori* infection and gastric cancer has so far been unsuccessful in India [35]. A recent study from India involving 279 patients with gastric neoplasms failed to show a higher prevalence of *H. pylori* infection in patients with gastric neoplasms as compared to the controls (101 non-ulcer dyspepsia and 355 healthy subjects) [36]. These observations challenge the versatility of simplified models of gastric carcinogenesis based on *H. pylori* infection. We believe that in Indian context, diet as a major environmental factor governs the dynamics of gastric cancer demography mainly by regulating physiological integrity of gastric mucosal niches. And that is where; the dietary practices and lifestyle factors become important in the context of progression from gastritis to gastric cancer. Diets low in vegetables, fibers and fruits and high in salt-preserved foods or salt-processed meat increase the risk of stomach cancer [37].

Accordingly, in such situations there seems to be a difference in the distribution frequency of gastric cancer incidence. The southern and eastern parts of India have higher frequency of gastric cancer than rest of the country [38]. Rice is the staple food in south, whereas fish, meat, spices and salts are the main food items in eastern part [37-39]. Contrastingly, the large vegetarian population in northern India is at lower risk of gastric cancer. But the times are changing; rapid flourish of post globalization corporate culture brought fast foods, germ-free bottled water, pasteurized milk and preserved meat items to the present day lifestyle in big Indian cities. However, it will be too early to link it with rising gastric cancer incidences in cities in India [39].

Conversely, low to negligible incidence of gastric cancer as recorded for rural areas in India by the national cancer registry [39] leads us to speculate why rural communities have distinct advantages in terms of less damages from *H. pylori* infection. It needs to be investigated if these advantages are due to their diet based on fresh farm produce and their 'friendship' with the so-called "old friends", the group of bacteria that might be maintaining levels of regulatory immune cell populations and have been intricately associated during most of the mammalian evolution.

Competing interests

The author(s) declare that they have no competing interests.

Acknowledgements

Authors would like to thank Prof. Seyed E. Hasnain for his guidance and for discussions.

YA is recipient of Junior Research Fellowship from Council of Scientific & Industrial Research (CSIR), Govt. of India. Research in the laboratory of NA was supported by grants from the Department of Biotechnology, Govt. of India.

References

1. Dunn BE, Cohen H, Blaser MJ: **Helicobacter pylori.** *Clin Microbiol Rev* 1997, **10:**720-741.
2. Colding H, Hartzen SH, Roshanisefat H, Andersen LP, Krogfelt KA: **Molecular methods for typing of *Helicobacter pylori* and their applications.** *FEMS Immunol* 1999, **24:**193-199.
3. Atherton JC: **The pathogenesis of *Helicobacter pylori* -induced gastro-duodenal diseases.** *Annual Reviews in Pathology* 2006, **1:**63-96.
4. Devi SM, Ahmed I, Khan AA, Rahman SA, Alvi A, Sechi LA, Ahmed N: **Genomes of *Helicobacter pylori* from native Peruvians suggest admixture of ancestral and modern lineages and reveal a western type *cag*-pathogenicity island.** *BMC Genomics* 2006, **7:**191.
5. Franchini M, Veneri D: ***Helicobacter pylori* infection and immune thrombocytopenic purpura: an update.** *Helicobacter* 2004, **9:**342-346.
6. Jackson S, Beck PL, Pineo GF, Poon MC: ***Helicobacter pylori* eradication: novel therapy for immune thrombocytopenic purpura? A review of the literature.** *Am J Hematol* 2005, **78:**142-150.
7. Pellicano R, Fagoonee S, Rizzetto M, Ponzetto A: ***Helicobacter pylori* and coronary heart disease: Which directions for future studies?** *Crit Rev Microbiol* 2003, **29:**351-359.
8. Gasbarrini A, Carloni E, Gasbarrini G, Chisholm SA: ***Helicobacter pylori* and extragastric diseases – other *Helicobacters*.** *Helicobacter* 2004, **9:**57-66.
9. Uemura N, Okamoto S, Yamamoto S, Matsumura N, Yamaguchi S, Yamakido M, Taniyama K, Sasaki N, Schlemper RJ: ***Helicobacter pylori* infection and the development of gastric cancer.** *N Engl J Med* 2001, **345:**784-789.
10. Kuipers EJ, Perez-Perez GI, Meuwissen SG, Blaser MJ: ***Helicobacter pylori* and atrophic gastritis: importance of the *cagA* status.** *J Natl Cancer Inst* 1995, **87:**1777-1780.
11. Blaser MJ, Perez-Perez GI, Kleanthous H, Cover TL, Peek RM, Chyou PH, Stemmermann GN, Nomura A: **Infection with *Helicobacter pylori* strains possessing *cagA* is associated with an increased risk of developing adenocarcinoma of the stomach.** *Cancer Res* 1995, **55:**2111-2115.
12. Parsonnet J, Friedman GD, Orentreich N, Vogelman H: **Risk for gastric cancer in people with CagA positive or CagA negative *Helicobacter pylori* infection.** *Gut* 1997, **40:**297-301.
13. Hatakeyama M: **Oncogenic mechanisms of *Helicobacter pylori cag*A protein.** *Nature Rev Cancer* 2004, **4:**688-694.
14. Yamaoka Y, Kodama T, Kashima K, Graham DY, Sepulveda AR: **Variants of the 3' region of the *cag*A gene in *Helicobacter pylori* isolates from patients with different *H. pylori*-associated diseases.** *J Clin Microbiol* 1998, **36:**2258-2263.
15. Azuma T, Yamakawa A, Yamazaki S, Ohtani M, Ito Y, Muramatsu A, Suto H, Yamazaki Y, Keida Y, Higashi H, Hatakeyama M: **Distinct diversity of the *cag* pathogenicity island among *Helicobacter pylori* strains in Japan.** *J Clin Microbiol* 2004, **42:**2508-2517.

16. Argent RH, Kidd M, Owen RJ, Thomas RJ, Limb MC, Atherton JC: **Determinants and consequences of different levels of CagA phosphorylation for clinical isolates of *Helicobacter pylori*.** *Gastroenterology* 2004, **127:**514-523.
17. Sunny L, Yeole BB, Hakama M, Shiri R, Mathews S, Falah Hassani K, Advani SH: **Decreasing trend in the incidence of stomach cancer in Mumbai, India, during 1988 to 1999.** *Asian Pac J Cancer Prev* 2004, **5:**169-174.
18. Figueiredo C, Quint W, Nouhan N, van den Munckhof H, Herbrink P, Scherpenisse J, de Boer W, Schneeberger P, Perez-Perez G, Blaser MJ, van Doorn LJ: **Assessment of *Helicobacter pylori vac*A and *cag*A genotypes and host serological response.** *J Clin Microbiol* 2001, **39:**1339-1344.
19. Gebert B, Fischer W, Weiss E, Hoffmann R, Haas R: ***Helicobacter pylori* Vacuolating Cytotoxin Inhibits T Lymphocyte Activation.** *Science* 2003, **301:**1099-1102.
20. Oliveira MJ, Costa AC, Costa AM, Henriques L, Suriano G, Atherton JC, Machado JC, Carneiro F, Seruca R, Mareel M, Leroy A, Figueiredo C: ***Helicobacter pylori* Induces Gastric Epithelial Cell Invasion in a c-Met and Type IV Secretion System-dependent Manner.** *J Biol Chem* 2006, **281:**34888-34896.
21. Aviles-Jimenez F, Letley DP, Gonzalez-Valencia G, Salama N, Torres J, Atherton JC: **Evolution of the *Helicobacter pylori* vacuolating cytotoxin in a human stomach.** *J Bacteriol* 2004, **186:**5182-5185.
22. Prouzet-Mauleon V, Hussain MA, Lamouliatte H, Kauser F, Megraud F, Ahmed N: **Pathogen evolution *in vivo* : genome dynamics of two isolates obtained 9 years apart from a duodenal ulcer patient infected with a single *Helicobacter pylori* strain.** *J Clin Microbiol* 2005, **43:**4237-4241.
23. Aras RA, Fischer W, Perez-Perez GI, Crosatti M, Ando T, Haas R, Blaser MJ: **Plasticity of repetitive DNA sequences within a bacterial (Type IV) secretion system component.** *J Exp Med* 2003, **198:**1349-1360.
24. Kersulyte D, Chalkauskas H, Berg DE: **Emergence of recombinant strains of *Helicobacter pylori* during human infection.** *Mol Microbiol* 1999, **31:**31-43.
25. Marshall BJ: **The future of *Helicobacter pylori* eradication: a personal perspective.** *Aliment Pharmacol Ther* 1997, **1:**109-115.
26. Kuipers OP, Buist G, Kok J: **Current strategies for improving food bacteria.** *Res Microbiol* 2000, **151:**815-822.
27. Carroll IM, Ahmed N, Beesley SM, Khan AA, Ghousunnissa S, Morain CA, Habibullah CM, Smyth CJ: **Microevolution between paired antral and paired antrum and corpus *Helicobacter pylori* isolates recovered from individual patients.** *J Med Microbiol* 2004, **53:**669-677.
28. El-Omar EM, Oien K, El-Nujumi A, Gillen D, Wirz A, Dahill S, Williams C, Ardill JE, McColl KE: ***Helicobacter pylori* infection and chronic gastric acid hyposecretion.** *Gastroenterology* 1997, **113:**15-24.
29. Hansson LE, Nyren O, Hsing AW, Bergstrom R, Josefsson S, Chow WH, Fraumeni JF Jr, Adami HO: **The risk of stomach cancer in patients with gastric or duodenal ulcer disease.** *N Engl J Med* 1996, **335:**242-249.
30. Chakravorty M, Ghosh A, Choudhury A, Santra A, Hembrum J, Roychoudhury S: **Interaction between IL1B gene promoter polymorphisms in determining susceptibility to *Helicobacter pylori* associated duodenal ulcer.** *Hum Mutat* 2006, **27:**411-419.
31. Abasiyanik MF, Tunc M, Salih BA: **Enzyme immunoassay and immunoblotting analysis of *Helicobacter pylori* infection in Turkish asymptomatic subjects.** *Diagn Microbiol Infect Dis* 2004, **50:**173-177.
32. Ahmed N: **23 years of the discovery of *Helicobacter pylori* : is the debate over?** *Ann Clin Microbiol Antimicrob* 2005, **4:**17.
33. Singh K, Ghoshal UC: **Causal role of *Helicobacter pylori* infection in gastric cancer: an Asian enigma.** *World J Gastroenterol* 2006, **12:**1346-1351.
34. Miwa H, Go MF, Sato N: ***H pylori* and gastric cancer: the Asian enigma.** *Am J Gastroenterol* 2002, **97:**1106-1112.
35. Khanna AK, Seth P, Nath G, Dixit VK, Kumar M: **Correlation of *Helicobacter pylori* and gastric carcinoma.** *J Postgrad Med* 2002, **48:**27-28.
36. Ghoshal UC, Guha D, Bandyopadhyay S, Pal C, Chakraborty S, Ghoshal U, Ghosh TK, Pal BB, Banerjee PK: **Gastric adenocarcinoma of MALT lymphoma with successful anti-*H pylori* therapy and in a patient re-infected with *H pylori* after regression gastric resection: a case report.** *BMC Gastroenterol* 2002, **2:**6.
37. Gajalakshmi CK, Shanta V: **Lifestyle and risk of stomach cancer: a hospital-based case-control study.** *Int J Epidemiol* 1996, **25:**1146-1153.
38. Mathew A, Gangadharan P, Varghese C, Nair MK: **Diet and stomach cancer: a case-control study in South India.** *Eur J Cancer Prev* 2000, **9:**89-97.
39. **National Cancer Registry Programme, India (Indian Council of Medical Research)** [http://www.canceratlasindia.org/about.htm]

BMC Genomics

BioMed Central

Research article

Open Access

Ancestral European roots of *Helicobacter pylori* in India

S Manjulata Devi[†1], Irshad Ahmed[†2,3], Paolo Francalacci[4], M Abid Hussain[1], Yusuf Akhter[1], Ayesha Alvi[1], Leonardo A Sechi[5,6], Francis Mégraud[5,7] and Niyaz Ahmed[*1,5]

Address: [1]Pathogen Evolution Group, Centre for DNA Fingerprinting and Diagnostics, Hyderabad, India, [2]Centre for Liver Research and Diagnostics, Deccan College of Medical Sciences and allied Hospitals, Hyderabad, India, [3]Department of Microbiology, Shri Shivaji College of Arts, Commerce and Science (SGB Amravati University), Akola, MS, India, [4]Dipartimento di Zoologia e Genetica Evoluzionistica, University of Sassari, Sassari, Italy, [5]ISOGEM Collaborative Network on Genetics of Helicobacters (The International Society for Genomic and Evolutionary Microbiology, University of Sassari, Sassari, Italy), [6]Dipartimento de Scienze Biomediche, University of Sassari, Sassari, Italy and [7]INSERM U853 and Centre National de Référence des Campylobacters et Hélicobacters, Laboratoire de Bactériologie, Université Victor Segalen Bordeaux 2, France

Email: S Manjulata Devi - manju@cdfd.org.in; Irshad Ahmed - sirfirshadahmed@gmail.com; Paolo Francalacci - pfrancalacci@uniss.it; M Abid Hussain - abid@cdfd.org.in; Yusuf Akhter - yusuf.akhter@gmail.com; Ayesha Alvi - ayesha@cdfd.org.in; Leonardo A Sechi - sechila@uniss.it; Francis Mégraud - Francis.Megraud@chu-bordeaux.fr; Niyaz Ahmed* - niyaz.cdfd@gmail.com

* Corresponding author †Equal contributors

Published: 20 June 2007

BMC Genomics 2007, **8**:184 doi:10.1186/1471-2164-8-184

Received: 5 January 2007
Accepted: 20 June 2007

This article is available from: http://www.biomedcentral.com/1471-2164/8/184

Abstract

Background: The human gastric pathogen *Helicobacter pylori* is co-evolved with its host and therefore, origins and expansion of multiple populations and sub populations of *H. pylori* mirror ancient human migrations. Ancestral origins of *H. pylori* in the vast Indian subcontinent are debatable. It is not clear how different waves of human migrations in South Asia shaped the population structure of *H. pylori*. We tried to address these issues through mapping genetic origins of present day *H. pylori* in India and their genomic comparison with hundreds of isolates from different geographic regions.

Results: We attempted to dissect genetic identity of strains by multilocus sequence typing (MLST) of the 7 housekeeping genes (*atp*A, *efp*, *ure*I, *ppa*, *mut*Y, *trp*C, *yph*C) and phylogeographic analysis of haplotypes using MEGA and NETWORK software while incorporating DNA sequences and genotyping data of whole *cag* pathogenicity-islands (*cag*PAI). The distribution of *cag*PAI genes within these strains was analyzed by using PCR and the geographic type of *cag*A phosphorylation motif EPIYA was determined by gene sequencing. All the isolates analyzed revealed European ancestry and belonged to *H. pylori* sub-population, hpEurope. The *cag*PAI harbored by Indian strains revealed European features upon PCR based analysis and whole PAI sequencing.

Conclusion: These observations suggest that *H. pylori* strains in India share ancestral origins with their European counterparts. Further, non-existence of other sub-populations such as hpAfrica and hpEastAsia, at least in our collection of isolates, suggest that the hpEurope strains enjoyed a special fitness advantage in Indian stomachs to out-compete any endogenous strains. These results also might support hypotheses related to gene flow in India through Indo-Aryans and arrival of Neolithic practices and languages from the Fertile Crescent.

Background

Analysis of genetic diversity in microorganisms normally reflects patterns of their own evolution although it is very rare that this can portray their hosts' evolution. Co-evolution between host and pathogens can be explained only if pathogens are not horizontally transmitted, and this supports a possible phylogenetic and evolutionary parallel of the host and pathogens. Sadly, in many cases frequent horizontal transmission separates the evolution of the bacterium from that of the host. However, for some pathogens, such as *H. pylori* [1-3], and JC viruses [4], transmission is faithfully restricted to families within specific communities. This phenomenon has in recent times provided evidence regarding patterns of human migration [2,4,5] in different continents.

The human gastric pathogen *H. pylori* is presumed to have co-evolved with its host [6] and established itself in the human stomach possibly millions of years ago [7]. It has been recognized recently as a reliable biological marker of host-pathogen co-evolution and ancient human migration based on sequence variation in select gene loci. *H. pylori* are genetically diverse to the extreme, providing about 1,400 informative sites within 3.5 to 4.5 kb of sequence from housekeeping genes, and their global genetic structure based on such sequence-haplotypes parallels that of humans [2]. Moreover, epidemiological studies have shown that transmission occurs predominantly within families [8-11]. *H. pylori* therefore, could provide a window into human origins and migration [1,3] and the impact of religions and social systems on stratification of human ethnic groups [12].

A landmark study based on PCR based DNA motif analysis proposed that *H. pylori* jumped recently from animals to humans and, therefore, the acquisition of *H. pylori* by humans may be a recent phenomenon [13]. This study has been the basis for the idea of '*H. pylori* free New World' [13]. However, several independent studies based on large-scale analyses of candidate gene polymorphisms contrasted the idea of recent acquisition and suggest that *H. pylori* might have co-evolved with humans [1,6,14].

Using the same set of Peruvian isolates described earlier by Kersulyte *et al.* [13], Devi *et al*. [3], from our group have suggested that the genetic make up of south American isolates could be an admixture of ancestral and modern lineages of *H. pylori*. They clearly highlighted presence of ancestral *H. pylori* in Peruvians that possibly survived influxes of Spanish strains from Iberian expansions in Peru about 500 years ago. Also, according to this study, the survival advantage of indigenous strains was possibly due to the acquisition of western type *cag*PAIs from newly arrived Spanish strains.

Previous genotyping studies on Indian isolates have largely targeted molecular epidemiological issues. However, Wirth *et al.* [12], for the first time, using *H. pylori* genotypes, addressed issues such as impact of two different religions and societal systems on stratification of human ethnic groups [12] in the remotest north eastern Ladakh area of India. In view of intriguing ideas on ancient origin of *H. pylori*, and the fact that ancient origins and arrival of *H. pylori* are hardly known in the context of the vast South Asian continent, additional evidences based on strains from different geographical regions of Asia are clearly needed.

In this study, we attempted to unravel population genetic structure and gene pool diversity of Indian isolates of *H. pylori* from culturally and linguistically diverse ethnic Indians. The main objective behinds the study has been to explore genetic features of the strains that might explain their ancestral origin and might help reconstruct different waves of pre-historic human migration in India. We also looked if it is possible to link some of the native strains to their ancestors in West Asia, Eurasia or Europe.

Results

DNA isolates, diagnostic PCR and epidemiological genotyping

DNA quality and purity was confirmed by agarose gel electrophoresis and diagnostic PCRs revealed presence of *cag*A, *ice*A, *vac*A, *glm*M, *bab*B and *oip*A genes in all the Indian isolates we tested. The molecular epidemiological features of all the 63 strains we analyzed have been elaborated in Figure 1. Our isolates were quite diverse with respect to the plasticity region ORFs that we analyzed and no specific signature was seen dominant as regards to the arrangement or rearrangement of these ORFs. This validated that all the isolates that we looked at were in fact independent and did not represent any derivatives of clonal evolution.

Specific primers amplifying different alleles (see methods section) were used to analyze the *vac*A allelic diversity. The sizes of the amplified products for *vacA* s1 and *vacA* s2 were 259 bp and 286 bp respectively. Of the 63 isolates analyzed, the s1 allele was detected in 33 (52.3%) and the s2 allele type was detected in 11 (17.4%) strains. The m1 variant was detected in 34 (53.9%) and the m2 variant in 37 (58.7%). The highly toxigenic *vac*A allele combination s1m1 was found to be dominant (33.3%) as compared to other *vac*A allele types. The *vac*A genotype s1m2 was detected in 9 isolates (14.2%) whereas *vacA* s2m1 and *vacA* s2m2 genotypes were detected in 4 isolates (6.3%) each. Not all the isolates yielded full *vac*A amplicons, as regions of *vac*A gene, in particular, the signal region posed difficulty in amplification. This is a very common phenomenon observed in *H. pylori* owing to frequent recom-

Strains ID	Disease type	Religion	Phylogenetic placement	*vacA* allele	*cag*-RJ	Plasticity region ORFs							
						986	947	912	926	944	931	945	933
MS1	DU	Muslim	hpEurope	s2m2	-								
MS2	DU	Hindu	hpEurope	s2m2	-								
MS3	DU	Muslim	hpEurope	s1m1b	IIIa								
MS4	DU	Muslim	hpEurope	s1m1b	IIIa								
MS5	DU	Hindu	hpEurope	s1m1b	IIIa								
MS6	DU	Hindu	hpEurope	s2m2	IIIb								
MS7	NUD	Hindu	hpEurope	s1m1a	IIIa								
MS8	DU	Muslim	hpEurope	s1m1a	IIIb								
MS9	GU	Muslim	hpEurope	s1m2	IIIa								
MS10	PU	Muslim	hpEurope	s1m2	IIIb								
MS11	DU	Muslim	hpEurope	s1m2	-								
MS12	NUD	Muslim	hpEurope	s1m2	IIIb								
MS13	GC	Muslim	hpEurope	s1	-								
MS14	DU	Muslim	hpEurope	s1	IIIa								
MS15	GU	Muslim	hpEurope	s1m2	IIIa								
MS16	DU	Hindu	hpEurope	nd	IIIb								
MS17	DU	Muslim	hpEurope	s1m1a	IIIb								
MS18	DU	Hindu	hpEurope	s2m1b	IIIa								
MS19	unknown	Muslim	hpEurope	nd	-								
MS21	DU	unknown	hpEurope	s1m1a	-								
MS22	DU	Hindu	hpEurope	-	IIIa								
MS24	GU	Muslim	hpEurope	-	IIIa								
MS26	GC	Muslim	hpEurope	s1m1a	IIIb								
MS27	NS	Muslim	hpEurope	m1a/m1b	-								
MS28	DU	unknown	hpEurope	s1	IIIb								
MS33	DU	Hindu	hpEurope	s1m1a/m1b	IIIb								
MS34	GU	Muslim	hpEurope	s1m1a	IIIb								
MS37	unknown	Muslim	hpEurope	s1m2	IIIa								
MS38	unknown	Hindu	hpEurope	s1m1b/ m2	-								
MS40	unknown	unknown	hpEurope	s1m2	IIIa								
3C	DU	Hindu	hpEurope	s1m1b/m2	-								
3K	PUD	Muslim	hpEurope	s1m1a/m2	IIIa								
3S	DU	Hindu	hpEurope	s1m1b/m2	IIIb								
3E	DU	Muslim	hpEurope	s1m1b/m2	-								
3G	DU	Muslim	hpEurope	s1m2	IIIb								
4J	CDU	Hindu	hpEurope	s1m1b/m2	IIIb								
4L	DU	Hindu	hpEurope	s1m1b/m2	-								
4K	DU	unknown	hpEurope	s1m1b/m2	IIIa								
4R	DU	Hindu	hpEurope	s2m2	-								
L8	G	Buddhist	hpAsia2	s2	IIIa								
L22	G	Buddhist	hpAsia2	m2	III								
L36	G	Muslim	hpAsia2	s2m1a	III								
L44	G	Muslim	hpEurope	m2	IIIa								
L45	G	Muslim	hpAsia2	s2m1b	III								
L60	G	Muslim	hpEurope	m2	III								
L67	G	Muslim	hpAsia2	s2m1a	IIIa								
L79	G	Buddhist	hpAsia2	m1b	IIIa								
L133	G	Buddhist	hpAsia2	s2	-								
L172	G	Buddhist	hpAsia2	s2	-								
NI1133	DU	Hindu	hpEurope	m2	III b								
NI5	GU	Hindu	hpEurope	s1m1b/ m2	III b								
NI6	DU	Hindu	hpEurope	s1m1b/ m2	III b								
NI7	DU	Hindu	hpEurope	s1m2	III a								
NI8	DU	Hindu	hpEurope	s1m1b/ m2	III a								
AA1	DU	Hindu	hpEurope	m1b/ m2	III b								
AA2	DU	Hindu	hpEurope	m1b/ m2	III a								
NI01	GC	Hindu	hpEurope	m1b/ m2	-								
NI02	DU	Hindu	hpEurope	m1b/ m2	III a								
AA5	GC	Hindu	hpEurope	m1b/ m2	III a								
NI03	PHT	Hindu	hpEurope	m2	I								
AA7	DU	Hindu	hpEurope	m1b/ m2	III a								
AA8	NUD	Hindu	hpEurope	m1b/ m2	III a								
NI04	NUD	Hindu	hpEurope	m2	-								

Figure 1
Detailed characteristics of Indian *H. pylori* isolates used in the study. [Yellow, region amplified or present; Blue, region absent or rearranged; -, region failed to amplify].

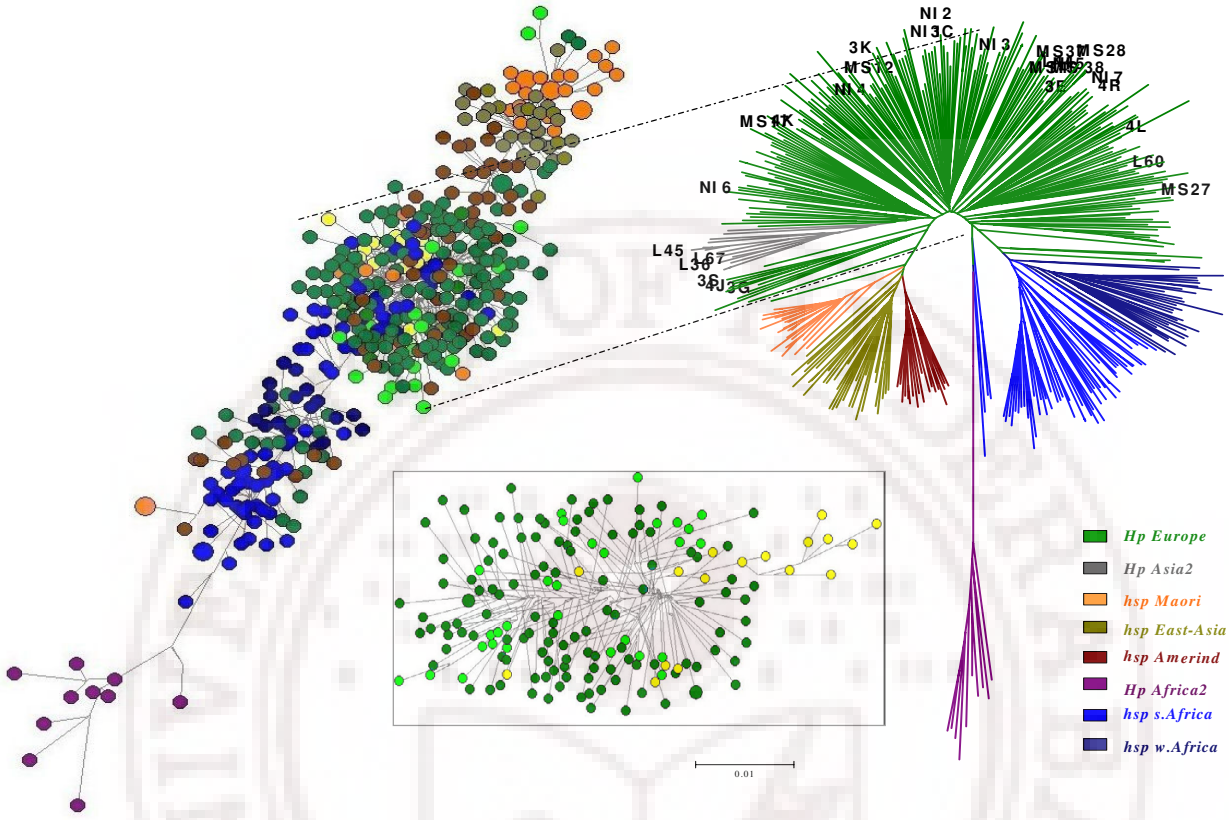


Figure 2
Neighbor joining tree (Kimura 2-parameter) (right) showing the global population structure of *H. pylori* wherein Indian isolates are highlighted. The phylogenetic tree was based on a total of 23 sequence records of South and North Indian isolates while incorporating ~400 other sequence records from pubMLST database representing different *H. pylori* populations and sub populations in the world. The population genetic structure was investigated by determining the multilocus haplotypes based on concatenated sequences of seven unlinked housekeeping genes that are scattered around the *H. pylori* chromosome. Individual isolates were assigned to bacterial populations called hpEastAsia (sub-populations: hspEAsia, hspMaori, hspAmerind), hpEurope, hpAfrica1 (hspSAfrica, hspWAfrica), hpAsia2 and hpAfrica2 [11]. Representatives from each of these (sub)-populations were chosen for subsequent analysis of the *cag*PAI. Isolates from the population hpAfrica2 do not contain *cag*PAI. Phylogenetic relationships were also estimated through NETWORK analysis (left) based on 665 mutating positions that revealed the co-evolution of the *H. pylori* genome. The Ladakhi (yellow) and other Indian (light green) lineages were more clearly discerned within the European (dark green) cluster (centre box), when analyses based on the remaining 650 mutating positions were performed. For the Neighbor-joining tree (right), the bootstrap values of the interior branches as calculated in MEGA, were significantly high to indicate the correct topology of the branches within the clades.

bination. The *vacA* alleles have been shown to differ in frequency and type among East Asian isolates [15], for instance, s1c is the predominant signal sequence allele among East Asian isolates [16]. Typically, the *vac*A s1c was found to be completely absent in the Indian isolates.

Multilocus sequence analysis

We report that almost all of the *H. pylori* strains from India share significant homology to the members of sub-population hpEurope. A total of 33 MLST profiles based on DNA sequence of a concatenated multigene comprising of 7 individual gene loci (*atpA, efp, mutY, ppa, trpC, ureI* and *yphC*) were generated from Indian isolates. Data comprising of these MLST profiles were subjected to comparative genomic analysis with ~400 other *H. pylori* sequences from different geographical and ethnic groups [11]. Such analyses upon construction of a neighbor-joining tree in MEGA 3.1 software using Kimura-2 parameter revealed clear geographic distribution of various *H. pylori* populations and sub-populations, essentially in accordance with the previous results [1,3,17]. All the Indian isolates from North and South India and 2 of them from Ladakh clustered under hpEurope. Seventeen Ladakhi isolates clustered tightly to form a separate branch, hpAsia2. Results of MLST analysis in MEGA3.1 were successfully reproduced using NETWORK based phylogeny, which revealed simi-

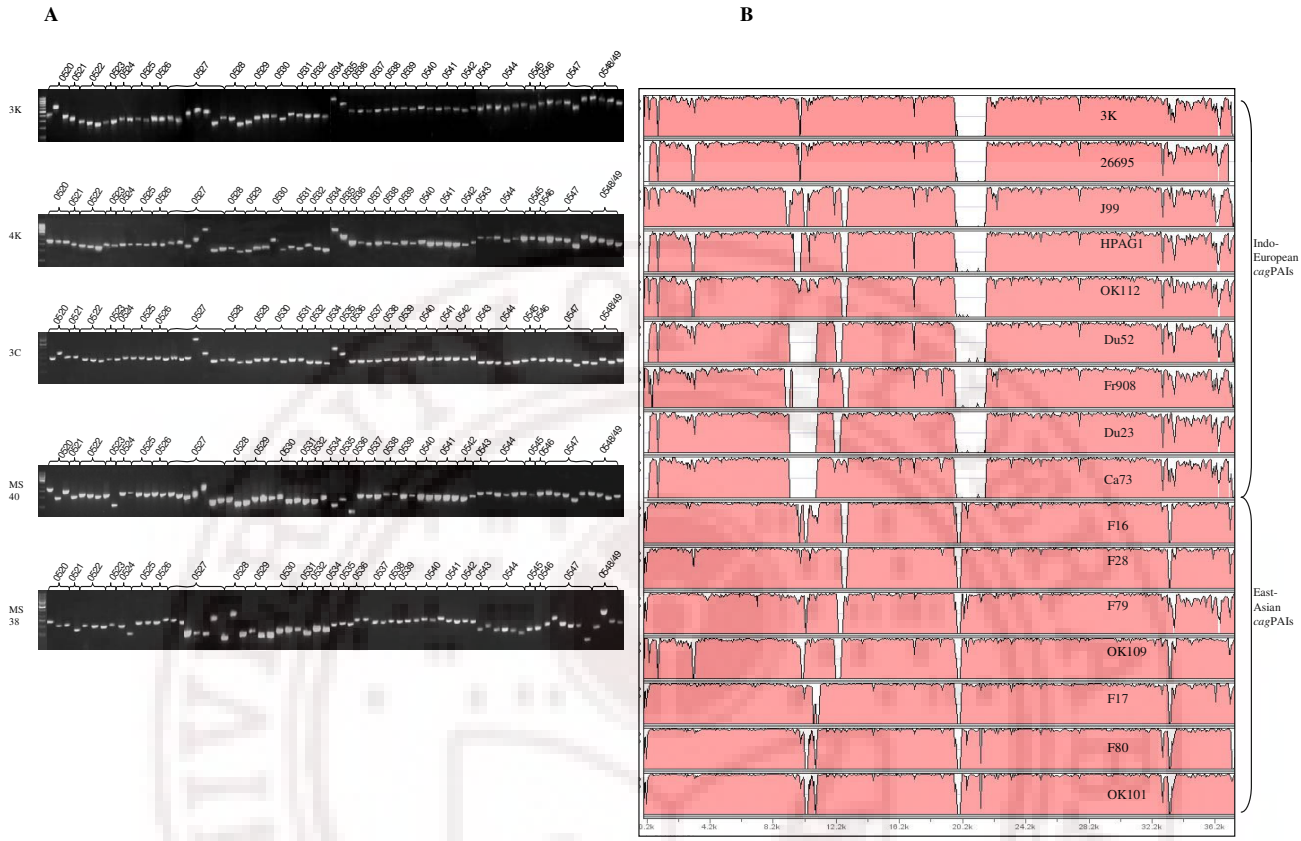


Figure 3
Comparative genomic analysis of the *cag*PAIs from Indian isolates. A) PCR based analysis of the complete *cag*PAI of 5 representative hpEurope Isolates: 3K, 4K, 3C, MS40 and MS38 from India. Overlapping PCR primers amplified the whole *cag*PAI indicating the intactness of the PAI in these isolates. B) Global pair-wise alignments of whole *cag*PAI sequences of different *H. pylori* isolates were generated by VISTA using default parameters [47]. The OK129 genome was taken as the base sequence (not shown) and rest of the sequences were aligned against it. The X-axis denotes length of the sequence under consideration and the Y-axis conveys homology in % with the base genome sequence). The Indian hpEurope isolate, 3K was aligned with other whole *cag*PAI sequences from GenBank along with the *cag*-PAIs of HP 26695, HPJ99 and HPAG1. The accession numbers for the public domain sequences of the *cag*PAIs from Europe [9] and Japan [49] that we used in our analyses, were as follows – Ca73 (AY330638 and AY330639), Du23:2 (AY330643 and AY330644), Du52:2 (AY330640, AY330641 and AY330642), F80 (AB120421), OK112 (AB120425), F16 (AB120416), F17 (AB120417), F28 (AB120418), F79 (AB120420), OK101 (AB120422), OK109 (AB120424). Sequence of the French isolate, Fr 908 was determined in this study (EF195721). While the *cag*PAI sequence of the Indian isolate 3K (hpEurope) was found to be genetically highly similar to and aligning closely with the 26695 sequence, it also revealed significant sequence similarities with other isolates of European origins (that harbor Western type of *cag* EPIYA sequences) such as HPAG1, OK112, Du52, Du23, Ca73, J99 and Fr908. It was however largely unrelated to the East Asian like isolates (mainly harboring Asian type *cag* EPIYA sequences) such as F16, F28, F79, OK109, F17, OK101 and F80.

lar acquaintances for *H. pylori* in India. Mirroring the spread of human populations from Africa, our network analysis suggests the co-evolution of *H. pylori* with *Homo sapiens*, as also suggested recently [6]. Both the domains of the Network tree based on 650 (data not shown) and 665 (Figure 2, left) mutating positions clearly separated African from non-African sequences. The second domain seemed to harbor higher phylogenetic information, since the resulting graph is more clearly structured, with a more accurate separation among European, Amerindian, Asian and Australasian lineages. The Indian *H. pylori* sequences were clustered within the European portion of the network, wherein the first domain identifies a separate branch, encompassing the majority of the Ladakhi samples, as a distinct sub-population of hpAsia2 within the European variability, and remarking the isolation of the human host population. However, many of the Ladakhi Muslim samples clustered in hpEurope and revealed a significant sequence similarity to the mainland Indian samples. These results are in agreement with previous studies on the hypervariable region of human mitochondrial DNA that showed the common origin of European and

Indian populations [18] and the relative homogeneity of Indian populations regardless of their ethnic and linguistic affiliation [19].

Analysis of the cagPAI and its Right Junction (RJ) motifs

Overlapping primer amplification to span entire *cag*PAI worked reproducibly with our isolates; Figure 3(A) reveals complete PCR output for the ~38 kb *cag*PAI region in 5 representative strains MS38, MS40, 3K, 4K and 3C. All the constituent genes of the PAI were successfully amplified for all the Indian isolates studied. To get more insights into composition and arrangement of the gene loci within the PAI, complete sequencing of the *cag*PAI of isolate 3K was performed. This isolate was from a patient with peptic ulcer disease (PUD) from South India. The size of complete *cag*PAI of this isolate was 36,876 bp with a G+C content of 35.9. The sequence composition and gene order in the *cag*PAI of 3K was compared to those of the three completely sequenced strains 26695, J99 and HPAG1 which revealed some minor differences such as fused HP0521 and HP0522 genes due to the deletion of a single nucleotide at the 3' end of HP0521. Similarly single or dinucleotide differences were observed in the *cagX* (HP0528), *cagN* (HP0538) and *cagE* (HP0544) and most of these insertions and deletions were observed in the intergenic regions. Broadly, the *cag*PAI genes were very conserved as regards to the amino acid sequences when compared with at least 15 different publicly available *cag*PAI sequences.

cag-RJ (the extreme right junction of the *cag*PAI, between 3' end of the *cag*A gene and the start of the glutamate racemase – *glr*) was studied for our 63 isolates where 99% isolates harbored type III motif. A total of 47 of 63 strains (75%) gave positive PCR results for *cag*-RJ (Figure 1). The type III motif was found in 27 of 39 South Indian isolates and 20 of 24 North Indian isolates. It is noteworthy that *cag*-RJ typeIII motifs are genetically close to European type I motifs probably due to an ancient insertion event, followed by recombinational scrambling among type I and III lineages [13]. We did not find in our Indian isolates any type II motifs, which constitute a signature characteristic of East Asian gene pool.

Genetic relationship of Indian isolates based on cagA and whole cagPAI sequences

A full-length *cag*PAI sequence based alignment was constructed using the Indo-European type 3K and Afro-European type Fr908 (French patient isolate) sequences determined in this study, along with 15 different whole *cag*PAI sequence from GenBank: Ca73, Du23: 2, Du52: 2, F16, F17, F28, F79, F80, OK101, OK109, OK112, OK129, 26695, J99 and HPAG1. Our South Indian isolate, 3K, was found to be aligning with the Western *cag*PAI sequences (Figure 3B).

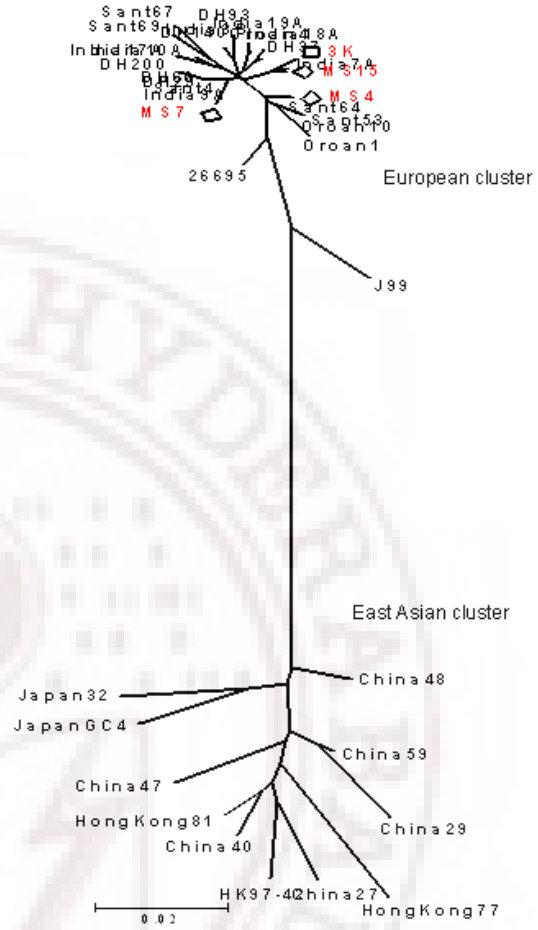


Figure 4

Phylogenetic tree based on the 5' end sequence of the *cag*A (an informative 219 bp segment of *cag*A was used to align sequences from unrelated isolates) suggests possible common origins for isolates from ethnic Indians and the tribal. Representative Indian genotypes (3K, MS4, Ms7 and MS15) based on this 219 bp sequence clustered tightly with previously determined genotypes of strains obtained from ethnic Bengalis [India3B (AF202219), India7A (AF202220), India9A (AF202221), India10A (AF202222), India17A (AF202223), India18A (AF202224), India19A (AF202225), DH140 (AY169293), DH200 (AY169294), DH29 (AY169295), DH37 (AY169296), DH60 (AY169297), DH93 (AY169298)] and Santhal and Oraon tribals [Sant4 (AY162446), Sant53 (AY162447), Sant64 (AY162448), Sant67 (AY162449), Sant69 (AY162450), Oraon1 (AY162451), Oraon10 (AY162452), Oraon4 (AY162453)] [20]. All the East Asian strains [China27 (AJ252979), China29 (AJ252980), China40 (AJ252982), China48 (AJ252983), China47 (AJ252985), China59 (AJ252986), Hongkong77 (AF198485), Hongkong81 (AF198486), Hongkong97-42 (AF239733), Japan GC4 (AF198484), Japan32 (AJ239726)], however, clustered together and formed a separate cluster.

We examined relatedness of the *cag*A gene sequences of tribal isolates from India to the mainstream Indian isolates and the European isolates by analyzing a 219 bp informative fragment near the 5' end of *cag*A which usually distinguishes the European and the East Asian strains [20]. Comparative sequence analysis was used to construct phylogenetic relationship in MEGA3.1. All the sequence records corresponding to the isolates of Santhal and Oraon tribals revealed homologies to the main stream Indian strains from Hyderabad, Lucknow and Bengal and also to all the representative European strains. These tribal isolates did not cluster with East Asian strains (Figure 4).

This makes it clear that the *cag*PAI of Indian strains is a completely evolved one and probably was acquired from a European source, well before the arrival of *H. pylori* in India. This is also evident from the fact that the Indian strains, though of a European descent, do not share characteristic features of Asian *cag*PAIs.

Discussion

Although the Indian peninsula has seen many different waves of population migration [21], the Paleolithic archaeological evidence is not clearly visible to understand peopling of this country [22]. Nonetheless, the Indus Valley and Harappan civilizations portray footprints of Neolithic period [23] suggestive of the arrival of Indo-European speakers who established the caste system, an anthropologically significant prehistoric event [24,25]. The cultural and historical importance of the arrival and settlement of the Indo-Aryans is undisputed, but it is not clear if this was established through 'replacement of the existing people by outsiders' [22] or did the 'people already in India changed their habits and cultures?' [22]. Such questions have never been addressed in an unambiguous manner, even though the potential of polymorphic DNA markers in reconstruction of human migration and phylogeography [26,27] has long been appreciated. It appears that even carefully planned geographic genomics studies remained largely speculative due to the lack of a universal 'gold standard' as the classical mitochondrial DNA markers offer too few informative polymorphisms and the newly developed Y – chromosome markers are even less polymorphic than mitochondrial hypervariable regions [2]. Lately, new genetic models were successfully harnessed based on parasites and pathogens that probably accompanied their human host during evolution and much of the human history including migrations and expansions [2,4,5] in different continents. Such approaches constitute an attractive alternative to reconstruct human origins and spreads, population dynamics and bottlenecks, wars and displacements, farming and plagues etc.

Our study was aimed at tracking ancient origins of the Indian *H. pylori* through a two-pronged approach to i) substantiate European link of the pathogen in India and ii) to prove that the pathogenicity island was also of European origin and this PAI has not been a 'recent' addition to the genome of Indian *H. pylori*. Our analyses, based on MLST and comprehensive genotyping of the *cag*PAI, linked about 100% of the Indian isolates to *H. pylori* subpopulation hpEurope. This perhaps conveys the message that *H. pylori* was most probably introduced to the Indian subcontinent by ancient Indo-European nomadic people and our findings, therefore, are consistent with the idea of a possible gene flow into India with the arrival of Indo-Aryans.

Overall, based on the MLST data (Figure 2) and the *cag*PAI patterns (Figure 3), we suggest that *H. pylori* might have arrived in India probably at the same time when Indo-European language speaking people crossed into India (~4000–10,000 years before present). Alternatively, the unquestionable common origin of Indian strains with the European ones could be actually more ancient, following the upper Paleolithic spread of *Homo sapiens* in Eurasia, as suggested by mtDNA variability [18], and our data on *H. pylori* MLST do not rule out this possibility.

Present day India represents a 'genetic playground' with tremendous diversity of cultures, and languages. However, the people are largely stratified as tribals and nontribals [25]. Four main language families are spoken, the largest being, Indo European (IE), which is prevalent in North, and the second largest Dravidian (DR) group represents languages spoken in the South [28]. The other two language groups include Tibeto-Burman (TB) of the Sino-Tibetan and the Austro-Asiatic (AA) families, largely spoken in far North and the North-east India. While most of the IE speakers belong to castes, the majority of the tribal communities (>450) speak about 750 different dialects that fit within any one of the other three language families (DR, TB, AA) [25,28]. Such an enormous cultural diversity might argue for many different populations and sub-populations of *H. pylori*. But until now, and including this study, *H. pylori* with genetic features of hpEurope have only been reported from India [29,30]. Even the newly described sub-population hpAsia2 from Ladakh is also a variant of hpEurope and many Ladakhi strains that we looked at in this study, clustered with European *H. pylori* clade (Figure 2). Also, the *cag*A sequences from *H. pylori* belonging to tribal Oraon and Santhals were indistinguishable from those of the mainstream Indians and Europeans (Figure 4), indicating sweeping spread of a single *H. pylori* genotype across the Indian peninsula. Moreover, we did not document presence of any other *H. pylori* populations and sub-populations such as hspAmerind, hspMaori, hpAfrica and hpEast Asia in the limited, but

representative culture collection that we looked at. However, the visible footprints of other migrations into India such as from the North Eastern corridor and the presence of phenotypic features resembling to Africans in the South, make it unwise to presume an '*H. pylori* free India' at the time of arrival of Indo-European speaking invaders. This issue and the fact that *H. pylori*'s first association with humans traces back to millions of years before present, in Africa [6,17], it is more realistic to hypothesize that *H. pylori* of African and Asian gene pool might have already been present in India. The predominance of a single *H. pylori* population might therefore, point to a distinct survival advantage conferred by a fully functional (western type) *cag*PAI. This analogy is consistent with the scenario we previously reported [3] for the South American, Amerindian strains, which were presumably out competed by their Spanish counterparts arriving with an intact and functional western *cag*PAI.

Finally, it is possible that phylogeny based on highly recombining gene loci [15,29,31-35] may not be completely foolproof to extract inheritance from different ancestral populations, especially when we use tools such as MEGA 3.0 [36], which do not support admixture analysis. Moreover, phylogenetic methods based on bifurcating trees, such as Neighbor joining analysis, may not be fully appropriate for analysis at the intra-species level [37,38], especially in case of hypervariable genomic regions, where multiple homoplasy due to reversions, recurrent mutations etc., or polytomy may sometimes confound the phylogenetic interpretation. However, the housekeeping genes used here are selectively neutral and uniform as compared to virulence associated loci such as the flagellins and *vac*A [10], and therefore, recombinant and hybrid alleles that blur lineage inferences, could be a rare occurrence and not a routine. Partly in view of this assumption and due to our previous experiences on dissecting complex ancestry of native Peruvian isolates using phylogenetic methods [3] we did not attempt admixture analysis with complicated Bayesian statistics. However, to ensure that our conclusions did not represent shortcomings of a single method, we adopted an integrated phylogenetic approach combining MEGA/NETWORK based analyses and genotyping strategy based on full *cag*PAI and its left and right end sequences. Interestingly, these approaches unambiguously show the Indian *H. pylori* genotypes scattered among the European ones. Although this would be consistent with gene flow into India with the Indo-Aryans, or even more ancient origins following the Paleolithic expansion of humans in Eurasia, but also consistent with another scenario: migration from India to Europe. However, the later scenario becomes insignificant due to the unavailability of supporting archeological, linguistic and historical data. Nonetheless, an understanding of the time-scale would be helpful for choosing between such explanations, with the estimation of divergence times between the *H. pylori* sequences in the different human populations. These issues therefore need to be addressed in future.

Conclusion

In summary, we found significant overlap among genetic identities of Indian and European *H. pylori* based on core and flexible genome markers. This remarkable genetic similarity points to their possible common genetic origins and could therefore be potentially useful in understanding entry, survival, spread and adaptation of *H. pylori* in Indian stomachs. Also, this study is consistent with the hypothesis of co-evolution of *H. pylori* with *H. sapiens* and therefore, could form a reliable foundation to test and reconstruct gene flow into India with the arrival of Indo-Aryans or otherwise.

Methods

Bacterial strains, genomic DNA and diagnostic PCR

All the strains were cultured by the Centre for Liver Research and Diagnostics, Deccan college of Medical Sciences, Hyderabad, from patient biopsies. All the biopsy material was collected with necessary ethical clearances and after obtaining informed consents. Template DNA was prepared from single colony picks as described previously [39]. Genomic DNAs of the 10 Ladakhi strains were received from Mark Achtman, Max-Planck Institute für Infektionsbiologie, Berlin, Germany. Genomic DNA was isolated from strains obtained from patients with different disease types including Duodenal Ulcer (DU); Gastric Ulcer (GU); Gastric Cancer (GC); Gastritis (G); Non Ulcer Dyspepsia (NUD); Peptic Ulcer Disease (PUD); Chronic Duodenal Ulcer (CDU); Portal Hyper Tension (PHT) etc. (Figure 1). However, in the current study, the clinical background of the individual isolates was not taken into account. The Indian isolates we looked at (n = 63) were originally from Native Indian people mainly of Aryan and Dravidian ancestry from India. PCR based analyses of genes namely *cag*A, *glm*M, *bab*B [14] and *oip*A were carried out to ascertain the quality of DNA samples we used. Also these PCR assays served as amplification level controls for the analysis of insertion, deletion and substitution in the *cag*PAI.

MLST analysis by MEGA 3.1 and NETWORK 4.2.0

A 600 bp region each from the 7 housekeeping genes spread throughout the genome *atp*A, *efp*, *ure*I, *ppa* and *mut*Y, *trp*C, *yph*C was amplified by PCR and sequenced for all the Indian isolates exactly as described previously [3]. Sequencing was performed on both the strands, using an ABI Prism 3100 DNA sequencer (Applied Biosystems, USA). PCR and direct sequencing were performed at least twice to determine and confirm the DNA sequences for each isolate. Consensus sequence for each of the samples

was generated using Genedoc (version 2.6.002). Multiple alignments of sequenced nucleotides were carried out using Clustal X (version 1.81). Neighbor joining trees were constructed in MEGA 3.0 [36] using bootstrapping at 10000 bootstrap trials and through Kimura-2 parameters. For beginning construction of phylogenetic trees based on MLST genotyping procedures, ~400 sequences of the 7 housekeeping genes of strains belonging to different established genotypes, including 40 sequences of isolates from Ladakh were obtained from the pubMLST database [40] (courtesy, Daniel Falush). The Indian *H. pylori* diversity represented in the final MEGA3.0 alignment and the tree thereof comprised of a total of 63 sequences inclusive of the 10 Ladakhi sequences generated in house along with the other 9 representative Ladakhi sequences from the database. We performed on MLST sequence data a network analysis using the program Network 4.2.0.0. [38,41]. In particular, the median-joining algorithm for multistate DNA data was used [42,43]. Because of a program limitation, which cannot handle more than 1000 polymorphic sites at once, we performed the analysis separately on two halves of the sequence (encompassing respectively 650 and 665 polymorphic sites). The input file (in *.rdf format) was obtained using the commercial software DNA Alignment 1.1.2.1.

Profiling of the cagA gene, the whole cagPAI and its right junction

The 5' end of the *cag*A gene was amplified using primers mentioned elsewhere [44] and the amplified products were sequenced with forward and reverse primers. The consensus sequences were then translated into amino acid sequences using GeneDoc software (version 2.6.002) and were then assigned to the Western or the East Asian group based on the C or D repeats present respectively in the EPIYA motif [45]. Genetic diversity of the *cag*A 5' end sequences for our Indian isolates: MS15, MS7, MS4 and 3K along with 26695 and J99 were compared to the other records from GenBank [20,30,46]. A phylogenetic neighbor-joining tree was constructed by MEGA 3.1 version using these sequences (Figure 4).

PCR analyses were carried out to find the status of the *cag*-PAI using 8 sets of primers that amplified the *cag*A gene, its promoter region, the *cag*E and *cag*T genes and the left end of the *cag*PAI [8,29,34]. We also analyzed whole *cag*-PAI of the representative isolates from India (3K, 4K, 3C, MS40 and MS38) by PCR using overlapping primers as described by Blomstergren and colleagues [9]. The entire *cag*PAI sequence of a single representative Indian isolate 3K was determined. The complete *cag*PAI sequence was aligned by VISTA programme [47] against other PAI sequences belonging to strains 26695, J99, HPAG1 and 13 other clinical isolates corresponding to *H. pylori* sub-populations hpEurope, hpEast Asia and hpAfrica1 (Figure 3B).

Chromosomal rearrangements are known to give rise to 5 types of insertion-deletion and substitution motifs in the region between the right end of *cag*A gene and the glutamate racemase (*glr*) gene (*cag*-RJ). We assessed these rearrangement profiles for all of the Indian isolates by PCR as described earlier by Kersulyte and colleagues [13].

Analysis of the chromosomal plasticity region cluster

Chromosomal plasticity region ORF's were assessed for all the 63 Indian isolates by PCR based typing to ensure that all the strains that we looked at were independent and non-clonal by descent. The PCR primers and the procedures used for evaluating the presence of the plasticity region ORF's (JHP912, HP986, JHP947, JHP926, JHP944, JHP931, JHP945 and JHP933) have been descried previously [48].

Nucleotide sequence accession numbers

The nucleotide sequences of the 7 housekeeping genes for the 23 representative Indian isolates have been deposited in the GenBank [Accession numbers, GenBank: DQ504165–DQ504183 and DQ927245–DQ927248 (*atpA*), DQ504184–DQ504202 and DQ927249–DQ927252 (*efp*), DQ504203–DQ504221 and DA927253–DA927256 (*mutY*), DQ504222–DQ504240 and DQ927257–DQ927260 (*ppa*), DQ504241–DQ504259 and DQ927261–DQ927264 (*trpC*), DQ504260–DQ504278 and DQ927265–DQ927268 (*ureI*), DQ504279–DQ504297 and DQ927269–DQ927272 (*yphC*)]. These sequences will also be made available through the pubMLST database maintained at the Max-Planck Institute für Infektionsbiologie, Berlin, Germany. The sequence of whole *cag*PAIs of the representative Indian isolate 3K and the French isolate Fr908 for which the sequence was determined in our laboratory, have been deposited in Genbank under accession nos. DQ985738 and EF195721 respectively. These and other sequences can also be requested from the authors.

Authors' contributions

SMD and IA performed and analyzed MLST, all other genotyping experiments and phylogenetic analysis. SMD also helped in analysis of *bab*B and *oip*A genotyping. MAA performed *vac*A genotyping. IA also performed *H. pylori* isolation and culture. YA carried out *in silico* analysis of the *cag*PAI sequences. PF performed Network analysis on MLST data and contributed to manuscript writing. LAS and FM provided expert clinical and epidemiological support and contributed to discussions and manuscript writing. NA planned and supervised the study, edited the final draft of the manuscript and provided overall leadership. All the authors read and approved the final manuscript.

Acknowledgements

We thank Director of the Centre for DNA Fingerprinting and Diagnostics (CDFD), Hyderabad for support and guidance. Our thanks are due to various collaborators in India and abroad, who contributed to our *H. pylori* DNA collections. We are grateful to Daniel Falush and Mark Achtman (pubMLST.org) for international MLST data and advice. We are grateful to Seyed E. Hasnain (University of Hyderabad) for his guidance and to Chris Tyler-Smith (Sanger Centre, UK) for his critical comments on our raw data. We are also thankful to the International Society for Genomic and Evolutionary Microbiology (ISOGEM) for supporting and endorsing the study. Financial support from the Department of Biotechnology, Government of India to NA (grant ref. *BT/PR2473/Med/13/106/2001*) is gratefully acknowledged. Help provided by our laboratory support staff, namely, Shaikh Zamir, B Krishnamurthy and Wasim Ahmad is thankfully appreciated. NA is the Corresponding Fellow of the European Helicobacter Study Group.

References

1. Falush D, Wirth T, Linz B, Pritchard JK, Stephens M, Kidd M, Blaser MJ, Graham DY, Vacher S, Perez-Perez GI, Yamaoka Y, Mégraud F, Otto K, Reichard U, Katzowitsch E, Wang X, Achtman M, Suerbaum S: **Traces of human migrations in *Helicobacter* pylori populations.** *Science* 2003, **299:**1582-1585.
2. Wirth T, Meyer A, Achtman M: **Deciphering host migrations and origins by means of their microbes.** *Mol Ecol* 2005, **14:**3289-3306.
3. Devi SM, Ahmed I, Khan AA, Rahman SA, Alvi A, Sechi LA, Ahmed N: **Genomes of *Helicobacter pylori* from native Peruvians suggest admixture of ancestral and modern lineages and reveal a western type *cag*-pathogenicity island.** *BMC Genomics* 2006, **7:**191.
4. Pavesi A: **Utility of JC polyomavirus in tracing the pattern of human migrations dating to prehistoric times.** *J Gen Virol* 2005, **86:**1315-1326.
5. Holmes EC: **The phylogeography of human viruses.** *Mol Ecol* 2004, **13:**745-756.
6. Linz B, Balloux F, Moodley Y, Manica A, Liu H, Roumagnac P, Falush D, Stamer C, Prugnolle F, van der Merwe SW, Yamaoka Y, Graham DY, Perez-Trallero E, Wadstrom T, Suerbaum S, Achtman M: **An African origin for the intimate association between humans and *Helicobacter pylori*.** *Nature* 2007, **445:**915-918.
7. Covacci A, Telford JL, Giudice GD, Parsonnet J, Rappuoli R: ***Helicobacter pylori* virulence and genetic geography.** *Science* 1999, **284:**1328-1333.
8. Ikenoue T, Maeda S, Gura KO, Akanuma M, Mitsuno Y, Imai Y, Yoshida H, Shiratori Y, Omata M: **Determination of *Helicobacter pylori* virulence by simple gene analysis of the *cag* pathogenicity island.** *Clin Diag Lab Immunol* 2001, **8:**181-186.
9. Blomstergren A, Lundin A, Nilsson C, Engstrand L, Lundeberg J: **Comparative analysis of the complete *cag* pathogenicity island sequence in four *Helicobacter pylori* isolates.** *Gene* 2004, **328:**85-93.
10. Achtman M, Azuma T, Berg DE, Ito Y, Morelli G, Pan ZJ, Suerbaum S, Thompson S, van der Ende A, van Doorn LJ: **Recombination and clonal groupings within *Helicobacter pylori* from different geographical regions.** *Mol Microbiol* 1999, **32:**459-470.
11. Falush D, Stephens M, Pritchard JK: **Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies.** *Genetics* 2003, **164:**1567-1587.
12. Wirth T, Wang X, Linz B, Novick RP, Lum JK, Blaser M, Morelli G, Falush D, Achtman M: **Distinguishing human ethnic groups by means of sequences from *Helicobacter pylori*: lessons from Ladakh.** *Proc Natl Acad Sci USA* 2004, **101:**4746-4751.
13. Kersulyte D, Mukhopadhyay AK, Velapatino B, Su WW, Pan ZJ, Garcia C, Hernandez V, Valdez Y, Mistry RS, Gilman RH, Yuan Y, Gao H, Alarcon T, Lopez-Brea M, Nair GB, Chowdhury A, Datta S, Shirai M, Nakazawa T, Ally R, Segal I, Wong BCY, Lam SK, Olfat F, Boren T, Engstrand L, Torres O, Schneider R, Thomas JE, Czinn S, Berg DE: **Differences in genotypes of *Helicobacter pylori* from different human populations.** *J Bacteriol* 2000, **182:**3210-3218.
14. Ghose C, Perez-Perez GI, Bello MGD, Pride DT, Bravi CM, Blaser MJ: **East Asian genotypes of *Helicobacter* pylori strains in Amerindians provide evidence for its ancient human carriage.** *Proc Natl Acad Sci USA* 2002, **99:**15107-15111.
15. Carroll IM, Ahmed N, Beesley SM, Khan AA, Ghousunnissa S, O'Morain CA, Smyth CJ: **Fine-structure molecular typing of Irish *Helicobacter pylori* isolates and their genetic relatedness to strains from four different continents.** *J Clin Microbiol* 2003, **41:**5755-5759.
16. Doorn VLJ, Figueiredo C, Mégraud F, Pena S, Midolo P, Queiroz DM, Carneiro F, Vanderborght B, Pegado MD, Sanna R, De Boer W, Schneeberger PM, Correa P, Nq EK, Atherton J, Blaser MJ, Quint WG: **Geographic distribution of *vacA* allelic types of *Helicobacter pylori*.** *Gastroenterology* 1999, **116:**823-830.
17. Gressmann H, Linz B, Ghai R, Pleissner KP, Schlapbach R, Yamaoka Y, Kraft C, Suerbaum S, Meyer TF, Achtman M: **Gain and loss of multiple genes during the evolution of Helicobacter pylori.** *PLoS Genet* 2005, **1(4):**e43.
18. Kivisild T, Bamshad MJ, Kaldma K, Metspalu M, Metspalu E, Reidla M, Laos S, Parik J, Watkins WS, Dixon ME, Papiha SS, Mastana SS, Mir MR, Ferak V, Villems R: **Deep common ancestry of Indian and western-Eurasian mitochondrial DNA lineages.** *Curr Biol* 1999, **9:**1331-1334.
19. Sharma S, Saha A, Rai E, Bhat A, Bamezai R: **Human mtDNA hypervariable regions, HVR I and II, hint at deep common maternal founder and subsequent maternal gene flow in Indian population groups.** *J Hum Genet* 2005, **50:**497-506.
20. Datta S, Chattopadhyay S, Nair GB, Mukhopadhyay AK, Hembram J, Berg DE, Saha DR, Khan A, Santra A, Bhattacharya SK, Chowdhury A: **Virulence genes and neutral DNA markers of *Helicobacter pylori* isolates from different ethnic communities of West Bengal, India.** *J Clin Microbiol* 2003, **41:**3737-3743.
21. Underhill PA, Jin L, Zemans R, Oefner PJ, Cavalli-Sforza LL: **A pre-Columbian Y chromosome-specific transition and its implications for human evolutionary history.** *Proc Natl Acad Sci USA* 1996, **93:**196-200.
22. Carvalho-Silva DR, Zerjal T, Tyler-Smith C: **Ancient Indian roots?** *J Biosci* 2006, **31:**1-2.
23. Kenoyer JM: **Ancient cities of the Indus valley civilization.** Karachi: Oxford University Press; 1998.
24. Bamshad M, Kivisild T, Watkins WS, Dixon ME, Ricker CE, Rao BB, Naidu JM, Prasad BV, Reddy PG, Rasanayagam A, Papiha SS, Villems R, Redd AJ, Hammer MF, Nguyen SV, Carroll ML, Batzer MA, Jorde LB: **Genetic evidence on the origins of Indian caste populations.** *Genome Res* 2001, **11:**994-1004.
25. Basu A, Mukherjee N, Roy S, Sengupta S, Banerjee S, Chakraborty M, Dey B, Roy M, Roy B, Bhattacharyya NP, Roychoudhury S, Majumder PP: **Ethnic India: a genomic view, with special reference to peopling and structure.** *Genome Res* 2003, **13:**2277-2290.
26. Cavalli-Sforza LL: **The DNA revolution in population genetics.** *TIG* 1998, **14:**60-65.
27. Cavalli-Sforza LL, Feldman MW: **The application of molecular genetic approaches to the study of human evolution.** *Nat Genet* 2003, **33(Suppl):**266-275.
28. Sahoo S, Singh A, Himabindu G, Banerjee J, Sitalaximi T, Gaikwad S, Trivedi R, Endicott P, Kivisild T, Metspalu M, Villems R, Kashyap VK: **A prehistory of Indian Y chromosomes: Evaluating demic diffusion scenarios.** *Proc Natl Acad Sci USA* 2005, **103:**843-848.
29. Kauser F, Khan AA, Hussain MA, Carroll IM, Ahmad N, Tiwari S, Shouche Y, Das B, Alam M, Ali SM, Habibullah CM, Sierra R, Megraud F, Sechi LA, Ahmed N: **The *cag* pathogenicity island of *Helicobacter pylori* is disrupted in the majority of patient isolates from different human populations.** *J Clin Microbiol* 2004, **42:**5302-5308.
30. Mukhopadhyay AK, Kersulyte D, Jeong J, Datta S, Ito Y, Chowdhury A, Chowdhury S, Santra A, Bhattacharya SK, Azuma T, Nair GB, Berg DE: **Distinctiveness of genotypes of *Helicobacter pylori* in Calcutta, India.** *J Bacteriol* 2000, **182:**3219-3227.
31. Ahmed N, Khan AA, Alvi A, Tiwari S, Jyothirmayee CS, Kauser F, Ali M, Habibullah CM: **Genomic analysis of *Helicobacter pylori* from Andhra Pradesh, south India: molecular evidence for three major genetic clusters.** *Curr Sci* 2003, **85:**101-108.
32. Carroll IM, Ahmed N, Beesley SM, Khan AA, Ghousunnissa S, O'Morain CA, Habibullah CM, Smyth CJ: **Microevolution between paired antral and paired antral and corpus *Helicobacter pylori* isolates recovered from individual patients.** *J Med Microbiol* 2004, **53:**669-677.

33. Kauser F, Hussain MA, Ahmed I, Ahmad N, Habeeb A, Khan AA, Ahmed N: **Comparing genomes of *Helicobacter pylori* strains from the high altitude desert of Ladakh, India.** *J Clin Microbiol* 2005, **43:**1538-1545.
34. Prouzet-Mauleon V, Hussain MA, Lamouliatte H, Kauser F, Megraud F, Ahmed N: **Pathogen evolution in vivo: genome dynamics of two isolates obtained nine years apart from a duodenal ulcer patient infected with a single *Helicobacter pylori strain*.** *J Clin Microbiol* 2005, **43:**4237-4241.
35. Ando T, Peek RM, Pride D, Levine SM, Takata T, Lee YC, Kusugami K, van der Ende A, Kuipers EJ, Kusters JG, Blaser MJ: **Polymorphisms of *Helicobacter pylori* HP0638 reflect geographic origin and correlate with *cagA* status.** *J Clin Microbiol* 2002, **40:**239-246.
36. Kumar S, Tamura K, Nei M: **Integrated software for molecular evolutionary genetics analysis and sequence alignment.** *Brief Bioinfor* 2004, **5:**150-163.
37. Herrnstadt C, Elson JL, Fahy E, Preston G, Turnbull DM, Anderson C, Ghosh SS, Olefsky JM, Beal MF, Davis RE, Howell N: **Reduced-median-network analysis of complete mitochondrial DNA coding-region sequences for the major African, Asian, and European haplogroups.** *Am J Hum Genet* 2002, **70:**1152-1171.
38. Posada D, Crandall KA: **Intraspecific gene genealogies: trees grafting into networks.** *Trends Ecol Evol* 2001, **16:**37-45.
39. Kauser F, Hussain MA, Ahmed I, Srinivas S, Devi SM, Majeed AA, Rao KR, Khan AA, Sechi LA, Ahmed N: **Comparative genomics of *Helicobacter pylori* isolates recovered from ulcer disease patients in England.** *BMC Microbiol* 2005, **5:**32.
40. **'pubMLST database'** [http://www.pubmlst.org]
41. **'Network package'** [http://www.fluxus-engineering.com]
42. Bandelt H-J, Forster P, Sykes BC, Richards MB: **Mitochondrial portraits of human populations.** *Genetics* 1995, **141:**743-753.
43. Bandelt H-J, Forster P, Röhl A: **Median-joining networks for inferring intraspecific phylogenies using median networks.** *Mol Biol Evol* 1999, **16:**37-48.
44. Yamaoka Y, Orito E, Mizokami M, Gutierrez O, Saitou N, Kodama T, Osato MS, Kim JG, Ramirez FC, Mahachai V, Graham DY: ***Helicobacter pylori* in north and south America before Columbus.** *FEBS Lett* 2002, **517:**180-184.
45. Hatakeyama M: **Oncogenic mechanisms of the Helicobacter pylori CagA protein.** *Nat Rev Cancer* 2004, **4:**688-694.
46. Rahman M, Mukhopadhyay AK, Nahar S, Datta S, Ahmad MM, Sarker S, Masud IM, Engstrand L, Albert MJ, Nair GB, Berg DE: **DNA-Level characterization of *Helicobacter pylori* strains from patients with overt disease and with benign infections in Bangladesh.** *J Clin Microbiol* 2003, **41:**2008-2014.
47. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I: **VISTA: computational tools for comparative genomics.** *Nucleic Acids Res* 2004, **32:**W273-W279.
48. Occhialini A, Marais A, Alm R, Akanuma M, Mitsuno Y, Imai Y, Yoshida H, Shiratori Y, Omata M: **Distribution of open reading frames of plasticity region of strain J99 in *Helicobacter pylori* strains isolated from gastric carcinoma and gastritis patients in Costa Rica.** *Infect Immun* 2000, **68:**6240-6249.
49. Azuma T, Yamakawa A, Yamazaki S, Ohtani M, Ito Y, Muramatsu A, Suto H, Yamazaki Y, Keida Y, Higashi H, Hatakeyama M: **Distinct diversity of the *cag* pathogenicity island among *Helicobacter pylori* strains in Japan.** *J Clin Microbiol* 2004, **42:**2508-2517.