In-silico prediction of regulons in bacterial genomes

Thesis Submitted for the Degree of

Doctor of Philosophy

To the Department of Biochemistry
School of Life Sciences, University of Hyderabad

By

Sailu Yellaboina
Centre for DNA Fingerprinting and Diagnostics
Hyderabad 500 076

2005

University of Hyderaad
School of Life Sciences
Department of Biochemistry
Hyderabad 500 046. India

Declaration

The research work embodied in this thesis entitled, "***In-silico*** **prediction of regulons in bacterial genomes**", has been carried out by me at the Centre for DNA Fingerprinting and Diagnostics, Hyderabad, under the guidance of Dr. 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.

Sailu Yellaboina

University of Hyderabad
School of Life Sciences
Department of Biochemistry
Hyderabad 500 046. India

Certificate

This is to certify that this thesis entitled, "***In-silico* prediction of regulons in bacterial genomes**", submitted by Mr. Sailu Yellaboina for the degree of Doctor of Philosophy to the 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 for any diploma or degree of any other university or institution.

Dr. Seyed E. Hasnain
Thesis supervisor
CDFD, Hyderabad

Prof. C. K. Mitra
Head, Department of Biochemistry
University of Hyderabad

Prof. A. S. Raghavendra
Dean, School of Life Sciences
University of Hyderabad

Contents

Acknowledgements

I wish to thank all the individuals and organizations that directly or indirectly assisted me in the completion of this research dissertation.

I would like to express my sincere gratitude to Dr. Seyed E. Hasnain, Director, CDFD, for his advice and whose prime motivation gave me never ending enthusiasm to work on this thesis work. I am also grateful for his encouragement to carry out the research work and to publish in peer reviewed journals. I also admire his generosity in sending students to various national and international conferences, which is quite encouragement for a young researcher like me.

I am grateful to Dr. Akash Ranjan, for key contributions to this thesis. His observations, comments and advice helped me to establish the overall direction of the research and to move forward. I also admire his help in making presentations and writing publications. I thank him for providing me resources to carry out the research work.

I would like to thank my former supervisor Prof. C.K Mitra, Head, Department of Biochemistry, University of Hyderabad, for introducing me bioinformatics, teaching computational skills and algorithms.

I would like to thank Prof. K. Subbarao, former Head, Department of Biochemistry, University of Hyderabad, for his generosity in sending me to CDFD to carry out the research work.

I would like to thank Dr. Shekher C. Mande for coordinating the PhD program and for his encouragement in carrying out the research work.

I am greatful to Mr. Bhoopal Reddy, one of my best friends for teaching me mathematics at the intermediate level and C/C^{++}, PERL programming during PhD. I am grateful to Dr. Prachee Prakash and Dr. Sarita Ranjan for carrying out the experimental validation of my predictions that resulted in cooperative joint publications.

I am greatful to Mr. Senthil Kumar for forcibly making me to shift from Windows to Linux and teaching me Linux, which has accelerated my research work. I am grateful to Jayasree Sheshadri whose expertise in computers helped me to get my first publication, which encouraged me in carrying out research work with great enthusiasm.

I thank my former colleague Abdul Wahid Ansari and one of my best friends Dr. Narasimha Rao for sending me the countless number of PDFs of various articles that helped me to gain in depth knowledge in my research work.

I thank my lab mates Mr. Vaibhav Vindal, Ms. Umadevi and former labmate, Mr. Rami Reddy for their help during my PhD work. I thank Dr. Rahul Siddhartan and Dr. Eric Van Neumann for providing me software tools that enabled me to carry out this work.

I gratefully acknowledge the financial support of several institutions, the Council of Scientific and Industrial Research (CSIR) for providing a research fellowship and SCIR (NMITLI) for funding the projects I worked for. I would like to acknowledge the institutes that provided me genome sequence data. Preliminary sequence data on *Mycobacterium smegmatis* was obtained from The Institute for Genomic Research through the website at http://www.tigr.org. Sequencing of *Mycobacterium smegmatis* was accomplished with support from National Institute of Allergy and Infectious Diseases (NIAID). The sequence data on *Mycobacterium marinum* were produced by the *Mycobacterium marinum* Sequencing Group at the Sanger Institute and has been obtained from ftp://ftp.sanger.ac.uk/pub/pathogens/mm/.

Lastly, and most importantly, I wish to thank my parents and brothers. They raised me, supported me, taught me and love me.

Dedication

This thesis is dedicated to my Intermediate as well as Bachelor of Science teacher, Ch.Suresh Reddy, whose excellent teaching of chemistry is my prime source of understanding the science.

Sailu Yellaboina

List of Abbreviations

ADP	Adenosine Diphosphate
ATP	Adenosine Triphosphate
B. subtilis	*Bacillus subtilis*
BBH	Bi-directional Best Hits
BLAST	Basic Local Alignment Search Tool
C. diptheriae	*Corynebacterium diphtheriae*
C. glutamicum	*Corynebacterium glutamicum*
CBS	Cystathionine Beta-synthase Domain
CDD	Conserved Domain Database
CDFD	Centre for DNA Fingerprinting and Diagnostics
COG	Cluster of Orthologous Group
DNA	Deoxyribonucleic Acid
dNTP	Deoxynucleotide Triphosphate
Dps	DNA - binding protein from starved cells
DTT	1, 4-Dithiothreitol
DtxR	Diphtheria toxin Repressor
E. coli	*Escherichia coli*
EMSA	Electrophoretic Mobility Shift Assay
GeSTer	Genome Scanner for Terminators
HTML	HyperText Markup Language
ICF	Index of Cluster Formation
IdeR	Iron dependent Regulator
IPTG	Isopropyl-b-D-Thiogalactopyranoside
IUPAC	International Union of Pure and Applied Chemistry
M. avium	*Mycobacterium avium sub sp. paratuberculosis*
M. bovis	*Mycobacterium bovis*
M. leprae	*Mycobacterium leprae*
M. marinum	*Mycobacterium marinum*

M. smegmatis	*Mycobacterium smegmatis*
M. tuberculosis	*Mycobacterium tuberculosis*
MEME	Multiple Expectation Maximization for Motif Elicitation
Mramp	Mycobacterium natural-resistance-associated macrophage protein
N. farcinica	*Nocardia farcinica*
NAD	Nicotinamide Adenine Dinucleotide
NCBI	National Centre for Biotechnology Information
Nramp	Natural resistance macrophage associated protein
ORF	Open Reading Frame
PAGE	Polyacrylamide Gel Electrophoresis
PBBH	The gene Pairs of Bi-directional Best Hits
PCBBH	The gene Pairs of Close Bi-directional Best Hits
PCR	Polymerase Chain Reaction
PERL	Practical Extraction and Reporting Language
PROCSE	Probabilistic Clusering of Sequences
RBS	Ribosome Binding Site
RNA	Ribonucleic Acid
RPS-BLAST	Reversed Position Specific - Basic Local Alignment Search Tool
RPS-BLAST	Reversed Position Specific - Basic Local Alignment Search Tool
SDS	Sodium Dodecyl Sulfate
TIGR	The Institute for Genomic Research

Chapter 1

Introduction

Cells having identical genomic content may exhibit differences in their metabolism and physiology. Such differences arise due to differential gene expression. Differential gene expression was first discovered in *Escherichia coli*, where an operon model of gene expression was proposed. Subsequent studies have demonstrated that these operons are ubiquitous in many bacteria.

1.1 Operon organization

An operon is a series of genes that are transcribed together as a single mRNA. The operon consists of transcription initiation signal (promoter), transcription regulatory sequence (operator) and a transcription termination signal (Jacob and Monod, 1998).

1.1.1 Structural genes and regulator genes

There are two classes of genes in operons: structural genes and regulator genes. The structural genes code protein and RNA molecules that are required for coordinated enzymatic or structural functions in the cell. The regulator genes code for proteins that bind to operator sequences close to promoter and modulate the transcriptional activity of RNA polymerase.

1.1.2 Transcription initiation signals

In bacteria, the rate of transcriptional initiation is the primary determinant of gene expression. Transcription initiation in bacteria requires RNA-polymerase and the initiation factor σ and promoter sequences. The promoters are DNA sequence elements that are present upstream of the site of transcriptional initiation and promote recognition of transcriptional start sites by RNA polymerase (Hawley and McClure 1983). The promoters vary in their affinities for RNA polymerase, a factor very important with regard to controlling the frequency of transcription and therefore the extent of gene expression. Multiple σ factors have been identified where each σ factor programs the core enzyme to transcribe from different class of promoters.

There are four notable features in most *E. coli* promoters: the transcriptional start site, the -10 hexamer, the -35 hexamer and the distance between the -10 and -35 sequences.

The transcriptional start site has been found to be purine in more than 90% of characterized promoters (Hawley and McClure, 1983). Just upstream of the start site, a six base pair (bp) region is recognizable in most promoters. The center of the hexamer is often close to 10 bp upstream of the transcriptional start site. This distance varies in known promoters from 9 to 18 from the transcriptional start site. Its consensus is TATAAT in *E.coli*. The other conserved hexamer is around ~35 bp upstream of the start site. The consensus for –35 has been universally accepted as TTGACA (Hawley and McClure, 1983). The distance separating the -35 and -10 sites has been found to be between 16 and 18 bp in 90% of the promoters (Hawley and McClure, 1983). The -35 region is said to provide the signal for recognition by RNA polymerase, while the -10 sequence allows the complex to convert from `closed` to `open` form (Hawley and McClure, 1982).

1.1.3 Transcription termination signals

In bacteria there are two mechanisms of transcription termination, intrinsic or Factor independent termination and Factor-dependent termination (Unniraman *et al*., 2002; Richardson 2002; Henkin *et al*, 1996). Usually either of these mechanisms is used to terminate transcription at the end of an operon.

Factor-independent terminator is composed of a GC rich RNA hairpin loop and a U-rich tail (Carafa *et al*., 1990). The hairpin-loop structure may stall the RNAP to proceed while the loose binding between RNA and DNA due to the rich "U" may result in the detachment of the RNA polymerase from the template, which leads to the termination of transcription process (Yarnell *et al*., 1999).

In factor-dependent termination, a protein complex containing Rho factor binds to an unstructured segment of a transcript and surveys that transcript in the $5^{'} – 3^{'}$ direction, searching for a paused RNA polymerase. If the Rho complex contacts paused RNA-polymerase, it directs RNA-polymerase to detach from transcription complex to terminate transcription (Richardson 2002).

1.2 Operon regulation

The control of gene expression can occur at many points in the transcription and translation of the genes of bacterial operons. Transcription can be regulated at the level of initiation where the activity of RNA polymerase at a given promoter is regulated by interaction with regulatory proteins, which affect its ability to recognize transcription start sites. Another processes that involve early termination of transcription are called attenuation and anti-termination.

1.2.1 Attenuation and anti-termination

Transcription of many operons that code for biochemical pathways in bacterial genomes are regulated by processes called attenuation and anti-termination (Yanofsky *et al*., 1981, 1996, 2000). Classically, attenuation occurs when the transcribed RNA upstream of an operon has the ability to fold into two mutually exclusive RNA-fold structures, one that is termed an antiterminator and the other a terminator. If the terminator hairpin loop is allowed to fold, transcription is ultimately halted. Alternatively, if the antiterminator structure folds, the terminator is precluded from folding and transcription of the operon proceeds. The mechanisms that alternate between these two RNA folds (terminators and antiterminators) are quite diverse.

Regulation by antitermination can be differentiated from attenuation by the fact that alteration of the transcription complex (rather than the alternate RNA structures) decreases the efficiency of downstream terminators. Though, in reality, the boundary between these two types of regulation is not distinct.

Attenuation and antitermination mechanisms have both been described in a wide variety of regulatory and biochemical pathways. These include operons involved in aminoacyl tRNA biosynthesis (Sarsero *et al*., 2000), amino-acid biosynthesis (Babitzke *et al*., 2003; Grundy *et al*., 1998) and several others.

1.2.2 Regulatory proteins

The regulatory protein that binds selectively to a particular DNA site in the genome is the foundation upon which transcriptional regulatory pathways are built. Hence, regulatory proteins play central role in the regulation of transcription. There are three classes of regulatory proteins: repressor, activator and dual regulator.

Repressors compete with RNA-polymerase for binding to the promoter, thereby preventing initiation. Activators interact with RNA-polymerase in a manner that can enhance binding of the RNA-polymerase to the promoter. A dual regulator either activates or represses binding of the RNA-polymerase to the promoters of two different classes of genes.

The activity of regulator proteins is modulated by small molecule such as metabolites. These molecules are considered to be the molecular signals that communicate cell metabolic state to the regulatory proteins. Signals that can be sent through this mechanism may be either negative or positive.

A negative signalling molecule would bind to a transcription regulator and allosterically modulate the protein conformation so that its affinity for DNA would decrease. This means that the transcription regulator will be less likely to be bound to the

DNA and, therefore, less likely to exert its role as a repressor or activator of transcription initiation.

A positive signaling molecule would bind to a transcription regulator (either and activator or repressor) and allosterically change the protein so that its affinity for DNA increases. This means that the regulatory protein will be more likely to be bound to the DNA and, thus, more likely to act as a repressor or activator.

1.3 Lac operon

The lac operon is one of the most basic examples of gene regulation (Figure 1.1). The lac operon contains series of structural genes, *lacZ, lacY*, and *lacA*. The *lacZ* gene codes for ß-glycosidase and the *lacY* gene codes for a lactose permease, which facilitates movement of lactose into the cell. The *lacA* gene codes thiogalactoside transacetylase. These genes are under the control of common promoter and regulatory sequences located upstream to the structural genes. RNA polymerase binds to promoter sequences to initiate the transcription of structural genes. Where as the regulatory protein (repressor) encoded by *lacI* gene binds to the regulatory sequence (operator) to repress the transcription of structural genes.

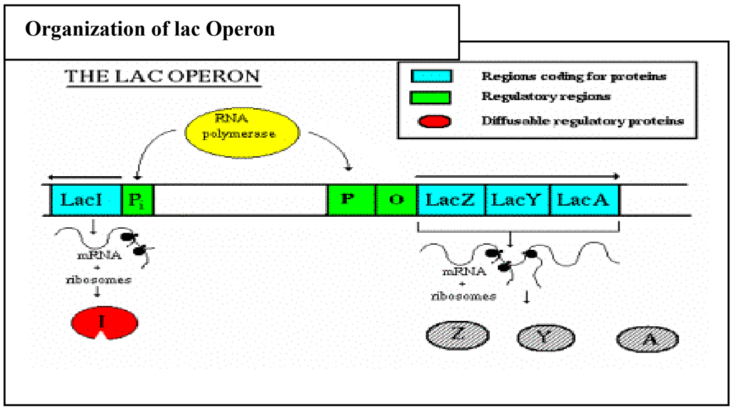
1.3.1 The lac operon is turned off by the action of the repressor

When there is no lactose present in the system, the repressor can bind to the operator region and prevent RNA polymerase from transcribing the structural genes, which are part of lac operon. As a result no mRNA for structural genes of lac operon is synthesized and corresponding protein products are not produced.

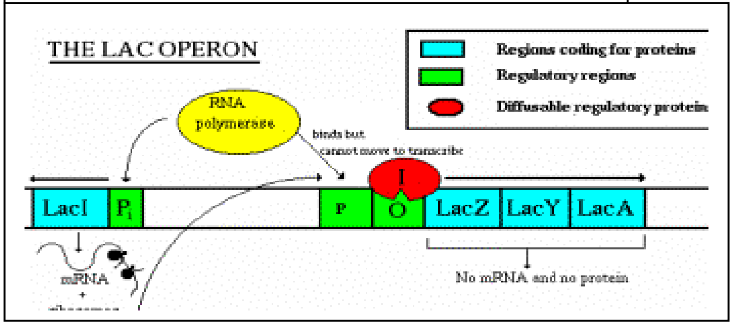
1.3.2 The lac operon is turned on in the presence of lactose

When lactose enters the cell, it binds to the repressor and changes its shape. Once this repressor-lactose complex has formed, it cannot bind to the operator region. Hence,

Organization of lac Operon



The *lac* operon is turned off by the action of the repressor



The lac operon is turned on in the presence of lactose

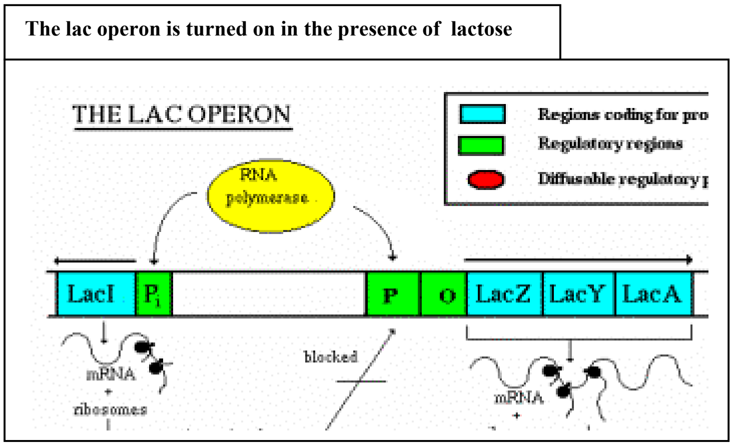


Figure 1.1: Organization of lac operon in *E. coli*

RNA-polymerase can transcribe the *lacZ, lacY*, and *lacA* genes. Translation of *lacZ* part of mRNA produces β-galactosidase, which can then convert lactose into glucose. Once all of the lactose has been utilized, the repressor can then bind to the operator region and turn the *lac* operon off.

1.4 Regulon – regulatory network

Operons are the main transcriptional regulatory units in bacteria. Often, many bacterial operons/genes contain similar upstream regulatory motifs, which are recognized by a single regulator in response to the levels of effector molecules. These operons/genes are co-regulated to form a higher-order regulatory unit called regulon.

Regulons lie at the center of gene regulation and physiological function of the organism. Regulatory interaction between regulons leads to formation of complex transcription regulatory network, which determines physiological state of the organism.

1.5 Objective and overview of the present work

One of the challenges of Functional Genomics is the identification of all the elements that take part in an organism's transcriptional regulatory network. The first step towards this goal is the identification of all the genes regulated by a transcription factor (TF), i.e. its regulon.

Identifying the Regulon is an important step towards elucidation of higher-level regulatory circuits at the whole genome level. Regulon provides useful information about gene function, and the way genes interact with each other to form molecular networks and pathways. It also helps in understanding the adaptation of bacteria to a particular environment. Most of the genes, that are part of a Regulon, code for the proteins that are collectively responsible for effective functioning of the organism. Comparative analysis of regulons in different bacteria can help in understanding of important differences and similarities between species.

Experimental efforts towards understanding the regulation of genes is laborious and expensive, but can be substantially accelerated with use of computational predictions.

In the past few years, a great amount of research has been dedicated to computational prediction of promoters (Huerta and Collado-Vides, 2003), operons (Salgado *et al*., 2000), regulatory proteins (Perez-Rueda and Collado-Vides, 2000) and transcription regulatory network (Thieffry *et al*., 1998) in the *E. coli* genome. As more and more bacterial genomes are sequenced, it is becoming more important to extend these efforts to other organisms, and decipher their transcriptional regulatory networks by means of comparative genomic studies.

Our primary goal in this study is to develop resources and algorithms to predict regulons in bacteria. The new resources and algorithms developed were used to identify regulons in two important actinobacteria pathogens, *M. tuberculosis* H37Rv and *C. diptheriae*.

An algorithm was developed for operon prediction in bacterial genomes. In bacteria the gene pairs can be grouped into convergent, divergent and co-directional category on the basis of their relative transcriptional direction. The gene pairs that belong to either convergent or divergent category are part of different operons. However, the gene pair with co-directional transcription belongs to either same operon or different operons. Conserved clusters of genes, Rho-independent transcription termination and intergenic distance were used as the signals for identification of operons from co-directionally transcribed genes. The method was used to predict operons in genomes of *E. coli K12* and *M. tuberculosis* H37Rv.

Next an algorithm was developed for genome wide prediction of potential binding sites of a regulatory protein based on Shannon relative entropy method. An interactive web server (http://www.cdfd.org.in/predictregulon/) was developed for predicting the potential binding sites and its target operons for a given regulatory protein in bacterial genomes. The program allows users to submit known or experimentally determined binding sites of a regulatory protein as ungapped multiple sequence alignments and computes the binding site recognition profile based on positional relative entropy of each base. Subsequently, this profile was used to scan the upstream regions of all genes in a user selected bacterial genome and returns the potential binding sites along with the downstream genes (operons).

The tool was applied to identify the binding sites and target genes regulated by DtxR family of transcription regulators in species of *Corynebacterium* and *Mycobacterium*. A few of the predicted binding sites were experimentally validated by electrophoretic mobility shift assay.

Further, I have shown that selection of orthologous upstream sequences on the basis of sequence similarity is a good choice for prediction of *cis*-regulatory elements by phylogenetic footprinting, a comparative genomics tool to predict *cis*-regulatory elements by finding unusually well conserved regions in orthologous upstream sequences (Bailey and Elkan, 1995; Sandelin *et al*., 2004). The basis for these tools is orthologous genes could have similar regulatory signals and the signals will be conserved during the evolution. McCue and coworkers (McCue *et al*., 2002) showed that selection of upstream sequences from 3 species is optimal for phylogenetic footprinting. He also showed that number of orthologues, phylogenetic distance, and similarity of habitat are important factors in the selection of species for phylogenetic footprinting. The orthologous upstream sequences can be completely identical, not identical but show identical regulatory signals and not identical. The first and latter types are not suitable for phylogenetic footprinting. To address this issue optimal similarity between the upstream sequences was computed to select the upstream sequences for phylogenetic footprinting irrespective to phylogenetic relationship of the species.

The approach was used to predict *cis*-regulatory elements, upstream to the operons of *M. tuberculosis* H37Rv by phylogenetic footprinting of *M. tuberculosis* H37Rv, *M. leprae TN, M. bovis AF2122/97, M. avium* subsp. paratuberculosis str. k10*,* *Nocardia farcinica* IFM 10152*, M. marinum, M. microti* and *M. smegmatis*. Novel regulatory modules were identified in *M. tuberculosis* genome via clustering of operons by predicted *cis*-regulatory elements.

Chapter 2

Prediction of Operons

The operon is a main transcription regulatory unit and the genes in the operon are usually involved in related function. The operons can be connected via regulatory proteins to form higher order regulatory circuits and functional networks. Thus, identifying the entire operon structure is an important step towards elucidating higher order regulatory circuits as well as functional networks at the whole genome level.

Experimental detection of operons using northern blot, reverse transcription polymerase chain reaction and primer extension analysis is although possible but it is costly, time-consuming and relatively difficult to implement at genomic level in the laboratory. As a result, only a modest number of operons have been documented for model organism, *E. coli* (Salgado *et al*., 2004).

Completion of many bacterial genomes has allowed the analysis of gene clusters and lead to the development of a number of algorithms for operon prediction. These algorithms differ mainly in the characteristics which are used to identify the operons: 1) Conserved clusters of genes (Overbeek *et al*., 1999); 2) Intergenic distance distributions and gene functional annotations (Salgado *et al*., 2000); 3) Genes that are within an operon contain related phylogenetic profiles and conservation of adjacency than the ones that are at the borders of transcription units (Moreno-Hagelsieb and Collado-Vides, 2002); 4) Genes in an operon tend to encode enzymes that catalyze successive reactions in metabolic pathway (Zheng *et al*., 2002); 5) Genes with in an operon shows coordinate regulation and the co-relation between the expression levels across a series of different array experiments should be equal to one (Sabatti *et al*., 2002); 6) Genes within an operon shows similar codon usage profile (Bockhorst *et al*., 2003); 7) Rho-independent transcription terminator (Chen et *al.,* 2004; Wang et *al.,* 2004).

The operon prediction methods based on first two features and Rho-independent transcription termination prediction have been relatively more successful than others. These methods are described in detail in the following sections.

The gene Pairs of Close Bi-directional Best Hits (PCBBH)

The PCBBH method detects conserved clusters of genes based on the following definitions: a set of genes occurring on a prokaryotic chromosome will be called a ‘‘run’’ if and only if they all occur on the same strand and the gaps between adjacent genes are 300 bp or less. Any pair of genes occurring within a single run is called ‘‘close’’ (Overbeek *et al.*, 1999).

Given two genes P^a and Q^a from two genomes P and Q, P^a and Q^a are called a ‘‘bidirectional best hit (BBH)’’ if and only if recognizable similarity exists between them, there is no gene P^c in P that is more similar than P^a is to Q^a, and there is no gene Q^c in Q that is more similar than Q^a is to P^a.

Genes (P^a, P^b) from P and (Q^a, Q^b) from Q form a ‘‘pair of close bidirectional best hits (PCBBH)’’ if and only if P^a and P^b are close, Q^a and Q^b are close, P^a and Q^a are a BBH, and P^b and Q^b are a BBH. The notion of a PCBBH is illustrated graphically in Figure 2.1.

After selecting a pair of genes from an organism and collecting the list of PCBBHs containing the pair, "score" the evidence that the two genes are co-occurring.

Given a pair of genes P^a and P^b from genome P ($P^{a,b}$), the score reflecting the evidence that they co-occur was computed by adding an increment for each pair ($R^{i,i}$) from genomes R_i for which ($P^{a,b}$) and ($R^{i,i}$) form a PCBBH. Add the phylogenetic distance between P and R_i to the score. The result of summing these increments is the score that offers a rough measure that the co-occurrence of P^a and P^b are meaningful.

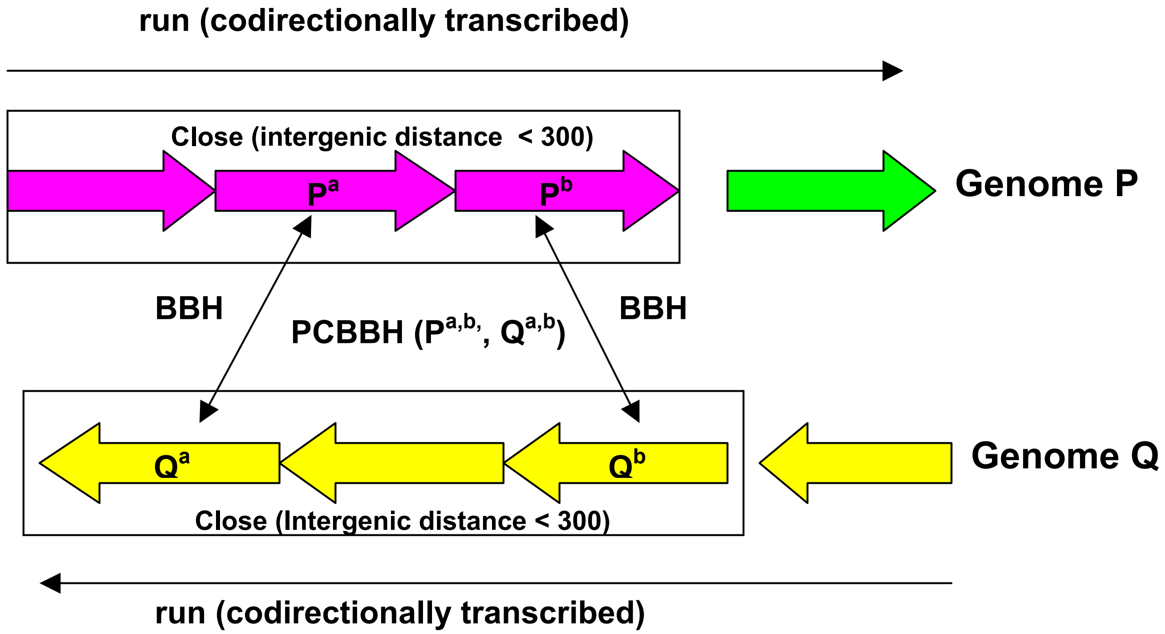

run (codirectionally transcribed)
Close (intergenic distance < 300)
P^a
P^b
Genome P
BBH
PCBBH (P^a,b,, Q^a,b)
BBH
Q^a
Q^b
Genome Q
Close (Intergenic distance < 300)
run (codirectionally transcribed)

Figure 2.1: Schematic representation of definitions of PCBBH and BBH

Intergenic distance distributions

Genes in an operon are closer than the genes at the borders of transcription units. The log-likelihood of a pair of neighboring genes being in the same operon as a function of distance was calculated with the formula:

$$LL(dist) = \log\frac{N_{\mathrm{op}}(dist)/TN_{\mathrm{op}}}{N_{\mathrm{nop}}(dist)/TN_{\mathrm{nop}}},$$

Where N_{op} and N_{nop} are pairs of genes in operons and at transcriptional boundaries, respectively, at a distance in 10-bp intervals, whereas TN_{op} and TN_{nop} are the total number of pairs of genes in operons and at the transcription unit boundaries, respectively.

Rho-independent transcription terminator prediction

There are various programs to predict Rho-independent terminators, which differ in characteristics of the Rho-independent terminators they use 1) TransTerm (Maria *et al.,* 2000) employed T weight measurement for the RNA-DNA hybrid binding site based on positional weight matrix and energy stability evaluation for the RNA hairpin structure to predict terminator; 2) RNAmotif (Lesnik *et al*., 2001) utilizes the thermodynamic parameters to measure the stability of hairpin-loop structure and its downstream sequence. The combined stability was assumed to be the determinant factor for the formation of an efficient intrinsic terminator; 3) GeSTer (Unniraman *et al*., 2002) assigned all the DNA palindrome sequences (which form RNA hairpin structures) at the intergenic regions as intrinsic terminators regardless of whether U-tails are present or not. The programs, Transterm and GCG Terminator software from the Wisconsin Package have been used to predict Rho-independent terminators, subsequently operons (Chen *et al*., 2004; Wang *et al*., 2004). But, these programs are reported to predict many false positive terminators (http://digbio.missouri.edu/~wanx/Rnall/).

In the present work, an efficient program Rnall was used to predict Rho-independent terminators. The program Rnall, first predicts the hairpin-loop structures and then filters the hairpin-loop structure using two U-tail parameters, i.e., T weight and hybridization energy.

A modified form of PCBBH, which is Index of Cluster Formation (ICF) to measure the degree of cluster formation for a pair of genes, was proposed. In addition, an efficient algorithm by combining the three characteristics, which are Rho-independent transcription terminator, intergenic distance and ICF to predict operons, was developed. The program is used to predict operons in *M. tuberculosis* genome.

2.1 Method

The complete genome sequence of *E. coli* and other bacterial genomes, used for comparative genome analysis were downloaded from NCBI (National Center for Biotechnology Information) ftp site (ftp.ncbi.nih.gov/genomes/Bacteria/).

2.1.1 Rho-independent Transcription termination prediction

The software, Rnall was used for Rho-independent transcription terminator prediction. To analyze the distribution of Factor-independent terminator relative to the translational start site, sequences of 50 bps upstream and 300 bps downstream from each stop codon of convergently transcribed genes of *E. coli* were extracted. Factor-independent terminators were predicted using the Rnall software. Figure 2.2 shows the distribution of predicted Factor-independent terminators relative to the translational stop site of convergently transcribed genes. The analysis shows that the predicted Rho-independent terminators are located with in the 50 bps upstream and 250 bps downstream from each stop codon. To predict the intrinsic terminators in entire genome of *E. coli*, sequences of 50 bps upstream and 250 bps downstream from each stop codon were extracted, based on statistics of intrinsic terminator distribution along the convergently transcribed genes. If the

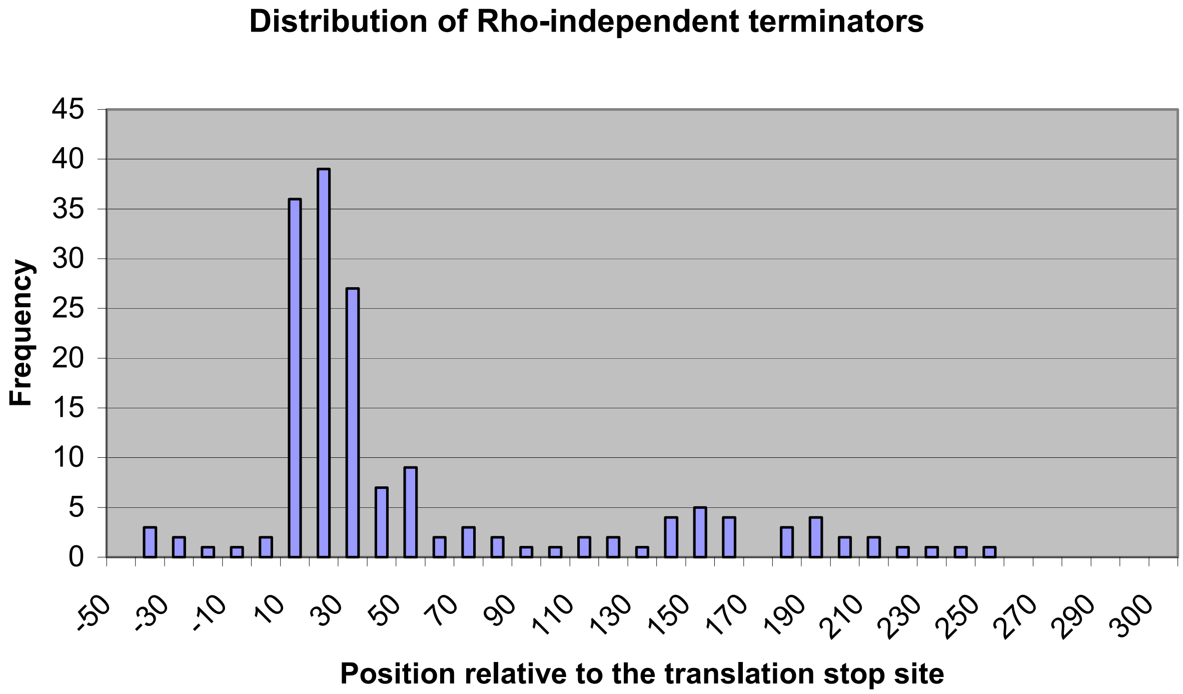


Figure 2.2: Positional distribution of predicted Rho-independent terminators relative to the translation stop site

intergenic region between the gene stop codon and its downstream gene was less than 250 bps, the intergenic sequence (together with the 50 bps upstream sequence) was extracted instead.

2.1.2 Analysis of Intergenic distance distribution

Known operons (509) of *E. coli* (http://ecocyc.org/) were taken to calculate the frequencies of intergenic distances between the gene pairs that are with in the operons. Intergenic distance between the convergently and divergently transcribed gene pairs was used as a model for the gene pairs that are at the borders of transcription units. Figure 2.3, shows the distribution of intergenic distances between the gene pairs that are with in the operons and the gene pairs that are at the borders of transcription units.

2.1.3 Index of Cluster formation (ICF)

I have proposed modified form of PCBBH, which is Index of Cluster Formation (ICF) to measure the degree of cluster formation for a pair of genes. Given a pair of genes, first identify their PCBBHs (Pair of Close Bi-directional Best Hits), then PBBHs, which are Pair of Bi-directional Best Hits, but may or may not "close".

For example a pair of genes a and b denoted by $P^{a,b}$ in a query genome P and their PCBBHs, $Q^{a,b}{}_{,}R^{a,b}$ in genomes Q and R respectively.

If an appropriate measure is given to estimate the distances between the genomes Q, R and S (correlation co-efficient of codon frequencies in two genomes), score of PCBBHs is defined as the following equation.

$$\mathrm{PCBBH_{score}}\,(P^{a,b}) = \mathrm{dist}\,(P^{a,b}{}_{,}\,Q^{a,b}) + \mathrm{dist}\,(P^{a,b}{}_{,}\,R^{a,b}) + \mathrm{dist}\,(Q^{a,b}{}_{,}\,P^{a,b}) + \mathrm{dist}\,(Q^{a,b}{}_{,}\,R^{a,b}) + \mathrm{dist}\,(R^{a,b}{}_{,}\,P^{a,b}) + \mathrm{dist}\,(R^{a,b}{}_{,}\,Q^{a,b})$$

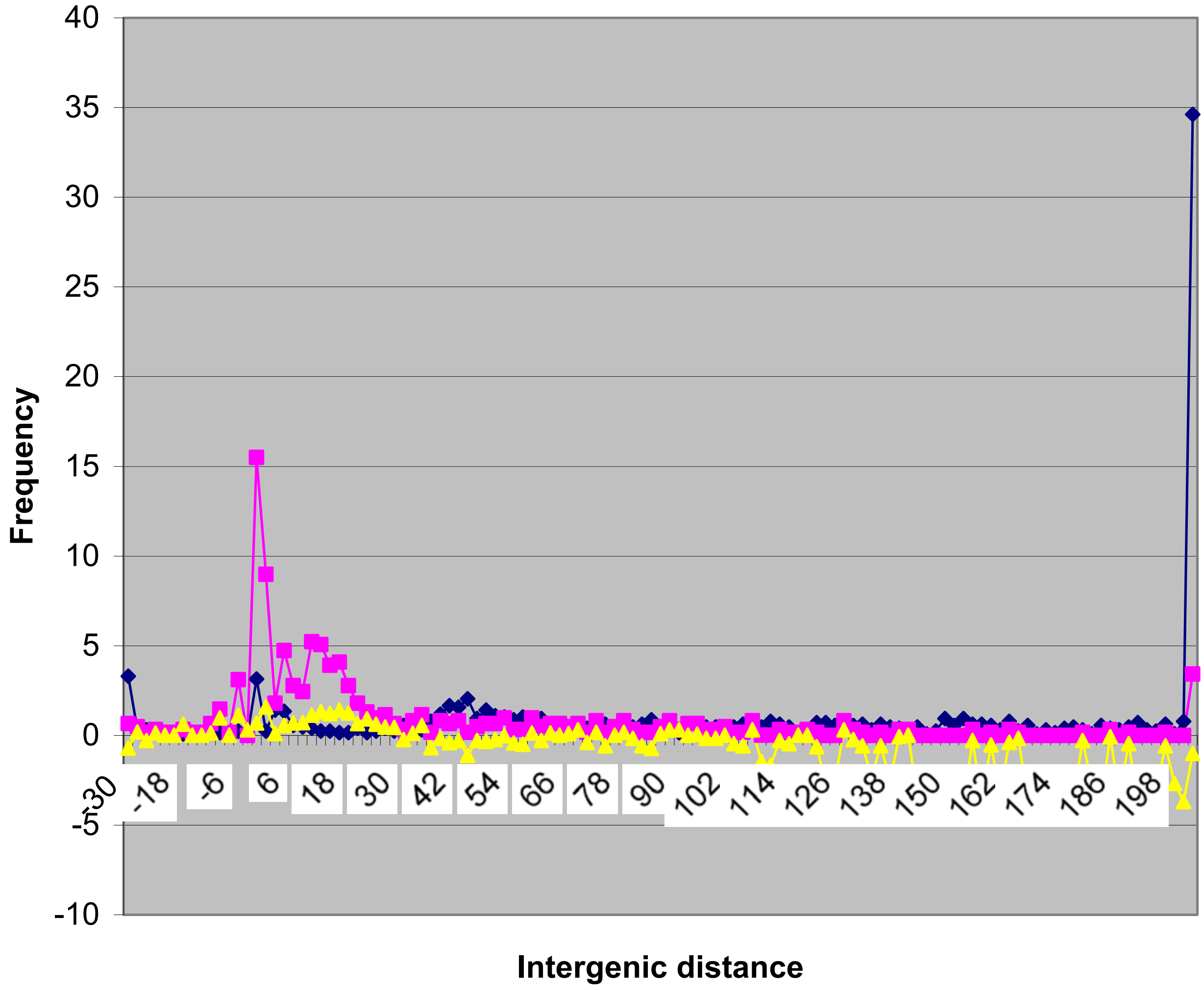


Figure 2.3: Intergenic distance distribution

Light red line represents the intergenic distances between the gene pairs that are located with in the operons; dark blue line represents the intergenic distances between the gene pairs that are at the borders of transcription units; yellow colour line represents the log likely hood scores for the distance distribution.

Similarly score the PBBHs ($PBBH_{score}$) and normalize the $PCBBH_{score}$ by dividing with $PBBH_{score}$, which gives rise to $PCBBH_{norm}$. Finally ICF was calculated by multiplying the $PCBBH_{norm}$ with $PCBBH_{score}$.

Orthologues of *E. coli* genes were identified in 106 sequenced genomes of bacteria by reciprocal best blast hits. ICF value for each gene pair was calculated as mentioned above. Known operons (509) of *E. coli* were taken (http://ecocyc.org/) to analyze the ICF values between the gene pairs that are with in the operons. ICF value of convergently transcribed gene pairs was used as a model for the gene pairs that are at the borders of transcription units. As shown in Figure 2.4, gene pairs that are with in the operons, there is an increase in ICF value in comparison to the gene pairs that are at the borders of transcription units.

2.1.4 Combined algorithm for operon prediction

In bacteria the gene pairs can be grouped into convergent, divergent and co-directional category on the basis of their relative transcriptional direction. The gene pairs that belong to either convergent or divergent category are part of different operons. However the gene pair with co-directional transcription belongs to either same operon or different operons. Rho-independent transcription termination, intergenic distance, Index of Cluster Formation (ICF) and similar gene names were considered as the as the signals for identification of operons from co-directionally transcribed genes.

Log likelihood scores were calculated for each gene pair based on distribution of intergenic distance and ICF (Figure 2.2 and 2.3). The sum of these two likelihood values gives an overall likelihood score for a candidate gene pair to be part of same operon. In absence of Rho-independent terminator, co-directionally transcribed genes were considered as part of an operon if overall likelihood score is greater than 1.1.

Similar method is applied to *M. tuberculosis*, where the log likelihood scores for intergenic distance were calculated using *E. coli* data. The log likelihood

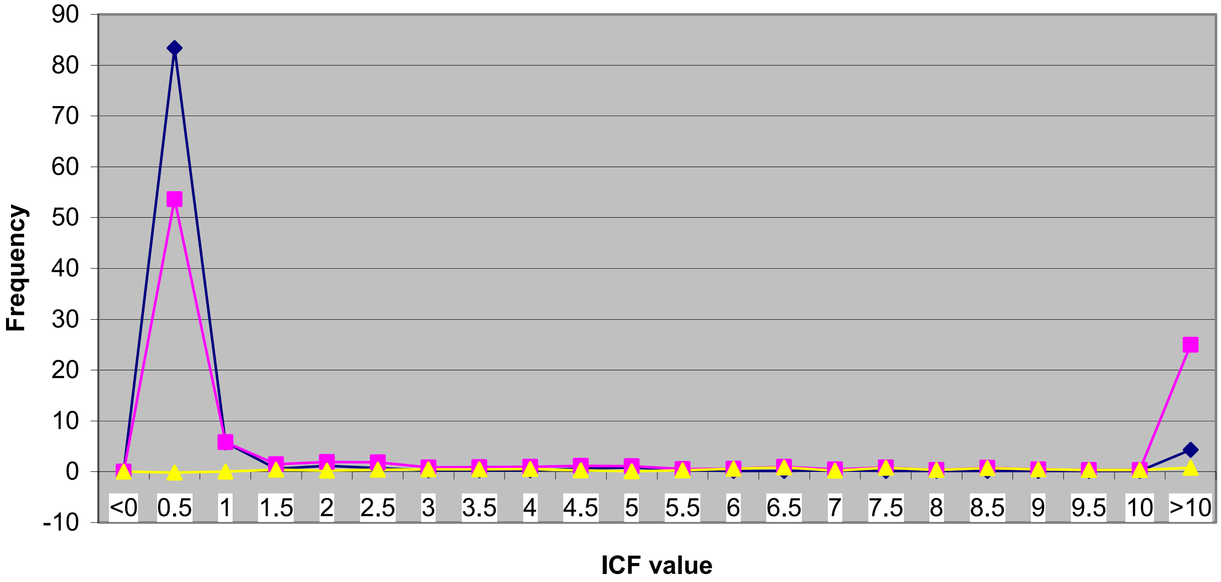


Figure 2.4: Comparative distribution of ICF value

Light red line colour represents the ICF of the gene pairs that are located with in the operons; dark blue line represents the ICF value of the gene pairs that are at the borders of transcription units; yellow colour line represents the log likely hood scores for the ICF distribution scores for ICF was calculated as following: 1) ICF value of gene pairs with high distance log likelihood scores (as a model for the gene pairs that are located with the operons) 2) ICF value of gene pairs that are convergently transcribed (as a model for the gene pairs that are located at the borders of transcription units).

2.2 Results

The combined algorithm developed by us was applied to predict operons in *E. coli*, the most studied bacterium and *M. tuberculosis*, an important pathogenic bacterium. The predicted operon list for each of these organisms can be accessed on http://www.cdfd.org.in/predictregulon/operons/.

2.2.1 Prediction of operons in *E. coli*

In order to evaluate the performance of the operon prediction method described above, the method is applied to the well-studied bacterial genome, *E. coli* K12. The method could predict the 450 of 509 (88%) experimentally identified operons (http://ecocyc.org/). There are 2651 predicted transcription units of which 771 are polycistronic. Among 2651 predicted transcription units, 846 contain Rho-independent transcription terminators and the rest of them are likely to have Rho-dependent transcription terminators.

Among 846 polycistronic units, 335 contain at least one gene that code for hypothetical protein with unknown function. Analysis of these genes suggests that they might fall into a similar functional category with their gene neighbors with in polycistron. In case in which these genes could not be annotated by a conventional sequence comparison method, identification of an operon around it may help us to unravel its functional role. For example Table 2.1 shows operons containing the genes (*ybdB, yabB, yafQ* and *ybaB*) that code for hypothetical proteins. The current annotation status for *ydbB* encodes for a protein, belonging to a large family of enzymes (pfam03061.11), which function primarily in thiol template-directed fatty acid and polyketide biosynthetic pathways. The

Table 2.1: Example of predicted *E. coli* operons containing hypothetical proteins

Gene	Synonym	Product
entC	b0593	Isochorismate Hydroxymutase 2, Enterochelin Biosynthesis
entE	b0594	2,3-Dihydroxybenzoate-AMP Ligase
entB	b0595	2,3-Dihydro-2,3-Dihydroxybenzoate Synthetase, Isochroismatase
entA	b0596	2,3-Dihydro-2,3-Dihydroxybenzoate Dehydrogenase, Enterochelin Biosynthesis
ybdB	**b0597**	**Orf, Hypothetical Protein**
yabB	**b0081**	**Orf, Hypothetical Protein**
yabC	b0082	Putative Apolipoprotein
ftsL	b0083	Cell Division Protein; Ingrowth Of Wall At Septum
ftsI	b0084	Septum Formation; Penicillin-Binding Protein 3; Peptidoglycan Synthetase
murE	b0085	Meso-Diaminopimelate-Adding Enzyme
murF	b0086	D-Alanine:D-Alanine-Adding Enzyme
mraY	b0087	Phospho-N-Acetylmuramoyl-Pentapeptide Transferase?
murD	b0088	UDP-N-Acetylmuramoylalanine-D-Glutamate Ligase
ftsW	b0089	Cell Division; Membrane Protein Involved In Shape Determination
murG	b0090	UDP-N-Acetylglucosamine Pyrophosphoryl-Undecaprenol N-Acetylglucosamine Traseferase
murC	b0091	L-Alanine Adding Enzyme, UDP-N-Acetyl-Muramate:Alanine Ligase
ddlB	b0092	D-Alanine-D-Alanine Ligase B, Affects Cell Division
ftsQ	b0093	Cell Division Protein; Ingrowth Of Wall At Septum
ftsA	b0094	ATP-Binding Cell Division Protein, Septation Process, Complexes With Ftsz,
ftsZ	b0095	Cell Division Protein Tubulin-Like GTP-Binding Protein And Gtpase
yafQ	**b0225**	**Orf, Hypothetical Protein**
dinJ	b0226	Damage-Inducible Protein J
dnaX	b0470	DNA Polymerase III, Tau And Gamma Subunits; DNA Elongation Factor III
ybaB	**b0471**	**Orf, Hypothetical Protein**
recR	b0472	Cog0353_16
sdhC	b0721	Succinate Dehydrogenase, Cytochrome B556
sdhD	b0722	Succinate Dehydrogenase, Hydrophobic Subunit
sdhA	b0723	Succinate Dehydrogenase, Flavoprotein Subunit
sdhB	b0724	Succinate Dehydrogenase, Iron Sulfur Protein
-	**b0725**	**Orf, Hypothetical Protein**
sucA	b0726	2-Oxoglutarate Dehydrogenase (Decarboxylase Component)
sucB	b0727	2-Oxoglutarate Dehydrogenase (Dihydrolipoyltranssuccinase E2 Component)

Note: Genes that are part of an operon are together

results show that *ybdB* is associated with the genes *entA*, *entB*, *entE* and *entC*. These genes encode the enzymes that are involved in siderophore (enterochelin) biosynthesis. Hence, it is likely that YbdB might also be involved in the siderophore biosynthesis pathway. Orthologues of YbdB are widely distributed in sequenced bacterial genomes including species of mycobacteria. It was described in following chapters, that orthologues of YbdB are present across the predicted iron dependent regulons of *Mycobacterium* and speculate that they could be involved in biosynthesis of *Mycobacterium* siderophores.

2.2.2 Prediction of operons in *M. tuberculosis*

There are 2255 predicted transcription units of which 743 are polycistronic. Among 2255 predicted transcription units, 106 contain predicted Rho-independent transcription terminators. Table 2.2 shows some of the operons containing the hypothetical proteins, whose function, might fall into a similar functional category with their gene neighbors with in the operon.

For example the genes Rv1846c, predicted code for a transcription regulator and Rv1845c, codes for a protein with unknown function belong to same operon. RPS-BLAST search against CDD databases shows that the gene, Rv1846c codes for a BlaI family of transcription regulator and the other gene Rv1845c codes for BlaR1 family of protein. The two families of proteins together confer resistance to variety of β-lactum antibiotics and widely distributed in pathogenic bacteria. In *Staphylococcus aureas*, BlaR1 family of protein MecR1, present in the cytoplasmic membrane, detects the β-lactum by means of an extracellular penicillin binding-domain and transmits the signal via a second intracellular zinc metalloprotease signalling domain. Binding of a β-lactum to MecR1 stimulates the autocatalytic conversion of intracellular Zinc metaloprotease signalling domain of MecR1 from an inactive proenzyme to an active protease. The activated form of MecR1 cleaves BlaI family of transcription regulator, MecI and de-represses the transcription of β-lactamase (Hanique *et al*., 2004).

Table 2.2: Example of predicted *M. tuberculosis* operons containing hypothetical proteins

Gene	Synonym	Product
-	**Rv1845c**	**Hypothetical protein**
-	Rv1846c	Predicted transcription regulator
gyrB	Rv0005	Type IIA topoisomerase
gyrA	Rv0006	Type IIA topoisomerase
-	**Rv0007**	**Hypothetical protein**
-	**Rv0282**	**Hypothetical protein**
-	**Rv0283**	**Hypothetical protein**
FtsK	Rv0284	Dna segregation atpase ftsk/spoiiie and related proteins
PE	Rv0285	-
PPE	Rv0286	Ppe-repeat proteins
hemL	Rv0524	Glutamate-1-semialdehyde aminotransferase
-	**Rv0525**	**Hypothetical protein**
-	**Rv0526**	**Hypothetical protein**
ccsA	Rv0527	Cytochrome c biogenesis protein
resB	Rv0528	Resb protein required for cytochrome c biosynthesis
ccsB	Rv0529	Abc-type transport system involved in cytochrome c biogenesis
pyrR	Rv1379	Pyrimidine operon attenuation protein/uracil phosphoribosyltransferase
pyrB	Rv1380	Aspartate carbamoyltransferase
pyrC	Rv1381	Dihydroorotase and related cyclic amidohydrolases
-	**Rv1382**	**Hypothetical protein**
carA	Rv1383	Carbamoylphosphate synthase small subunit
carB	Rv1384	Carbamoylphosphate synthase large subunit (split gene in mj)
pyrF	Rv1385	Pyrf
-	**Rv3662c**	**Hypothetical protein**
dppD	Rv3663c	Atpase components of various abc-type transport systems
dppC	Rv3664c	Abc-type dipeptide/oligopeptide/nickel transport systems
dppB	Rv3665c	Abc-type dipeptide/oligopeptide/nickel transport systems
dppA	Rv3666c	Abc-type oligopeptide transport system
parE	Rv1959c	Plasmid stabilization system protein
-	**Rv1960c**	**Hypothetical protein**
parA	Rv3917c	Probable chromosome partitioning protein
parB	Rv3918c	Atpase, involved in chromosome partitioning protein
gid	Rv3919c	Probable glucose-inhibited division protein

Note: Genes that are part of an operon are together

Since, upstream sequence to the first gene of the transcription unit contains the regulatory sequence, the predicted transcription units further used for selection of upstream sequences to predict *cis*-regulatory elements as described in following chapters.

Chapter 3

Prediction of Regulons from Regulatory Sites

Regulatory proteins sense the environmental and cellular conditions and binds to the upstream regulatory site of operons to alter the expression level of operon-encoded genes according to the need of bacteria. A group of genes regulated by a given regulatory protein are called regulon. Genes that are part of a regulon code for proteins those are collectively responsible for physiological activities shown by the bacteria in a given cellular conditions/environment.

The advent of the genomic era has generated interest in developing computational methods to predict the regulons/co-regulated genes in prokaryote genomes. These methods rely upon identification of common regulatory sites in upstream sequences of different operons/genes that are part of a regulon. The computational methods use the features of known sites or depend on various other characteristics of regulatory sites such as statistics and sequence conservation.

The method based on consensus is assigns consensus nucleotide symbol to describe the nucleotide composition in each column of the aligned binding sites usually following IUPAC conventions (Schneider and Stephens, 1990). The disadvantage with this approach is that a single symbol cannot quantitatively describe the nucleotide preference at specific position on the DNA. The other methods based profile search, initially constructs a model of aligned binding sites by counting the frequency of nucleotides in each in alignment column. A matrix is built out of nucleotide frequency in this is referred as a positional frequency matrix. The matrix can be normalized by dividing each element of the matrix by total number motifs, which gives positional probability matrix. The chance of observing particular site is product of the relevant probability-matrix cell for each nucleotide (Schneider, 1997).

The thesis work describes a novel profile method based on Shannon relative entropy, which considers the positional probability nucleotides within the aligned binding sites and the probability of nucleotides in genome sequence. This method can utilize the available experimental data on binding sites of transcription regulatory proteins from various bacterial species (Salgado *et al*., 2004) for identification of regulatory elements in phylogentically related species.

3.1 Method

The program, first constructs the binding site recognition profile based on un-gapped multiple sequence alignment of known binding sites. This profile is calculated using Shannon's positional relative entropy approach (Shannon *et al*., 1948). The positional relative entropy Q_i at position *i* in a binding site is defined as

$$Q_i = \sum_{b=A,T,G,C} f_{b,i} \log_{10} \frac{f_{b,i}}{q_b}$$

Where *b* refers to each of the possible base (A, T, G, C), $f_{b,i}$ is observed frequency of each base at position *i* and q_b is the frequency of base *b* in the genome sequence. The contribution of each base to the positional Shannon relative entropy is calculated by multiplying each base frequency with positional relative entropy as follows,

$$W_{b,i} = f_{b,i} \cdot Q_i$$

Where $W_{b,i}$ refers to the weighted Shannon relative entropy of the base *b* (A, T, G, C) at position *i.* Finally, a 4 X L entropy matrix (L is the length of the binding site) is constructed representing the binding site recognition profile, where each matrix element is the weighted positional Shannon relative entropy of a base.

The profile, encoded as the matrix, is used to scan the upstream sequences of all the genes of user-selected genome. Entropy score of each site is calculated as the sum of the respective positional nucleotide entropy ($W_{b,i}$). Maximally scoring site is selected from the upstream sequence of each gene. The score may represent the strength of interaction between regulatory protein and binding site (Benos *et al*., 2002). Least score among the experimentally known binding sites is considered as cut-off score. The sites scoring higher than the cut-off value are reported as potential binding sites conforming to the consensus profile. The gene down stream to the predicted binding site is considered as start gene of the operon. Further downstream operon organization was predicted using the method described in Chapter 2. The operons/genes downstream to the predicted binding site were considered as a regulon.

3.2 Results and Discussion

LexA binding sites and target genes in *M. tuberculosis* were prediced using the LexA binding sites of *B. subtilis*. LexA regulators from *B. subtilis* and *M. tuberculosis* share a high sequence identity (45%) at protein level. Table 3.1 lists the known LexA binding sites from *B. subtilis* given as input to the program and Table 3.2 shows the output of predicted LexA binding sites in *M. tuberculosis*. The site column in Table 3.2 represents the predicted binding sites of LexA in *M. tuberculosis.* Eighteen of these genes (indicated by asterisk) belonging to the LexA regulon was also observed in data obtained by experimental means by others (Durbach *et al*., 1997, Brooks *et al*., 2001; Brooks *et al*., 2002, Boshoff *et al*., 2003). The rest of the matches are likely to be novel regulatory sites.

This method has two specific requirements: 1) The availability of a few experimentally determined regulatory protein binding sites for developing the binding site recognition profile 2) The profile should be applicable to the genome where the regulator or its homologue is present. In absence of any experimental information on the regulatory sites in a given genome one may lookup the known regulatory motifs from other related species.

A limitation of this approach is that it may predict few false positive sites as candidates. However this limitation can be overcome by experimental validations, either by *in vitro* binding studies with double strand oligonucleotides containing the binding sites (designed based on prediction) and regulatory proteins and Real Time PCR analysis of candidate co- regulated genes.

Table 3.1: Known LexA binding sites of *Bacillus subtilis from* PRODORIC database

Binding Site	Gene
AGAACAAGTGTTCG	*din*C
AGAACTCATGTTCG	*din*B
CGAACTTTAGTTCG	*din*A
CGAATATGCGTTCG	*rec*A
CGAACGTATGTTTG	*din*C
CGAACCTATGTTTG	*din*R
CGAACAAACGTTTC	*din*R
GGAATGTTTGTTCG	*din*R

Table 3.2: Output of Predictregulon web server (predicted LexA binding sites)

Score	Position	Site	Gene	Synonym	COG	Product
5.37	-8	CGAACGTATGTTCG	-	Rv3776*	-	Hypothetical protein Rv3776
5.32	-100	CGAACATGTGTTCG	-	Rv3073c*	COG3189	Uncharacterized conserved protein
5.32	-144	CGAACATGTGTTCG	pyrR	Rv1379*	COG2065	Pyrimidine operon attenuation protein
5.22	-8	CGAACACATGTTCG	-	Rv3074*	-	Hypothetical protein Rv3074
5.2	-142	CGAACAATTGTTCG	-	Rv3371*	-	Hypothetical protein Rv3371
5.2	-64	CGAACAATTGTTCG	dnaE2	Rv3370c*	COG0587	DNA polymerase III
5.19	-36	CGAACGATTGTTCG	ruvC	Rv2594c*	COG0817	ruvC
5.14	-32	CGAAAGTATGTTCG	-	Rv0336*	-	Hypothetical protein Rv0336
5.14	-32	CGAAAGTATGTTCG	-	Rv0515*	-	Hypothetical protein Rv0515
5.14	-105	CGAACACATGTTTG	lexA	Rv2720*	COG1974	SOS-response transcriptional repressors
5.11	-122	CGAACAGGTGTTCG	recA	Rv2737c*	COG1372	recA
5.08	-87	CGAACAATCGTTCG	-	Rv2595*	COG2002	Hypothetical protein Rv2595
5.06	-44	CGAATATGCGTTCG	dnaB	Rv0058*	COG0305	Replicative DNA helicase
5.04	-263	GGAACTTGTGTTGG	UbiE	Rv3832c	COG2226	Methylase involved in ubiquinone biosynthesis
5.04	-23	AGAACGGTTGTTCG	SplB	Rv2578c*	COG1533	DNA repair photolyase
5.02	-6	CGAATATGAGTTCG	-	Rv0071*	COG3344	Retron-type reverse transcriptase
5.01	-255	CGAACAAGTGTTGG	-	Rv1414	COG3616	Predicted amino acid aldolase or racemase
4.99	-181	GGAACGCGTGTTTG	-	Rv0750	-	Hypothetical protein Rv0750
4.98	-105	CGAACAACAGTTCG	BaeS	Rv0600c	COG0642	Signal transduction histidine kinase
4.98	-186	CGAAGATGCGTTCG	rpsT	Rv2412	COG0268	Ribosomal protein S20
4.95	-242	TGAACGCAAGTTCG	fbpB	Rv1886c	COG0627	fbpB
4.95	-192	CGAACGGGAGTTCG	-	Rv1455	-	Hypothetical protein Rv1455
4.94	-270	AGAACCACCGTTCG	Phd	Rv3181c	COG4118	Antitoxin of toxin-antitoxin stability system
4.94	-213	CGAACGACGGTTCG	PE	Rv2099c*	-	PE
4.92	-118	CGAACAGGTGTTGG	-	Rv0004	COG5512	Zn-ribbon-containing
4.92	-163	CGAACTTGCGTTCA	-	Rv1887	-	Hypothetical protein Rv1887
4.91	-239	GGAACGCGAGTTCG	fadB2	Rv0468	COG1250	3-hydroxyacyl-CoA dehydrogenase
4.91	-7	TGAACGAATGTTCC	-	Rv0039c	-	Hypothetical protein Rv0039c
4.9	-237	CGAAGCCTTGTTCG	DltE	Rv3174	COG0300	Short-chain dehydrogenase
4.89	-225	GGAAGGTGCGTTCG	FrnE	Rv2466c	COG2761	Predicted dithiol-disulfide isomerase
4.88	-8	GGAAGCCATGTTCG	-	Rv0769	COG1028	Hypothetical protein Rv0769
4.88	-186	CGAAGAGGTGTTCG	CoxS	Rv0374c	COG2080	Aerobic-type carbon monoxide dehydrogenase
4.88	-186	CGAACCGCAGTTCG	LeuA	Rv3534c	COG0119	Isopropyl malate/citramalate synthases
4.85	-195	CGAACGGCTGTTGG	-	Rv2061c	COG3576	Hypothetical protein Rv2061c
4.85	-85	AGAACGGTTGTTGG	accA1	Rv2501c	COG4770	accA1
4.84	-151	CGAAATTGTGTTCC	nuoB	Rv3146	COG0377	NADH:ubiquinone oxidoreductase
4.84	-217	CAAACATGTGTTCG	-	Rv2719c*	-	Hypothetical protein Rv2719c
4.84	-5	CGAACATGTATTCG	-	Rv1702c*	-	Hypothetical protein Rv1702c
4.84	-199	CGAAATCTTGTTTG	-	Rv1375	COG1944	Hypothetical protein Rv1375

Note: Score: score of the binding sites, Position: position of the binding site relative to the translation start site, Site: binding site of a regulatory protein, Gene: gene downstream to the binding site, Synonym: synonym of the gene, COG: Cluster of Orthologous Gene code, Product: Gene product; * represents the ORFs known to be regulated by the transcription regulator, lexA.

Chapter 4

Preidictregulon Webserver

As a service to a wider scientific community, webserver called Predictregulon was devolved for prediction of binding sites and target operons a regulatory protein.. Predictregulon is accessible to all through Internet via CDFD website (http://www.cdfd.org.in/predictregulon/).

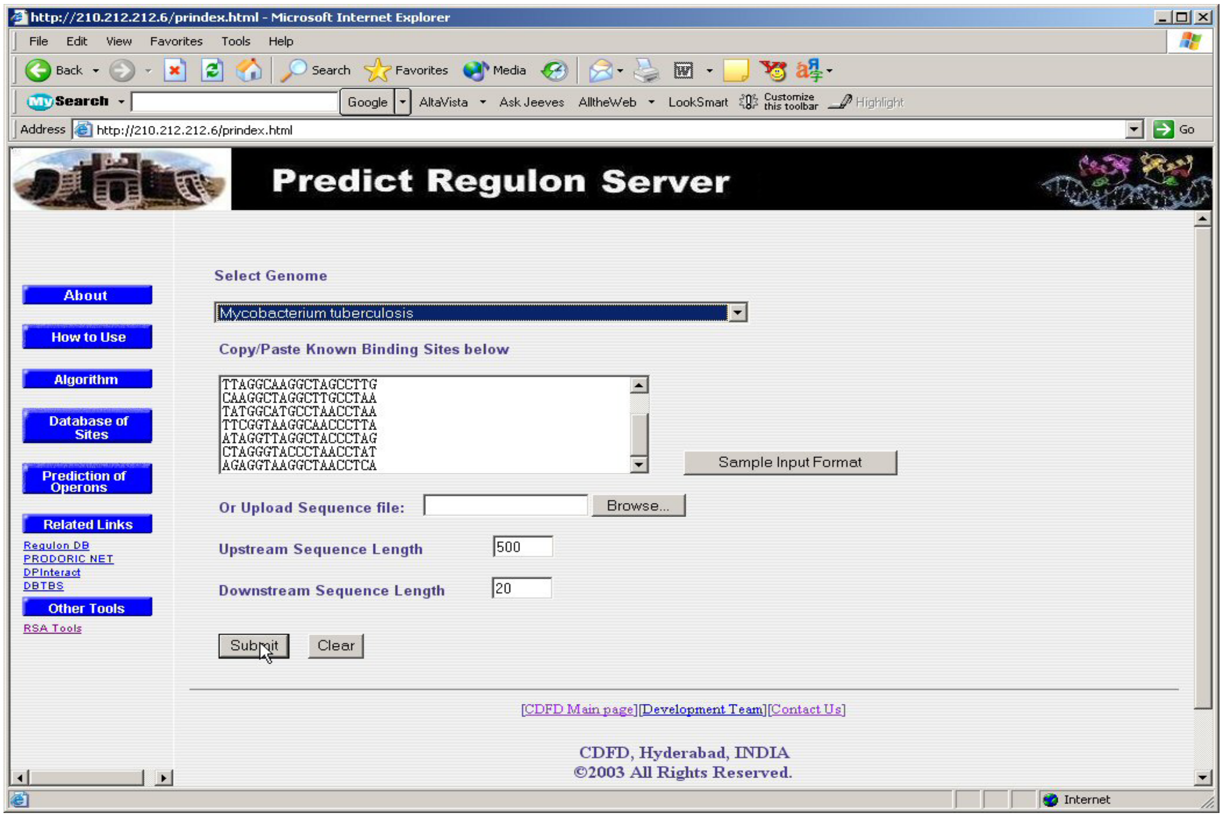
4.1 Web implementation

Predictregulon consists of an HTML interface form. This form accepts the parameters to be used with Predictregulon algorithm. These include: 1) the genome to be scanned 2) known or experimentally determined binding sites of a regulatory protein as un-gapped multiple sequence alignments and 3) definition of start and end of the upstream region with respect to translation start site of a gene.

The parmeters filled in through the prdictregulon form is passed to regulon search program implemented through a CGI script interface. The regulon search analyses the upstream regions of all genes in a user-selected prokaryote genome and returns the potential binding sites along with the downstream co-regulated genes (operons). The known binding sites of a regulatory protein can also be used to identify its orthologue binding sites in phylogentically related genomes where the trans-acting regulator protein and cognate *cis*-acting DNA sequences could be conserved.

4.2 Using Predictregulon

Use of the Predictregulon system is illustrated using Figure 4.1-4.4. This figures represents screenshots of all the analyses that can be performed, and the options available at each step. The *M. tuberculosis* Iron dependent regulon (IdeR) is used as an example.
In the first step, select a genome of given species in which binding-sites of regulatory protein is to be identified. The selection of genome is made through a drop down combobox list of genomes.



Address: http://210.212.212.6/prindex.html

Select Genome: Mycobacterium tuberculosis

Copy/Paste Known Binding Sites below

```
TTAGGCAAGGCTAGCCTTG
CAAGGCTAGGCTTGCCTAA
TATGGCATGCCTAACCTAA
TTCGGTAAGGCAACCCTTA
ATAGGTTAGGCTACCCTAG
CTAGGGTACCCTAACCTAT
AGAGGTAAGGCTAACCTCA
```

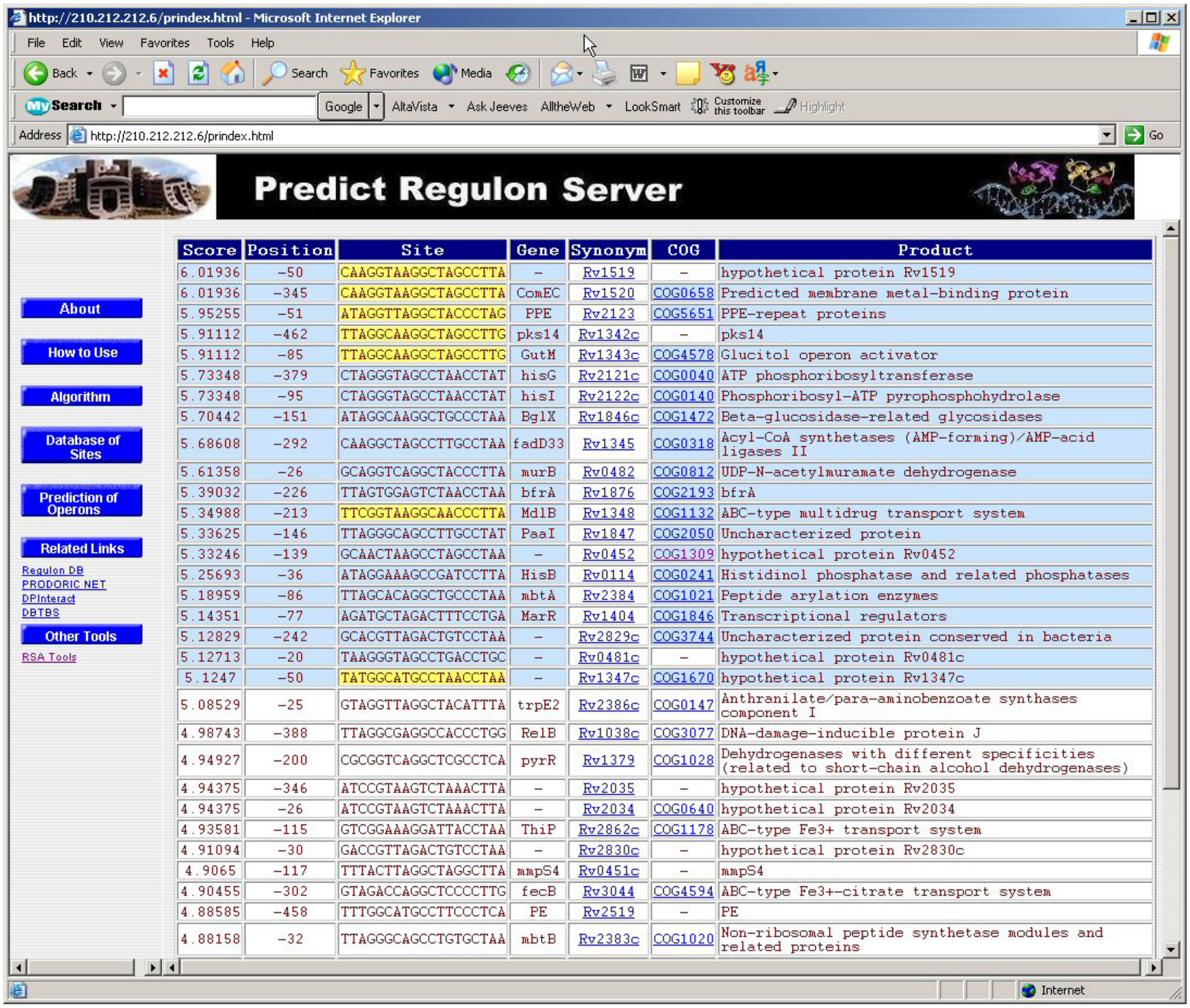
Or Upload Sequence file:

Upstream Sequence Length: 500

Downstream Sequence Length: 20

Figure 4.1: the web submission form

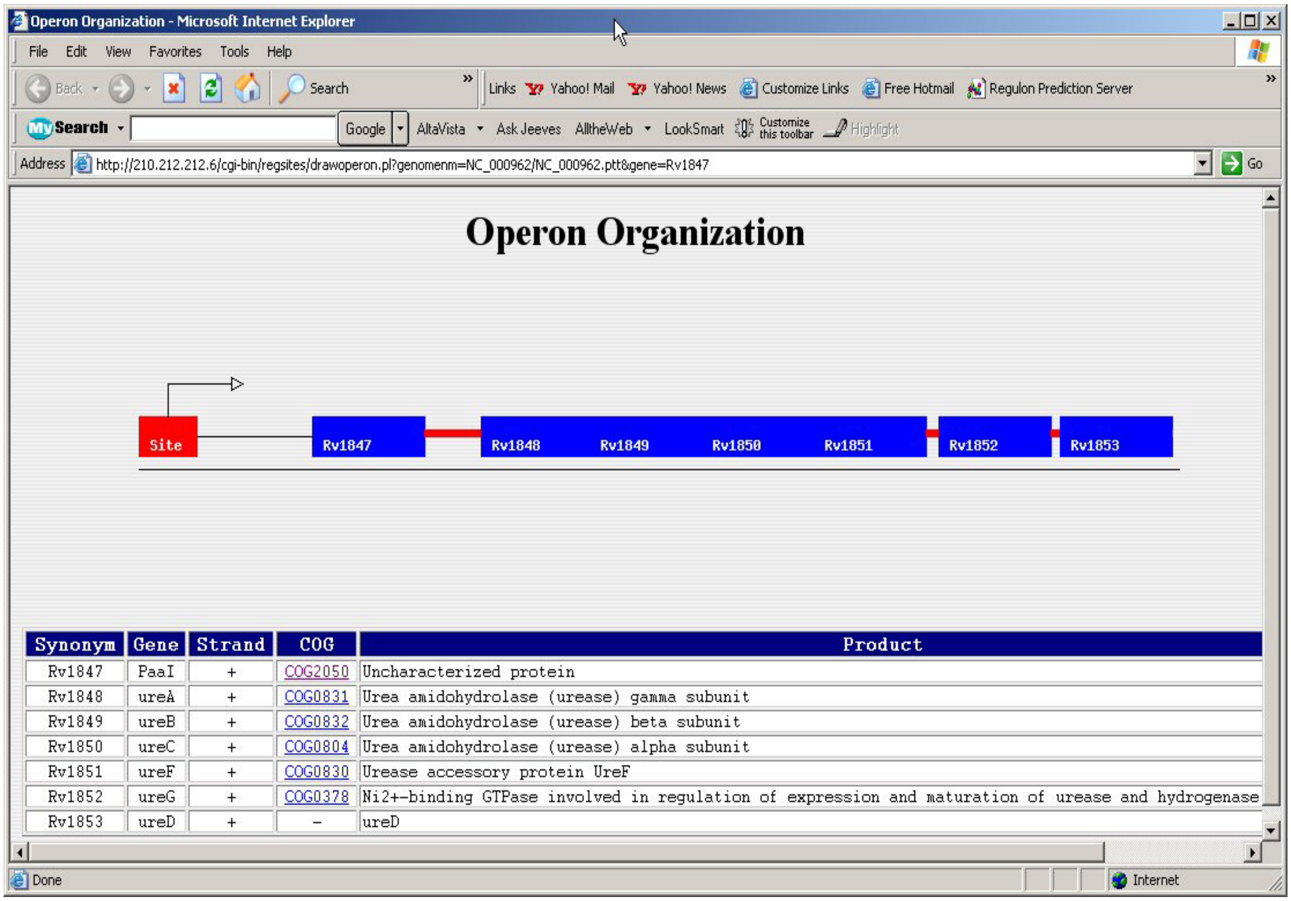
The user has to select the species name, length of upstream and downstream sequence relative to the translation start site of the gene.



Score	Position	Site	Gene	Synonym	COG	Product
6.01936	-50	CAAGGTAAGGCTAGCCTTA	–	Rv1519	–	hypothetical protein Rv1519
6.01936	-345	CAAGGTAAGGCTAGCCTTA	ComEC	Rv1520	COG0658	Predicted membrane metal-binding protein
5.95255	-51	ATAGGTTAGGCTACCCTAG	PPE	Rv2123	COG5651	PPE-repeat proteins
5.91112	-462	TTAGGCAAGGCTAGCCTTG	pks14	Rv1342c	–	pks14
5.91112	-85	TTAGGCAAGGCTAGCCTTG	GutM	Rv1343c	COG4578	Glucitol operon activator
5.73348	-379	CTAGGGTAGCCTAACCTAT	hisG	Rv2121c	COG0040	ATP phosphoribosyltransferase
5.73348	-95	CTAGGGTAGCCTAACCTAT	hisI	Rv2122c	COG0140	Phosphoribosyl-ATP pyrophosphohydrolase
5.70442	-151	ATAGGCAAGGCTGCCCTAA	BglX	Rv1846c	COG1472	Beta-glucosidase-related glycosidases
5.68608	-292	CAAGGCTAGCCTTGCCTAA	fadD33	Rv1345	COG0318	Acyl-CoA synthetases (AMP-forming)/AMP-acid ligases II
5.61358	-26	GCAGGTCAGGCTACCCTTA	murB	Rv0482	COG0812	UDP-N-acetylmuramate dehydrogenase
5.39032	-226	TTAGTGGAGTCTAACCTAA	bfrA	Rv1876	COG2193	bfrA
5.34988	-213	TTCGGTAAGGCAACCCTTA	MdlB	Rv1348	COG1132	ABC-type multidrug transport system
5.33625	-146	TTAGGGCAGCCTTGCCTAT	PaaI	Rv1847	COG2050	Uncharacterized protein
5.33246	-139	GCAACTAAGCCTAGCCTAA	–	Rv0452	COG1309	hypothetical protein Rv0452
5.25693	-36	ATAGGAAAGCCGATCCTTA	HisB	Rv0114	COG0241	Histidinol phosphatase and related phosphatases
5.18959	-86	TTAGCACAGGCTGCCCTAA	mbtA	Rv2384	COG1021	Peptide arylation enzymes
5.14351	-77	AGATGCTAGACTTTCCTGA	MarR	Rv1404	COG1846	Transcriptional regulators
5.12829	-242	GCACGTTAGACTGTCCTAA	–	Rv2829c	COG3744	Uncharacterized protein conserved in bacteria
5.12713	-20	TAAGGGTAGCCTGACCTGC	–	Rv0481c	–	hypothetical protein Rv0481c
5.1247	-50	TATGGCATGCCTAACCTAA	–	Rv1347c	COG1670	hypothetical protein Rv1347c
5.08529	-25	GTAGGTTAGGCTACATTTA	trpE2	Rv2386c	COG0147	Anthranilate/para-aminobenzoate synthases component I
4.98743	-388	TTAGGCGAGGCCACCCTGG	RelB	Rv1038c	COG3077	DNA-damage-inducible protein J
4.94927	-200	CGCGGTCAGGCTCGCCTCA	pyrR	Rv1379	COG1028	Dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases)
4.94375	-346	ATCCGTAAGTCTAAACTTA	–	Rv2035	–	hypothetical protein Rv2035
4.94375	-26	ATCCGTAAGTCTAAACTTA	–	Rv2034	COG0640	hypothetical protein Rv2034
4.93581	-115	GTCGGAAAGGATTACCTAA	ThiP	Rv2862c	COG1178	ABC-type Fe3+ transport system
4.91094	-30	GACCGTTAGACTGTCCTAA	–	Rv2830c	–	hypothetical protein Rv2830c
4.9065	-117	TTTACTTAGGCTAGGCTTA	mmpS4	Rv0451c	–	mmpS4
4.90455	-302	GTAGACCAGGCTCCCCTTG	fecB	Rv3044	COG4594	ABC-type Fe3+-citrate transport system
4.88585	-458	TTTGGCATGCCTTCCCTCA	PE	Rv2519	–	PE
4.88158	-32	TTAGGGCAGCCTGTGCTAA	mbtB	Rv2383c	COG1020	Non-ribosomal peptide synthetase modules and related proteins

Figure 4.2: Output of Predictregulon

Column 1 - Score of the binding site, sites with the score above the cut-off score are highlighted with blue background. Column 2 – Position of binding site relative to the translation start site. Column 3 - shows the list of known binding sites as well as predicted binding sites. Known binding sites are highlighted with Yellow background. Column 4 - gene downstream to the predicted binding site. Column 5 - Synonym of the gene. Column - 6 Cluster of orthlogous gene code - (COG). Column 7 - Function of the gene product.

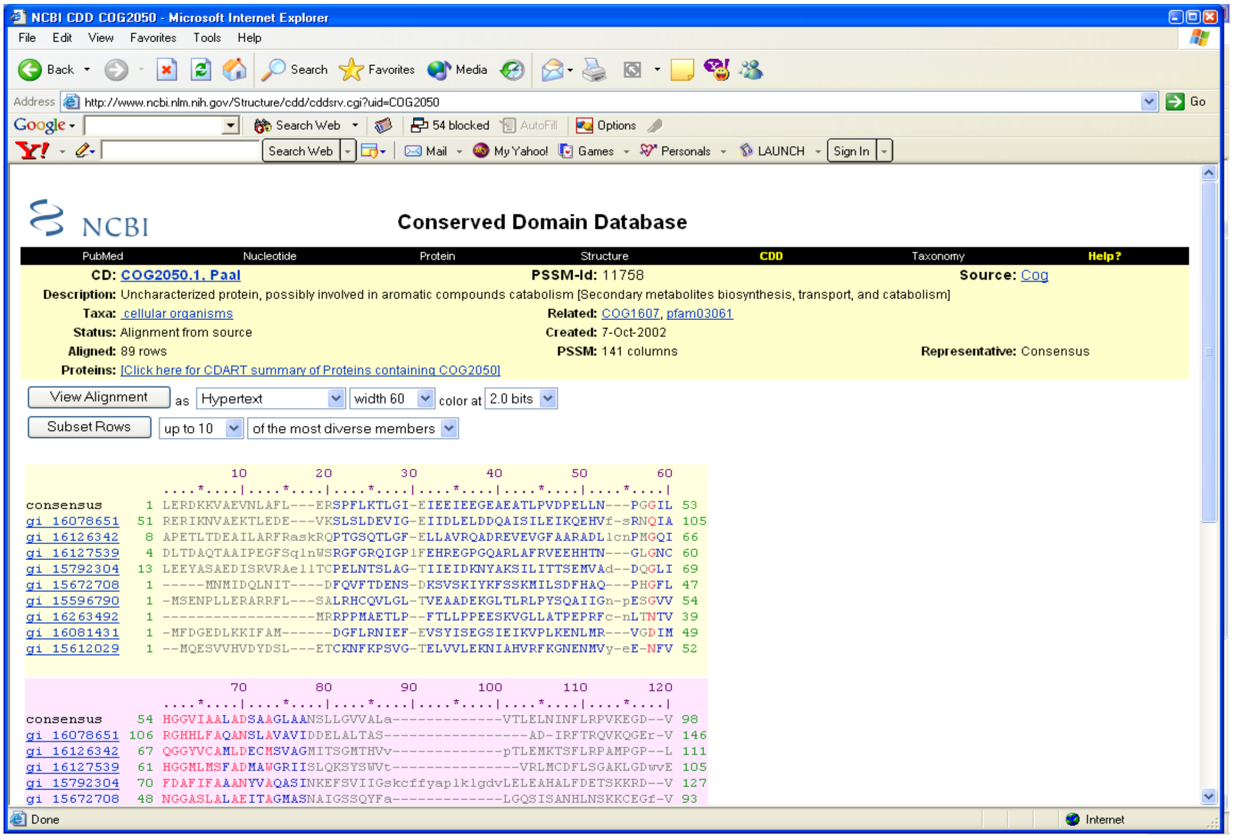

Operon Organization - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search
Links Yahoo! Mail Yahoo! News Customize Links Free Hotmail Regulon Prediction Server
MySearch Google AltaVista Ask Jeeves AlltheWeb LookSmart Customize this toolbar Highlight
Go
Site
Rv1847
Rv1848
Rv1849
Rv1850
Rv1851
Rv1852
Rv1853
Done
Internet

Address http://210.212.212.6/cgi-bin/regsites/drawoperon.pl?genomenm=NC_000962/NC_000962.ptt&gene=Rv1847

Operon Organization

Synonym	Gene	Strand	COG	Product
Rv1847	PaaI	+	COG2050	Uncharacterized protein
Rv1848	ureA	+	COG0831	Urea amidohydrolase (urease) gamma subunit
Rv1849	ureB	+	COG0832	Urea amidohydrolase (urease) beta subunit
Rv1850	ureC	+	COG0804	Urea amidohydrolase (urease) alpha subunit
Rv1851	ureF	+	COG0830	Urease accessory protein UreF
Rv1852	ureG	+	COG0378	Ni2+-binding GTPase involved in regulation of expression and maturation of urease and hydrogenase
Rv1853	ureD	+	-	ureD

Figure 4.3: Operon organization



NCBI CDD COG2050 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?uid=COG2050
Go
Google Search Web 54 blocked AutoFill Options
Search Web Mail My Yahoo! Games Personals LAUNCH Sign In
NCBI
PubMed Nucleotide Protein Structure CDD Taxonomy Help?
Done
Internet

Conserved Domain Database

CD: COG2050.1, PaaI **PSSM-Id:** 11758 **Source:** Cog
Description: Uncharacterized protein, possibly involved in aromatic compounds catabolism [Secondary metabolites biosynthesis, transport, and catabolism]
Taxa: cellular organisms **Related:** COG1607, pfam03061
Status: Alignment from source **Created:** 7-Oct-2002
Aligned: 89 rows **PSSM:** 141 columns **Representative:** Consensus
Proteins: [Click here for CDART summary of Proteins containing COG2050]

View Alignment as Hypertext width 60 color at 2.0 bits
Subset Rows up to 10 of the most diverse members

```
                        10        20        30        40        50        60
                ....*....|....*....|....*....|....*....|....*....|....*....|
consensus     1 LERDKKVAEVNLAFL---ERSPFLKTLGI-EIEEIEEGEAEATLPVDPELLN---PGGIL 53
gi 16078651  51 RERIKNVAEKTLEDE---VKSLSLDEVIG-EIIDLELDDQAISILEIKQEHVf-sRNQIA 105
gi 16126342   8 APETLTDEAILARFRaskRQPTGSQTLGF-ELLAVRQADREVEVGFAARADLlcnPMGQI 66
gi 16127539   4 DLTDAQTAAIPEGFSqlnWSRGFGRQIGPlFEHREGPGQARLAFRVEEHHTN---GLGNC 60
gi 15792304  13 LEEYASAEDISRVRAellTCPELNTSLAG-TIIEIDKNYAKSILITTSEMVAd--DQGLI 69
gi 15672708   1 -----MNMIDQLNIT----DFQVFTDENS-DKSVSKIYKFSSKMILSDFHAQ---PHGFL 47
gi 15596790   1 -MSENPLLERARRFL---SALRHCQVLGL-TVEAADEKGLTLRLPYSQAIIGn-pESGVV 54
gi 16263492   1 ------------------MRRPPMAETLP--FTLLPPEESKVGLLATPEPRFc-nLTNTV 39
gi 16081431   1 -MFDGEDLKKIFAM------DGFLRNIEF-EVSYISEGSIEIKVPLKENLMR---VGDIM 49
gi 15612029   1 --MQESVVHVDYDSL---ETCKNFKPSVG-TELVVLEKNIAHVRFKGNENMVy-eE-NFV 52

                        70        80        90       100       110       120
                ....*....|....*....|....*....|....*....|....*....|....*....|
consensus    54 HGGVIAALADSAAGLAANSLLGVVALa-------------VTLELNINFLRPVKEGD--V 98
gi 16078651 106 RGHHLFAQANSLAVAVIDDELALTAS-----------------AD-IRFTRQVKQGEr-V 146
gi 16126342  67 QGGYVCAMLDECMSVAGMITSGMTHVv-------------pTLEMKTSFLRPAMPGP--L 111
gi 16127539  61 HGGMLMSFADMAWGRIISLQKSYSWVt---------------VRLMCDFLSGAKLGDwvE 105
gi 15792304  70 FDAFIFAAANYVAQASINKEFSVIIGskcffyaplklgdvLELEAHALFDETSKKRD--V 127
gi 15672708  48 NGGASLALAEITAGMASNAIGSSQYFa-------------LGQSISANHLNSKKCEGf-V 93
```

Figure 4.4: Conserved domain database link

Next, input a block-aligned binding sites of a regulatory protein belonging to the species of selected genome, via a web-based form (Figure 4.1). Alternatively the user can upload a file containing the block-aligned binding sites by clicking on the ‘Browse’ button. The binding sites of a regulatory protein can also be used to identify its orthologue binding sites in phylogentically related genomes where the trans-acting regulatory proteins are conserved.

Further, input the upstream and downstream sequence length to scan for the binding sites. This length refers to the relative position to the translation start site.

On submit button press the parameters are sent to server, the input binding are is used to construct a profile based model of binding site using the Shannon relative entropy and the model is then used to scan the upstream sequences of all the genes of user selected genome. Finally the results will be shown as a web page. Figure 4.2 shows the output of IdeR binding sites in *M. tuberculosis.* The site column in Figure 4.2 represents the binding sites of IdeR in *M. tuberculosis*. In a typical output the perfect match to the known binding sites and the downstream genes are highlighted with a yellow background, and the rest with score greater than cut-off is shown with a light blue background.

The web output of Predictregulon also contains the hyperlinked gene-synonym and COG number. A click on the former shows the predicted operon context of the regulatory motif (Figure 4.3) while a click on the latter opens a new page showing a description of this gene in the NCBI Conserved Domain Database (Figure 4.4), which is in turn linked to Pubmed for published information on this gene.

These additional links provides users a simple way to browse and understand the functional/physiological implication of the genes that are part of predicted regulon.

4.3 Conclusion

The Predictregulon system integrates two different computational approaches operon prediction and regulatory element prediction to identify the regulons in prokaryote genomes. For the end user, Predictregulon alleviates the need for an expensive computer setup and familiarity with computer programming. With a robust engine written in PERL and C^{++}, the system is user-friendly, with simple menus and easy to understand results.

Chapter 5

Prediction of DtxR Regulon in *C. diphtheriae*

Iron is an important inorganic component of a cell. Iron is required as co-factor for various essential enzymes and proteins some of which are involved in electron transport (Cytochromes), redox reactions (oxidoreductases) and regulation of gene expression (fumarate-nitrate reduction regulatory protein, iron-binding protein) (Castagnetto *et al*., 2002). However a higher level of intracellular iron can catalyze formation of hydroxyl radicals and reactive oxygen species through Fenton's reaction, which could be lethal to the cell (Urbanski *et al*., 2000). Hence, a careful regulation of iron-requiring enzymes/proteins and iron uptake proteins/enzymes is required for the survival of bacteria.

Inorganic iron is also known to influence virulence in many pathogenic bacteria such as *C. diphtheriae, E. coli*, and *Bordetella bronchiseptica* (Tao *et al*., 1994; Russo *et al*., 2001; Register *et al*., 2001).

The diphtheria toxin repressor DtxR is known as an iron-activated global transcription regulator that represses the transcription of various iron- dependent genes in *C. diphtheriae* (Qian *et al*., 2002; Kunkle *et al*., 2003). Eight DtxR-binding sites in upstream sequences of operons/genes named as *tox, hmuO, irp1, irp2, irp3, irp4, irp5 and irp6* have been identified by DNA footprinting methods (Table 5.1). The product of *tox* gene is diphtheria toxin, which catalyzes the NAD-dependent ADP ribosylation of eukaryotic aminoacyl-transferase-II, thereby causing inhibition of protein synthesis and subsequent death of the host. The *hmuO* gene, which encodes a haem oxygenase, oxidizes the haem to release free iron. The operons *irp1* and *irp6* encode the products with homology to ABC-type ferric-siderophore transport systems. The gene *irp3* encodes a homologue of AraC-type transcriptional activators. The products of *irp2, irp4 and irp5* do not show any homology to the other known proteins. In addition, *C. diphtheriae* with inactive DtxR has been shown to be sensitive to killing by exposure to high iron conditions or hydrogen peroxide than the wild type (Oram *et al*., 2002).

Table 5.1. Known DtxR-binding sites from *C. diptheriae*

Binding site	Gene	Product
TTAGGATAGCTTTACCTAA	*tox*	Diphtheria toxin
TTAGGTTAGCCAAACCTTT	*Irp1*	Periplasmic protein of siderophore transport system
GCAGGGTAGCCTAACCTAA	*Irp2*	Hypothetical protein
TTAGGTGAGACGCACCCAT	*Irp3*	AraC-type transcription regulator
ATTACTAACGCTAACCTAA	*Irp4*	Hypothetical protein
CTAGGATTGCCTACACTTA	*Irp5*	Hypothetical protein
TTTCCTTTGCCTAGCCTAA	*Irp6*	Periplasmic protein of siderophore transport system
TGAGGGGAACCTAACCTAA	*hmuO*	Haem oxygenase

This work uses an *in silico* method to identify additional DtxR-binding sites and target genes to understand the role of DtxR in virulence and patho-physiology of *C. diphtheriae*.

5.1 Methods

The complete genome sequence of *C. diphtheriae* was downloaded from NCBI ftp site (ftp.ncbi.nih.gov/genomes/Bacteria/Corynebacterium_diphtheriae), and the DtxR-binding sites identified by experimental methods were collected from literature (Qian *et al*., 2002). DtxR binding sites and target operons were predicted by the method mentioned in chapter three and two.

5.1.1 Functional assignment of genes

The function of predicted genes was inferred using the RPS-BLAST search against conserved domain database (Marchler-Bauer *et al*., 2003). These genes were further classified according to their function.

5.1.2. Expression and purification of IdeR

The iron-dependent regulator IdeR from M. *tuberculosis* was expressed from a recombinant pRSET vector containing the IdeR gene fused to a six His affinity tag. The expressed protein was first purified using Ni-NTA Metal Chelate Affinity chromatography; later it was desalted and concentrated using Centricon Ultra filtration device. The concentration of the recombinant protein was estimated using Bradford method.

5.1.3 Electrophoretic mobility shift assay

Double-stranded oligonucleotides containing the predicted binding motif (19 bp long) were end labeled with T4 polynucleotide kinase and [γ^{32}P]-ATP and were incubated with the recombinant purified IdeR protein in a binding reaction mixture. The binding reaction mixture (20-µl total volume) contain the DNA-binding buffer (20 mM Tris-HCl [pH 8.0], 2 mM DTT, 50mM NaCl, 5mM $MgCl_2$, 50 % glycerol, 5 µg of bovine serum albumin per ml), 10 µg of poly (dI-dC) per ml (for nonspecific binding) and 200µM. $MnCl_2$. The reaction mixture was incubated at room temperature for 30 min. Approximately 2 µl of the tracking dye (50% sucrose, 0.6% bromophenol blue) was added to the reaction mixture at the end of incubation and was loaded onto 7% polyacrylamide gel containing 150µM $MnCl_2$ in 1× Tris-borate-EDTA buffer. The gel was electrophoresed at 200 V for 2 hours. Subsequently the gel was dried and exposed to Fuji Storage Phosphor Image Plates for 16 hours. The image plates were subsequently scanned in Fuji Storage Phosphor Imaging workstation.

5.2 Results

5.2.1 *In silico* identification of putative DtxR-binding sites

Experimentally characterized DtxR-binding motifs were collected from the literature (Table 5.1) (Qian *et al*., 2002). These binding sites were used to identify additional putative DtxR-binding sites along with associated operons in C. *diphtheriae* NCTC13129 genome (see materials and methods). Table 5.2 shows the predicted DtxR-binding sites with score 3.7438 or more. I could identify five (tox, irp4, irp5, irp6 and hmuO) of the eight known DtxR-binding sites, in sequenced *C. diphtheriae* NCTC13129 genome. I could not find irp1 and irp2 motifs, as the corresponding genes (*irp1, irp2*) are not present in the sequenced strain, NCTC13129 (Cerdeno-Tarraga *et al*., 2003). The regulator binding sites of *irp3*, *irp4* and *irp6* genes in the strain NCTC13129 shows one base change from the binding sites reported in strain C7 (Qian *et al*., 2002). Binding site of *irp3* gene (TTAGGTGAGACGCACCCAT) although exists in strain NCTC13129, but

Table 5.2. Predicted DtxR-binding sites in *C. diphtheriae*

Score	Position	Site	Gene	Synonym	Product
4.45904	-80	TGAGGGGAACCTAACCTAA	*hmuO*	DIP1669**	Heme Oxygenase
4.39003	-52	TTAGGATAGCTTTACCTAA	*Tox*	DIP0222**	Diphtheria Toxin Precursor
4.25877	-60	ATAGGCTACACTTACCTAA	-	DIP0624	Putative Membrane Protein
4.21068	-168	TTGGATTAGCCTACCCTAA	-	DIP2162**	ABC-Type Peptide Transport System Periplasmic Component
4.2033	-21	*TTAGGGTAGCTTCGCCTAA*	*iucA*	DIP0586	Putative Siderophore Biosynthesis Related Protein
4.17632	-78	ATAGGCATGCCTAACCTCA	-	DIP2330	Putative Membrane Protein
4.07921	-130	TTAGGTCAGGGTACCCTAA	-	DIP0370	Putative Succinate Dehydrogenease Cytochrome B Subunit
4.03559	-30	TTAGCTTAACCTTGCCTAT	*arsR*	DIP0415	Putative Arsr Family Regulatory Protein
4.01967	-239	*TTAGGGTAGGCTAATCCAA*	*sidA**	DIP2161	Nonribosomal Peptide Synthase
3.99985	-74	TTTTCTTTGCCTAGCCTAA	*irp6A*	DIP0108**	Ferrisiderophore Receptor Irp6A
3.99195	-241	TTAGGCACCCCTAACCTAG	-	DIP0539	Putative Sugar ABC Transport Syste ATP-Binding Protein
3.98554	-72	TTAGCTTAGCCCTAGCTAA	-	DIP0169	Putative Secreted Protein
3.9296	-26	CTAGGATTGCCTACACTTA	*Irp5*	DIP0894**	Conserved Hypothetical Protein
3.9073	-93	GTTGGGTTGCCCAACCTAC	-	DIP2106	Putative ABC Transport System, ATP-Binding Subunit
3.89763	-86	ATAGGTTAGGTTAACCTTG	*chtA**	DIP1520	Putative Membrane Protein
3.89676	-130	*TTGTGTTAGCCTAGGCTAA*	*secA*	DIP0699	Translocase Protein
3.89169	-26	*TTGGGGTGGCCTATCCTTA*	-	DIP2304	Putative DNA-Repair Glycosylase
3.88042	-172	TTAGGTAAGTGTAGCCTAT	*htaA**	DIP0625	Putative Membrane Protein
3.86534	-69	ATTACTAATGCTAACCTAA	*Irp4*	DIP2356**	Putative Conserved Membrane Protein
3.85539	-173	TTAGGGTGGGCTAACCTGC	*deoR**	DIP1296	Putative DNA-Binding Protein
3.84889	-75	TTAGGGAACTCTTGCCTTA	*piuB**	DIP0124	Putative Membrane Protein
3.83816	-121	TTAGCTAGGGCTAAGCTAA	-	DIP0168	Putative Glycosyl Transferase
3.83576	-219	GTAACAAAGGCAAGCCTAA	*xerD*	DIP1510	Putative Integrase/Recombinase
3.8224	-216	ATAGGCAAGGTTAAGCTAA	-	DIP0417	Putative Membrane Protein
3.81905	-47	GTTGGACAGGTTACCCTAA	*frgA**	DIP1061	Putative Iron-Siderophore Uptake System Permease
3.8148	-37	*TGTGGGCACACCAACCTAA*	-	DIP2272	Possible Sortase-Like Protein
3.76235	-136	TTGGGGTTGCCCTTCCTAA	-	DIP0142	Hypothetical Protein
3.76233	-268	CTAGGTTAGGGGTGCCTAA	*secY**	DIP0540	Preprotein Translocase Secy Subunit
3.74673	-110	TAAACATAGCCAAACCAAA	*nrdF1*	DIP1865	Ribonucleotide Reductase Beta-Chain 1
3.7438	-81	TAAGGATAGGCCACCCCAA	*Dps*	DIP2303	Starvation Inducible DNA-Binding Protein

Note: **Indicate the gene synonym with experimentally identified binding site in *C. diphtheriae* [6]. * Indicates the genes known to be regulated by DtxR [7]. The binding sites in Italics were verified by EMSA. The gene pairs, DIP0624-DIP0625, DIP2161-DIP2162, DIP0168-DIP0169, DIP0539-DIP0540 and DIP2303-DIP2304 are divergently transcribed and contain common regulatory regions.

not there in the predicted sites, because it is located within the coding region of *irp3* ORF. The predicted ORF of *irp3* in the sequenced strain NCTC13129 has different start position and is larger than what was previously reported in strain C7 (Cerdeno-Tarraga *et al*., 2003; Lee *et al*., 1997).

In addition, binding sites in upstream sequences of eight genes that are reported to be regulated by DtxR were identified (Kunkle *et al*., 2003). However, our prediction differs from the previous report for five (secY, deoR, chtA, frgA, sidA) of the seven sites which were identified by BLAST search (Table 5.2). Our prediction agreed with the previous report that the genes such as *recA* (DIP1450) and *ywjA* (DIP1735) are not under a direct DtxR regulation, as I could not detect any motif upstream to these genes with scores above the cutoff value (Kunkle *et al*., 2003).

5.2.2 Experimental validation of predicted binding sites

Since our approach to identify DtxR-regulated genes is purely computational in nature, I decided to test the validity of our predictions. A sample of predicted regulator binding motifs (Table 5.2) (upstream to ORFs: DIP2161, DIP0699, DIP0586, DIP2304, DIP2272) were experimentally verified by EMSA using IdeR, an orthologue of DtxR from *M. tuberculosis.* DtxR and IdeR are iron-dependent regulators. A pair wise sequence comparison of the two proteins shows a high (58%) overall sequence identity (similarity 72%), which increases further to 92% identity and 100% similarity in DNA recognition domain. In addition, the structural comparison of two regulators also shows a very similar 3D organization, suggesting that the IdeR regulator would be able to recognize the DtxR motif (Feese 2001).

Synthetic double stranded oligonucleotides corresponding to DNA-binding sites were labeled with ^{32}P and mixed with purified IdeR in presence of manganese ions and was assayed for the formation of DNA-protein complex using EMSA. Manganese was used as the divalent metal in the binding reactions on account of its redox stability compared with ferrous ion. Electrophoretic mobility of all five double stranded oligonucleotides has been tested was retarded by IdeR (Figure 5.1). However a synthetic motif (TTTTCATGACGTCTTCTAA) used as a negative control did not show any complex formation. These results indicate that the predicted DtxR-binding sites can indeed bind to DtxR.

5.2.3 Identification and annotation of DtxR-regulated genes

In addition to the binding site prediction, I have also identified co-regulated genes (operons) downstream to the predicted DtxR-binding site (Table 5.3). Function of the proteins encoded by the putative genes in Table 5.2 and Table 5.3 was predicted by RPS-BLAST search against conserved domain database (Marchler-Bauer *et al*., 2003).

5.3 Discussion

Our analysis identified putative DtxR motifs upstream to various operons/genes which could be involved in siderophore biosynthesis, ABC-type transport systems, iron storage, oxidative stress defense and iron-sulfur cluster biosynthesis. In addition, I have also identified the motifs upstream of operons that could be involved in anchoring of host-interacting proteins to the cell wall and secretion of various virulence factors. Important functions of some of these DtxR-regulated genes and their role in *C. diphtheriae* physiology are discussed here.

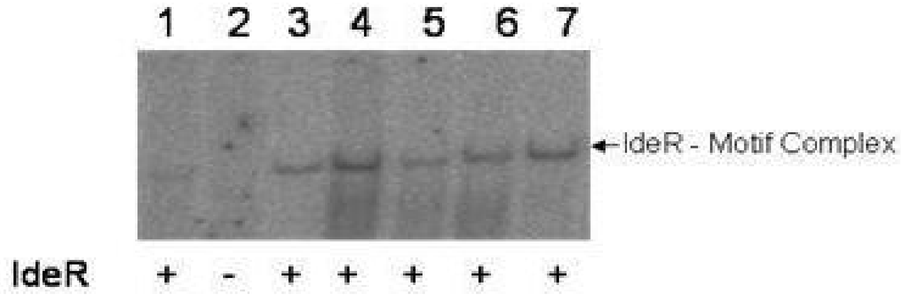


Figure 5.1: IdeR binds to the predicted DtxR-binding DNA fragments

30 pmoles of IdeR was added to ^{32}P-labelled DNA probes in the presence of 200 μM Mn^{2+}, and complexes were resolved on a 7% Tris-borate polyacrylamide gel containing 150 μM Mn^{2+}; Lane 1: Control gel retardation using Radiolabeled DNA motif without DtxR-binding site. Lane 2: Radiolabeled DIP2161 motif without IdeR. Lane 3: Radiolabeled DIP2161 motif with IdeR. Lane4: Radiolabeled DIP0699 motif with IdeR. Lane 5: Radiolabeled DIP0586 motif with IdeR. Lane 6: Radiolabeled DIP2304 motif with IdeR. Lane 7: Radiolabeled DIP2272 motif with IdeR

Table 5.3. Predicted DtxR-regulated operons in *C. diphtheriae*

Synonym	Gene	Orthologue	Product
DIP2158		COG1131	ABC-type transport system permease and ATPase component
DIP2159		COG1131	ABC- type transport system permease and ATPase component
DIP2160	-	COG3321	Polyketide synthase modules and related proteins
DIP2161*	-	COG1020	Non-ribosomal peptide synthetase modules and related proteins
DIP0586	*iucA*	Pfam04183	Catalyse discrete steps in biosynthesis of the siderophore aerobactin
DIP0587	-	-	Putative membrane protein
DIP0588	-	-	Putative membrane protein
DIP1059	*fepC*	COG1120	ABC-type cobalamin/Fe3+-siderophores transport systems
DIP1060	*fepG*	COG4779	ABC-type enterobactin transport system
DIP1061*	*fepD*	COG0609	ABC-type Fe3+-siderophore transport system
DIP2162	*ddpA*	COG0747	ABC-type peptide transport system periplasmic component
DIP2163	*ddpB*	COG0601	ABC-type peptide/nickel transport systems permease components
DIP2164	*ddpC*	COG1173	ABC-type peptide/nickel transport systems permease components
DIP2165	*dpdD*	COG0444	ABC-type peptide/nickel transport systems ATPase component
DIP0169	*lraI*	COG0803	ABC-type metal ion transport system, periplasmic component
DIP0170	*znuC*	COG1121	ABC-type Mn/Zn transport systems, ATPase component
DIP0171	*znuB*	COG1108	ABC-type Mn2+/Zn2+ transport systems, permease components
DIP0172	*znuB*	COG1108	ABC-type Mn2+/Zn2+ transport systems, permease components
DIP0173	*lraI*	COG0803	ABC-type metal ion transport system, periplasmic component
DIP2106	*mdlB*	COG1131	ABC-type multidrug transport system, ATPase and permease component
DIP2107	*mdlB*	COG1131	ABC-type multidrug transport system, ATPase and permease component
DIP0625	*htaa*	Pfam04213	Haemin transporter associated protein
DIP0626	*hmuT*	COG4558	ABC-type haemin transport system
DIP0627	*hmuU*	COG0609	ABC-type Fe3+-siderophore transport system
DIP0628	*hmuV*	COG4559	ABC-type haemin transport system
DIP0629*	*htaa*	Pfam04213	Haemin transporter associated protein
DIP1519*	*htaa*	pfam04213	Haemin transporter associated protein
DIP1520*	*htaa*	pfam04213	Haemin transporter associated protein
DIP2303	*dps*	COG0783	Starvation inducible DNA-binding protein
DIP2304	-	COG0266	Formamidopyrimidine-DNA glycosylase
DIP2305	-	COG0063	Predicted sugar kinase
DIP1510	*xerD*	COG4974	Site-specific recombinase
DIP1288	-	-	Conserved hypothetical protein
DIP1289	*uup*	COG0488	ATPase components of ABC transporters with duplicated ATPase domains
DIP1290	-	COG2151	Predicted metal-sulfur cluster biosynthetic enzyme
DIP1291	*iscU*	COG0822	NifU homolog involved in Fe-S cluster formation
DIP1292	*csd*	COG0520	Selenocysteine lyase
DIP1293	*sufC*	COG0396	ABC-type transport system involved in Fe-S cluster assembly
DIP1294	-	COG0719	ABC-type transport system involved in Fe-S cluster assembly
DIP1295	*sufB*	COG0719	ABC-type transport system involved in Fe-S cluster assembly
DIP1296*	*deoR*	COG2345	DeoR family transcriptional regulator
DIP0370	-	-	Putative succinate dehydrogenease (cytochrome b)
DIP0371	-	COG1053	Succinate dehydrogenase/fumarate reductase
DIP0372	-	COG0479	Succinate dehydrogenase/fumarate reductase
DIP0373	-	-	Putative membrane protein
DIP0374	-	-	Putative membrane protein
DIP0375	-	-	Putative membrane protein
DIP0376	-	-	Putative membrane protein
DIP0377	-	-	Putative membrane protein

Table 5.3. Contnd.

Synonym	Gene	Orthologue	Product
DIP1864	*ctaD*	COG0843	Heme/copper-type cytochrome/quinol oxidases
DIP1865	*nrdF1*	COG0208	Ribonucleotide reductase
DIP2330	-	-	Putative membrane protein
DIP2331	-	COG1012	NAD-dependent aldehyde dehydrogenases
DIP0124*	-	Pfam03929	Uncharacterized iron-regulated membrane protein (DUF337)
DIP0622	-	-	Putative membrane protein
DIP0623	*metA*	COG2021	Homoserine acetyltransferase
DIP0624	-	-	Putative membrane protein
DIP0415	-	Pfam01022	Bacterial regulatory protein
DIP0539	-	COG3839	ABC-type sugar transport systems
DIP0168	-	-	Putative glycosyl transferase
DIP0417	-	-	Putative membrane protein
DIP0142	-	-	Hypothetical protein
DIP0143	-	-	-
DIP0144	*tra8*	COG2826	Transposase and inactivated derivatives
DIP2271	-	-	Putative membrane protein
DIP2272	-	COG3764	Sortase (surface protein transpeptidase)
DIP0699	*secA*	COG0653	Preprotein translocase subunit SecA (ATPase
DIP0700	-	-	Hypothetical protein
DIP0540*	*secY*	Pfam00344	Eubacterial secY protein
DIP0541	*Adk*	COG0563	Adenylate kinase and related kinases
DIP0542	*mapA*		Methionine aminopeptidase
DIP0543	-	-	Sialidases or neuraminidases;
DIP0544	*erfK*	Pfam03734	This family of proteins contains a conserved histidine and cysteine
DIP0545	*infA*	COG0361	Translation initiation factor 1 (IF-1)
DIP0546	*rpsM*	COG0099	Ribosomal protein S13
DIP0547	*rpsK*	COG0100	Ribosomal protein S11
DIP0548	*rpsD*	COG0522	Ribosomal protein S4 and related proteins
DIP0549	*rpoA*	COG0202	DNA-directed RNA polymerase
DIP0550	*rplQ*	COG0203	Ribosomal protein L17
DIP0551	*truA*	COG0101	Pseudouridylate synthase

Note: * Indicate the genes reported be regulated by DtxR. Genes listed together belongs to same operon.

5.3.1 Regulation of siderophore biosynthesis and ABC- type transport systems

Predicted member of the DtxR regulon, the gene DIP0586, codes for the IucA/IucC family of enzymes that catalyze discrete step in the biosynthesis of the aerobactin (de Lorenzo and Neilands, 1986). In addition to known DtxR-regulated siderophore transport genes (irp1, irp6), DtxR could also regulate other ABC-type transport systems similar to Manganese/Zinc, peptide/Nickel and multidrug subfamilies of ABC transporters. The peptide/nickel transport system (DIP2162-DIP2165) suggested to be recently acquired by pathogenic *C. diphtheriae* (Cerdeno-Tarraga *et al*., 2003).

5.3.2 Regulation of iron storage and oxidative stress defense

I predict that DtxR could regulate divergently transcribed genes DIP2303 and DIP2304 whose products are similar to starvation inducible DNA-binding protein (Dps) and Formamidopyrimidine-DNA glycosylase (Fpg), respectively. Dps in *Escherichia coli* is induced in response to oxidative or nutritional stress and protects DNA from oxidative stress damage by nonspecific binding (Martinez *et al*., 1997). Dps also catalyzes oxidation of ferrous iron to ferric iron by hydrogen peroxide ($2Fe^{2+} + H_2O_2 + 2H_2O \rightarrow 2Fe^{+3}OOH_{(core)} + 4H^{+}$), which in turn prevents hydroxyl radical formation by Fenton's reaction ($Fe^{2+} + H_2O_2 \rightarrow Fe^{+3} + HO^{-} + HO^{\cdot}$) and thereby prevents subsequent DNA damage (Zhao *et al*., 2002)). The enzyme, formamidopyrimidine-DNA glycosylase is a primary participant in the repair of 8-oxoguanine, an abundant oxidative DNA lesion (Zaika *et al*., 2004). The gene DIP1510, which codes for the site-specific recombinase XerD could also be regulated by DtxR. The *xerD* gene in *E. coli* belongs to the oxidative stress regulon (Gaudu and Weiss, 2000).

5.3.3 Regulation of proteins involved in iron-sulfur cluster biosynthesis

The prediction shows that DtxR could regulate the operon DIP1288-DIP1296, which is similar to the suf operon of E.coli. The *suf* operon in bacteria encodes the genes for Fe-S cluster assembly machinery (Outten *et al*., 2003). In addition, genes encoding the iron-sulfur containing proteins such as succinate dehydrogenase (Sdh), cytochrome oxidase (CtaD) and Ribonucleotide reductase (NrdF1) in *C. diphtheriae* also show DtxR motif in their upstream sequences.

5.3.4 Regulation of sortases

The prediction shows that DtxR could regulate the recently acquired pathogenic island DIP2271-DIP2272, encoding the sortase srtA and hypothetical protein, respectively (Cerdeno-Tarraga *et al*., 2003). Sortases are membrane-bound trans-peptidases that catalyze the anchoring of surface proteins to the cell wall peptidoglycan (Cerdeno-Tarraga *et al*., 2003). Such systems are often used by gram-positive pathogens to anchor host-interacting proteins to the bacterial surface (Ton-That *et al*., 2003).

5.3.5 Regulation of protein translation and translocation system

DtxR could regulate two operons that contain genes DIP0699 (*secA*) and DIP0540 (*secY*) that code for the protein translocation system. The *sec*Y-containing operon, which is similar to the streptomycine operon spc from *B. subtilis* and other bacteria, encodes the genes required for protein translation and translocation (Suh *et al*., 1996). The operon contains additional sialidase gene (DIP0543) in comparison to non-pathogenic Corynebacterium species. Activity of sialidase has been linked to virulence in several other microbial pathogens and may enhance fimbriae mediated adhesion in *C. diphtheriae* by unmasking receptors on mammalian cells (Cerdeno-Tarraga *et al*., 2003).

The Sec system can both translocate proteins across the cytoplasmic membrane and insert integral membrane proteins into it. The former proteins but not the latter possess N-terminal, cleavable, targeting signal sequences that are required to direct the proteins to the Sec system. Some of the DtxR-regulated genes including diphtheria toxin (Table 5.4) show predicted signal sequences by SignalP 3.0 (Jannick *et al*., 2004)

Table 5.4: DtxR-regulated genes containing the potential signal sequence

Gene	Product
DIP0222	Diphtheria toxin
DIP0109	IRP6B
DIP2356	IRP4
DIP2162	ABC-type peptide transport system periplasmic component
DIP0172	Putative membrane protein
DIP2107	Putative integral membrane transport protein
DIP0625	Haemin transporter associated protein
DIP0626	ABC-type haemin transport system
DIP0627	ABC-type haemin transport system
DIP1519	Haemin transporter associated protein
DIP0629	Haemin transporter associated protein
DIP1520	Haemin transporter associated protein
DIP2330	Putative membrane protein
DIP0543	Sialidases or neuraminidases

and hence they may play an important role in host interaction and virulence of *C. diphtheriae* (Cerdeno-Tarraga *et al*., 2003).

5.4 Conclusions

The bioinformatics method used to predict the targets of DtxR in C. *diphtheriae* NCTC13129 genome is promising, as some of the predicted targets were experimentally verified. The approach identified novel DtxR-regulated genes, which could play an important role in physiology of *C. diphtheriae* NCTC13129. DtxR, generally known as a repressor of diphtheriae toxin and iron siderophore/transport genes, can also regulate other metal ion transport genes, iron storage, oxidative stress, DNA-repair, biosynthesis of iron-sulfur cluster, Fe-S-cluster containing proteins, and even protein sortase and translocation systems.

Chapter 6

Prediction of DtxR Regulon in *C. glutamicum*

This study aims to identify the DtxR regulated genes and their role in cellular physiology of *C. glutamicum* in comparison to pathogenic *C. diptheriae*. The 'Predictregulon' method was applied to identify the genes that are controlled by regulatory protein-DtxR. Reported DtxR binding sites from *C. diphtheriae* (Table 5.1) were used to generate a recognition profile based on Shannon relative entropy, which was used to predict potential DtxR sites in the genome of *C. glutamicum*. A sample of predicted motif was experimentally verified using recombinant IdeR (Iron dependent Regulator), an orthrolog of DtxR from *M. tuberculosis* - using EMSA. Since the transcription of the genes in prokaryotes can occur as an operon, I have also predicted the other co-expressed genes that are potentially part of DtxR regulated operons.

The study identifies DtxR regulated operons/genes, which code for proteins involved in iron release and uptake systems, such as hemolysins, hemin transport system, and ferric-siderophore transport system. The analysis also predicted few other DtxR regulated genes, whose products are orthologs of ferritin and starvation inducible DNA binding protein (Dps). These proteins are involved in iron storage and oxidative stress defense in many other bacteria.

In addition, the genes that code for the orthologs of adaptive response regulator (Ada) and endonuclease VIII (Nei) involved in DNA repair could also be regulated by DtxR. Analysis of DtxR regulated genes shows that DtxR has an important role in iron acquisition, uptake, iron storage, oxidative stress defense and DNA repair.

6.1 Method

6.1.2 Source of genome sequence

The complete genome sequence of *C. glutamicum* was downloaded from NCBI (ftp.ncbi.nih.gov/genomes/Bacteria/Corynebeacterium_glutamicum) and the DtxR binding sites identified by experimental methods were collected from literature (Table

5.1). DtxR biniding sites and target operons were predicted using the method described in chapter three and two.

6.1.3 Expression and purification of IdeR

IdeR Protein was expressed from a recombinant pRSET vector containing the IdeR gene fused to a six His Affinity tag. The expressed protein was first purified using Ni-NTA Metal Chelate Affinity chromatography; later it was desalted and concentrated using Centricon Ultra filtration device. The concentration of the recombinant protein was estimated using Bradford method.

6.1.4 Electrophoretic Mobility Shift Assay

Double-stranded oligonucleotides containing the binding motif (19 bp long) were incubated with the recombinant IdeR protein in a binding reaction mixture. The binding reaction mixture (20-µl total volume) contains the DNA binding buffer (20 mM Tris-HCl [pH 8.0], 1 mM DTT, 5mM $MgCl_2$, 10% glycerol, 50 µg of bovine serum albumin per ml), 50 µg of poly (dI-dC) per ml (for nonspecific binding), 200 µM Ni^{2+} as substitute for Fe^{2+} ion. The reaction mixture was incubated at room temperature for 30 min. Approximately 2 µl of the tracking dye (50% sucrose, 0.6% bromophenol blue) was added to the reaction mixture at the end of incubation and was loaded onto 7% polyacrylamide gel in 1× Tris-borate-EDTA buffer. The gel was electrophoresed at 200 V for 2 hours. Subsequently the gel was dried and exposed to Fuji Storage Phosphor Image Plates for 16 hours. The image plates were subsequently scanned in Fuji Storage Phosphor Imaging workstation.

6.2 Results

6.2.1 *In-silico* identification of potential DtxR binding sites

A recognition profile of eight known DtxR binding sites from *C. diphtheriae* was used to identify the potential DtxR binding sites and downstream operons/genes in *C. glutamicum* genome. Table 6.1 lists the predicted DtxR binding sites and Table 6.2 lists the predicted operons/genes downstream to the predicted DtxR binding sites.

6.2.2 Experimental verification of predicted DtxR binding sites

A sample of predicted motif (upstream to ORF NCgl0123, NCgl0377, NCgl0381, NCgl1394 and NCgl2439) was experimentally verified using EMSA. For this, I have used IdeR, an orthrolog of DtxR from *M. tuberculosis.*

The 5' ^{32}P phosphate labeled DtxR binding sites (19 bps) regulating the *tox* gene as well as two of the predicted sites, upstream to the ORF numbers (1394, 0639) showed an IdeR concentration dependent EMS in the assay (Figure 6.1). The EMS was abolished when the cold ds-oligo representing the binding motif was used as cold competitor. Figure 6.1 also showed that IdeR can show EMS with Tox motif as well as predicted motif with very low protein concentration (20 picomoles).

In order to test other remaining motifs simultaneously, similar assay was done with *C. diphtheria* Tox as radio labeled ds-oligo and other ds-oligos, representing the predicted sites, upstream to the ORF numbers (0381, 0123, 1394, 2439, 0377), as cold competitor. Increasing concentration of cold competitor was used, which resulted in a concentration dependent inhibition of EMS of Tox regulatory motif of *C. diphtheria* (Figure 6.2). These experiments demonstrate that DtxR homologue IdeR can indeed bind to these motifs and in turn are likely to regulate the downstream genes.

Table 6.1. Predicted DtxR binding sites in *Corynebacterium glutamicum*

Score	Position	Site	Gene	Product
4.38738	-59	***GTCGGGCAGCCTAACCTAA***	NCgl0639	ABC-type transporter, periplasmic component
4.24182	-116	***TATGGCTTGCCTAACCTAA***	NCgl1394	Ptative helicase
4.18776	-110	TTAGTAAAGGCTCACCTAA	NCgl0484	ABC-type transporter, permease component
4.12834	-267	TTAGGTGAGCCTTTACTAA	NCgl0485	Aetyl-CoA hydrolase
4.09952	-178	CACGGTGAACCTAACCTAA	NCgl2718	Ptative nitrite reductase
4.09864	-52	TGAGGTTAGCGTAACCTAC	NCgl0943	AraC-type DNA-binding domain-containing protein
4.08744	-84	TTTAGGTAACCTAACCTCA	NCgl0776	ABC-type Fe3+-siderophore transport system, periplasmic component
4.0873	-1	***AATGGTTAGGCTAACCTTA***	NCgl0123	Hpothetical protein
4.08124	-30	TTAGGCTTGCCATACCTAT	NCgl0430	Pedicted arsR family transcriptional regulator
4.05993	-139	GTAGGTGTGGGTAACCTAA	NCgl2146	Haem oxygenase
4.05748	-45	ATAGGATAGGTTAACCTGA	NCgl0618	ABC-type Fe3+-siderophores transport system, periplasmic component
4.05676	-174	AAAAGGTAGCCTTGCCTAA	NCgl1958	Sgnal peptidase I
4.05475	-131	TAAAGTAAGGCTATCCTAA	NCgl0359	Hpothetical membrane protein
4.03484	-161	***TTAAGTTAGCATAGCCTTA***	NCgl0377	Haemin transport system associated protein
3.99875	-132	ATAACGCACCCTAACCTTA	NCgl2902	NADPH:quinone reductase
3.99813	-210	TTAACTTTGCCCTACCTAA	NCgl2766	Hpothetical membrane protein
3.98788	-89	GCACGATGGCCAAACCTAA	NCgl0903	Pedicted lactoylglutathione lyase
3.96268	-52	***TTAGGTTAAGCTAATCTAG***	NCgl0381	Haemin transport system associated protein
3.96227	-65	CTACTGTGCCCTAACCTAA	NCgl1949	Translation elongation factor Ts
3.95735	-80	TCAGGATAGGACAACCTAA	NCgl2897	Sarvation-inducible DNA-binding protein
3.93746	-48	TAAGGATAACCTTGCCTTA	NCgl0329	ABC-type Fe3+-citrate transport, periplasmic component
3.93563	-85	TTAGGTTGTCCTATCCTGA	NCgl2898	Frmamidopyrimidine-DNA glycosylase
3.92855	-194	TTAGGTAAAGCTTGCCTAT	NCgl1646	Hpothetical protein
3.88848	-102	TTAAGTCAGTGTTACCTAA	NCgl0914	ABC-type multidrug transporter
3.88598	-154	AGAAGTAAAACTTACCTAA	NCgl2990	Gucose-inhibited division protein B
3.87257	-25	GCTCAATAACCTAACCTAA	NCgl2729	ABC-type transporter, permease component
3.85588	-184	TTGCATTAGGCTATCCTAA	NCgl2971	Ptative oxidoreductase/dehydrogenase
3.85111	-57	***TTATGCTGCGCTAACCTAT***	NCgl2439	Frritin-like protein
3.84519	-240	TTAGGATTCTCTCAACTAA	NCgl1703	Ste-specific DNA methylase or
3.83489	-247	TTAACCAAGCCAAACCTTT	NCgl0775	Hypothetical membrane protein
3.80333	-66	TCAAAGTAGCCTCAACTAA	NCgl0851	Pedicted membrane protein
3.79838	-59	TTAGGTTAGGCAAGCCATA	NCgl1395	Sderophore-interacting protein
3.78619	-18	AGAGGGCACACTACCCTAT	NCgl1615	Hpothetical protein
3.77492	-18	TTGCGTTAGGATAGCCTAA	NCgl2970	ABC-type transport systems, periplasmic component
3.76956	-165	CTAGGACACTGGAACCTAA	NCgl2412	Hpothetical membrane protein
3.76689	-49	TAAGGTTTGCCTAATCTTT	NCgl0774	ABC-type Fe3+-siderophore transport system, periplasmic component

Note: The second column shows the position of binding site relative to the translation start site. The binding sites with bold was experimentally verified by electrophoretic mobility shift assay.

Table 6.2. Predicted DtxR regulated operons in *Corynebacterium glutamicum*

Gene	COG No.	Product
NCgl0639	COG0614	ABC type transporter, periplasmic component
NCgl0638	COG0609	ABC type transporter, permease component
NCgl0637	COG0609	ABC type transporter, permease component
NCgl0636	COG1120	ABC type transporter, ATPase component
NCgl0635	COG2375	Siderophore interacting protein
NCgl0634	COG2838	Monomeric isocitrate dehydrogenase
NCgl0633	-	Hypothetical membrane protein
NCgl1394	COG0513	Putative helicase
NCgl1393	COG1253	Hemolysin containing CBS domain
NCgl1392	COG1253	Hemolysin containing CBS domain
NCgl1391	-	Hypothetical protein
NCgl0484	COG0609	ABC type transporter, permease component
NCgl0483	COG4779	ABC type transporter, permease component
NCgl0482	COG1120	ABC type transporter, ATPase component
NCgl0485	COG0427	Acetyl CoA hydrolase
NCgl2718	COG0155	Putative nitrite reductase
NCgl0943	COG2207	AraC type DNA binding domain containing protein
NCgl0944	COG4760	Hypothetical membrane protein
NCgl0776	COG4607	ABC type cobalamin/Fe3+ siderophore transport system, periplasmic component
NCgl0123	-	Hypothetical protein
NCgl0122	-	Hypothetical protein
NCgl0121	COG0477	Permease of the major facilitator superfamily
NCgl0120	COG1940	Transcriptional regulator
NCgl0430	COG0640	Predicted arsR family transcriptional regulator
NCgl2146	COG5398	Haem oxygenase
NCgl0618	COG0614	ABC type Fe3+ siderophores transport system
NCgl1958	COG0681	Signal peptidase I
NCgl1957	COG0164	Ribonuclease HII
NCgl1956	-	Hypothetical protein
NCgl0359	-	Hypothetical membrane protein
NCgl0360	COG1053	Succinate dehydrogenase/fumarate reductase, flavoprotein subunit
NCgl0361	COG0479	Succinate dehydrogenase/fumarate reductase Fe-S protein
NCgl0362	-	Hypothetical membrane protein
NCgl0363	-	Hypothetical protein
NCgl0377	-	Haemin transport system associated protein
NCgl0378	-	ABC type transporter, periplasmic component
NCgl0379	COG0609	ABC type transporter, permease component
NCgl0380	COG4559	ABC type transporter, ATPase component
NCgl2902	-	NADPH:quinone reductase
NCgl2901	COG0350	Methylated DNA protein cysteine methyltransferase (Ada)
NCgl2766	COG1275	Hypothetical membrane
NCgl0903	COG3607	Predicted lactoylglutthione lyase
NCgl0381	-	Haemin transport system associated protein
NCgl0382	-	Haemin transport system associated protein
NCgl1949	COG0264	Translation elongation factor
NCgl2897	COG0783	Starvation inducible DNA binding protein

Table 6.2 Contnd.

Gene	COG No.	Product
NCgl0639	COG0614	ABC type transporter
NCgl0329	COG0614	ABC-type Fe3+-citrate transport system, periplasmic component
NCgl2898	-	Formamidopyrimidine DNA- glycosylase
NCgl1646	COG0265	Hypothetical protein
NCgl1647	-	Hypothetical protein
NCgl0914	COG1132	ABC-type multidrug transporter ATPase and permease component
NCgl0915	COG1132	ABC-type multidrug transporter, ATPase and permease component
NCgl2990		Glucose inhibited division protein B
NCgl2729	COG0477	ABC type transporter, permease component
NCgl2971	COG0604	Putative oxidoreductase/dehydrogenase
NCgl2972	COG3759	Hypothetical membrane protein
NCgl2439	COG1528	Ferritin like protein
NCgl1703	COG0270	Site specific DNA methylase
NCgl1704	-	
NCgl1705	-	
NCgl0775	COG4243	Hypothetical membrane protein
NCgl0851	COG2259	Predicted membrane protein
NCgl0852		Hypothetical membrane protein
NCgl0853	COG0366	Glycosidase
NCgl1395	COG2375	Siderophore interacting protein
NCgl2970	COG0614	ABC type transport systems, periplasmic component
NCgl2412	COG4578	Hypothetical membrane protein
NCgl2413	-	Hypothetical membrane protein
NCgl0774	COG0614	ABC type Fe3+ siderophore transport system, periplasmic component
NCgl0773	COG2375	Siderophore interacting protein

Note: Genes that are together belongs to same operon.

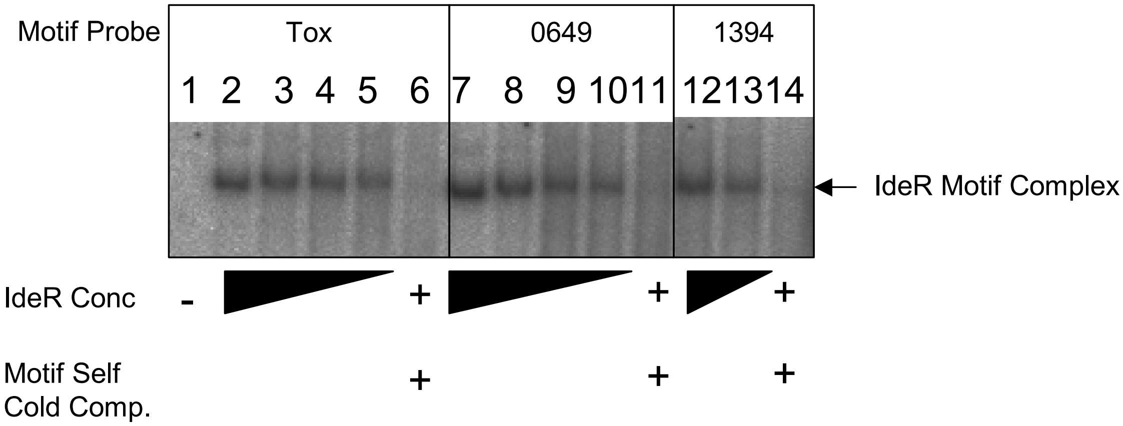


Figure 6.1: IdeR shows concentration dependent EMS with ds oligo representing the DtxR binding motifs

Lane 1: Radiolabeled Tox motif without IdeR. Lane 2-5: Radiolabeled Tox motif with decreasing concentration of IdeR (80,60,40,20 picomoles). Lane 6: Radiolabeled Tox motif with 20 picomoles of IdeR and 48 picomoles of cold Tox motif as cold competitor. Lane 7-10: Radiolabeled predicted motif, 0649, with decreasing concentration of IdeR (80,60,40,20 picomoles). Lane 11: Radiolabeled 0649 motif with 20 picomoles of IdeR and 48 picomoles of cold 0649 motif as cold competitor. Lane 12-13: Radiolabeled predicted motif, 1349, with decreasing concentration of IdeR (60, 20 picomoles). Lane 14: Radiolabeled 1394 motif with 20 picomoles of IdeR and 48 picomoles of cold 1394 motif as cold competitor.

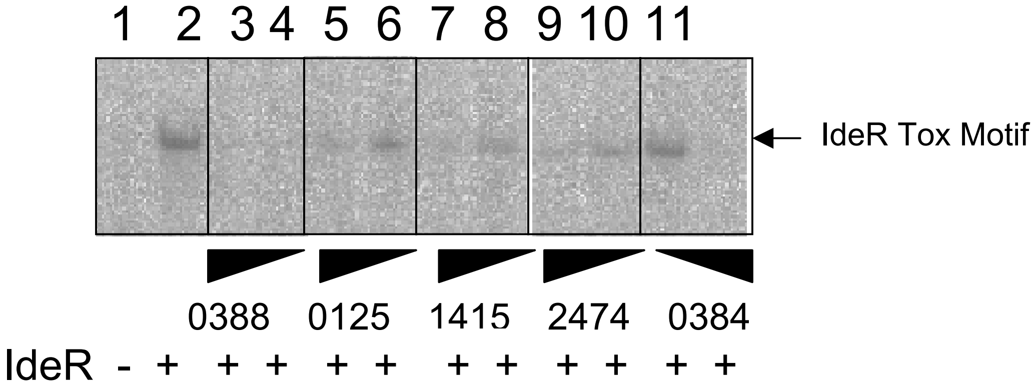


Figure 6.2: The predicted DtxR binding sites competes with radiolabeled Tox motif in its binding to IdeR

Lane 1: Radiolabeled Tox motif without IdeR. Lane 2: Radiolabeled Tox motif with 20 pico moles of IdeR Lane 3-4: Radiolabeled Tox motif with 20 picomoles of IdeR and decreasing concentration of cold 0388 motif (48, 24 picomples) as cold competitor. Lane 5-6: Radiolabeled Tox motif with 20 picomoles of IdeR and decreasing concentration of cold 0125 motif (48, 24 picomoles) as cold competitor. Lane 7-8: Radiolabeled Tox motif with 20 picomoles of IdeR and decreasing concentration of cold 1415 motif (48, 24 picomples) as cold competitor. Lane 9-10: Radiolabeled Tox motif with 20 picomoles of IdeR and decreasing concentration of cold 2474 motif (48, 24 picomples) as cold competitor. Lane 11-12: Radiolabeled Tox motif with 20 picomoles of IdeR and decreasing concentration of cold 0384 motif (48, 24 picomples) as cold competitor.

6.3 Discussion

Function for the proteins encoded by the genes in Table 6.2 was predicted by Reversed Position Specific-Basic Local Alignment Search Tool (RPS-BLAST) search against conserved domain database (Marchler-Bauer *et al.*, 2003). Some of the important genes/operons controlled by DtxR are described here.

6.3.1 Regulation of ABC type ferric siderophore transport systems

The genes NCgl0639, NCgl0638, NCgl0637 and NCgl0636 are part of an operon are similar to the *irp*1A, *irp*1B, *irp*1C and *irp*1D genes of *C. diphtheriae* respectively and belong to the ferric-siderophore transport system (Qian et al., 2002). In comparison to the *C. diphtheriae,* the operon contains additional genes that code for siderophore interacting protein (NCgl0635), isocitrate dehydrogenase (NCgl0634) and a predicted membrane protein (NCgl0633).

The operon with the genes, NCgl0484, NCgl0483, NCgl0483 and the gene NCgl0329 were similar to the *fag*A, *fag*B, *fag*C and *fag*D genes respectively of the ferric-siderophore transport system in *Corynebacterium pseudotuberculosis.* These four genes (*fag*A, *fag*B, *fag*C and *fag*D) are also identified as virulence genes in *Corynebacterium pseudotuberculosis* (Billington *et al*., 2002).

6.3.2 Regulation of Hemolysins

The genes NCgl1393 and NCgl1392 belong to the same orthologous gene group (COG1253) that code for Hemolysins containing Cystathionine Beta Synthase (CBS) domains. These genes were similar to the *tlyC* gene of other bacteria such as *Mycxococcus xanthus*, *Treponema hyodysenteriae* and *Rickettsiae typhi* (ter Huurne *et al*., 1993). The Hemolysin (*tlyC*) lyses host red blood cells and makes iron more available by releasing hemoglobin-bound iron as shown by Typhus group *Rickettsiae* (*R. typhi* and *R. prowazekii*), which adhere to and lyse human erythrocytes. Hemolysin (*tlyc*) is also identified as an important virulence gene in *Treponema hyodysenteriae* (ter Huurne *et al*., 1993).

6.3.4 Regulation of hemin transport

The genes NCgl0378, NCgl0379 and NCgl0380 belonging to an operon were similar to the *hmu*T, *hmu*U and *hmu*V genes respectively, of the hemin transport system in *C. diphtheriae and Corynebacterium ulcerons* (Drazek *et al*., 2000). The gene NCgl0378 associated with the same operon and other two genes, NCgl0381 and NCgl0382 of the adjacent operon are similar to the Hemin transport associated proteins in *C. diphtheriae* and *Corynebacterium ulcerons* (Schmitt *et al*., 2001). The gene NCgl2146 encodes a haem oxygenase (*hmu*O) homologue, which is involved in release of iron from haem in *C. diphtheriae* (Schmitt *et al*., 1997).

6.3.5 Regulation of Iron storage and oxidative stress defence

DtxR could regulate the genes NCgl2439 and NCgl2897 whose products are orthologous to ferrtin and starvation inducible DNA binding protein (Dps), respectively. In several bacteria, ferritin oxidizes and stores iron to supply iron under iron deficient conditions (Andrews *et al*., 1998). Dps in *E. coli* (*E. coli*) is induced in response to oxidative or nutritional stress and protects DNA from oxidative stress damage by nonspecific binding (Martinez *et al*., 1997). Dps also oxidizes ferrous iron to ferric iron by hydrogen peroxide, which in turn prevents hydroxyl radical formation by Fenton's reaction (Zhao *et al*., 2002). Ferritin (Rv3841) in *M. tuberculosis* induced by IdeR an ortholog of DtxR (Rodriguez *et al*., 2002). It is likely that DtxR like IdeR could also function as an activator of iron storage proteins.

6.3.6 Regulation of genes involved in DNA repair

Ferrous iron induced oxidative stress can damage the DNA. Our prediction shows that the genes, whose products are orthologs of DNA repairing proteins in *E. coli*- could be regulated by DtxR. The products of the two genes, NCgl2902 and NCgl2901 (Table 6.2) are orthologous to Ada and Nei proteins in *E. coli*, respectively. The Ada protein repairs alkylated guanine in DNA by transferring the alkyl group at the O-6 position to a cysteine residue in the protein. The methylated Ada protein acts as a positive regulator of its own synthesis, as well as the other iron containing proteins (AlkB) involved in DNA repair (Kleibl *et al*., 2002). The protein Nei in *E. coli* is a DNA-glycosylase, which removes oxidative products of thymine and 5-methyl cytosine from DNA (Hori *et al*., 2002).

6.4 Conclusions

C. glutamicum shows distinct subset of DtxR regulated genes in comparison to pathogenic *C. diphtheriae*. In *C. glutamicum*, DtxR regulates the genes that code for siderophore interacting protein (NCgl0635) and isocitrate dehydrogenase (NCgl0634), which are part of the operon that code for the proteins involved in siderophore transport. It also regulates the predicted operon containing the genes that code for hemolysins and iron storage proteins (BfrA) in *C. glutamicum.* In addition, the genes that code for the orthologs of adaptive response regulator (Ada) and endonuclease VIII (Nei) involved in DNA repair could also be regulated by DtxR.

Iron is although an essential element, it can catalyze formation of hydroxyl radicals and reactive oxygen species through Fenton’s reaction, which could be lethal to the cell. Hence, careful regulation of iron levels in cell is necessary for survival of bacteria. The data shows that DtxR regulates the iron homeostasis in *C. glutamicum* by controlling the genes involved in iron release, uptake and iron storage. In addition, it also regulates DNA repair enzymes to protect DNA in case there is oxidative stress affecting the DNA. Hence iron homeostasis and prevention of cellular damages due to Fenton's reaction could be the most important role of DtxR.

Chapter 7

Prediciton of IdeR Regulons in Mycobacteria

Homologues of DtxR family of transcription regulators, present in all the sequenced genomes of mycobacteria and related organism, *N. farcinica* (Urbanski and Beresewicz, 2000). The binding sites and target genes of DtxR homologue called Iron dependent regulator (IdeR) in *M. tuberculosis* are relatively better known. In *M. tuberculosis*, IdeR has been known to govern the expression of a wide variety of genes ranging from those involved in iron acquisition and oxidative stress response to ones that code for enzymes involved in aromatic amino acid biosynthesis (Gold *et al*., 2001; Rodriguez and Smith, 2003). Electrophoretic mobility shift assay and DNA footprinting analysis has lead to the identification of IdeR binding sites in upstream sequences of genes that code the proteins involved in biosynthesis of siderophores (MbtA, MbtB, MbtI), aromatic amino acids (PheA, HisE, HisG), lipopolysacaharide (Rv3402c), lipids (AcpP), peptidoglycon (MurB) and others annotated to be involved in iron storage (BfrA, BfrB) (Rodriguez *et al*., 1999; Gold *et al*., 2001). DNA microarray analysis of iron-dependent transcriptional profiles of wild-type and IdeR mutant of *M. tuberculosis* has lead to the identification of variety of other genes that code for the proteins, including putative transporters (Rv0282, Rv0283, Rv0284), members of the glycine-rich PE/PPE family (Rv2123), membrane proteins involved in virulence (MmpL4, MmpS4), transcriptional regulators, enzymes involved in lipid metabolism (Rv1344, Rv1345, Rv1346, Rv1347) and amino acid metabolism (TrpE2, PheA) (Rodriguez *et al*., 2002).

The work identifies common and unique Iron regulated genes in various sequenced *Mycobacterium* species and related organism *N. farcinica.* The ‘Predictregulon’ was used to identify the IdeR binding motifs upstream to the *Mycobacterium* genes and the operon context of that motif to identify IdeR dependent iron regulated genes in genomes *M. bovis*, *M. avium sub sp paratuberculosis, M. marinum* and *M. smegamtis*. Previously reported IdeR binding sites from *M. tuberculosis* were used to generate a recognition profile based on Shannon relative entropy, which was used to predict potential IdeR sites in the genomes of *M. bovis*, *M. avium sub sp paratuberculosis, M. marinum* and *M. smegamtis.* A sample of predicted motifs in *M.*

smegmatis was experimentally verified by EMSA using recombinant IdeR of *M. tuberculosis*.

7.1 Method

Published and annotated genome sequences of *M. tuberculosis, M. bovis* and *M. avium subsp. paratuberculosis* were downloaded from NCBI ftp site (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). Unpublished and un-annotated genome sequence of M. *marinum* was downloaded from sanger site (http://www.sanger.ac.uk/Projects/Microbes/) and *M. smegmatis* was from TIGR site (http://www.tigr.org/tdb/mdb/mdbinprogress.html). The genome sequences of *M. marinum and M. smegmatis* were annotated by GLIMMER software (Delcher *et al*., 1999). The gene name contains two letters followed by a number. First letter represent the name of genus and second letter represent the name of species. The number was given according to the order of gene appearance in the genome.

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

pQE30 expression vector (Qiagen) with an N terminal 6X His tag was used to clone the ORF Rv2711 of *M. tuberculosis* that encodes IdeR. Briefly, Rv2711 was taken out from pRSETIdeR construct with specific restriction enzyme sites (BamH1 and HindIII) and the insert was cloned into the corresponding sites of pQE30 expression vector. *E coli* M15 cells transformed with the 6xHis tagged chimeric construct were grown in 400mL of LB medium supplemented with 100µg/ml of ampicillin and 25µg/ml of kanamycin. IPTG (0.2mM) was added to a mid log phase culture. The cells were kept in an incubator shaker for another eight hours at 27^0C and 200 rpm to allow protein expression. Then, cells were harvested by centrifugation and resuspended in 10 ml of lysis buffer (50mM NaH_2PO_4, 300mM NaCl and 10mM imidazole, pH 8) with 1mM PMSF and disrupted using a sonicator. After a second round of centrifugation for 10 minutes at 10,000xg, the supernatant was applied to a Ni-NTA affinity column (Qiagen, USA). The supernatant was allowed to bind to Ni-NTA column. The recombinant protein was eluted with

200mM imidazole and analyzed by SDS PAGE after washing the column with 5 bed-volumes of lysis buffer containing 10mM imidazole.

7.1.2 Electrophoretic mobility shift assay

Double-stranded oligonucleotides containing the predicted binding motif (19 bp long) were end labeled with T4 polynucleotide kinase and [γ32P]-ATP and were incubated with the purified recombinant IdeR protein in a binding reaction mixture. The binding reaction mixture (20-µl total volume) contains the DNA-binding buffer (20 mM Tris-HCl [pH 8.0], 2 mM DTT, 50 mM NaCl, 5 mM MgCl2, 50% glycerol, 5 µg of bovine serum albumin per ml), 10 µg of poly (dI-dC) per ml (for nonspecific binding) and 200 µM $NiSO_4$. The reaction mixture was incubated at room temperature for 30 min and loaded onto 7% polyacrylamide gel containing 1 × Tris-borate-EDTA buffer. No dye was added for loading. The gel was electrophoresed at 200 Volts for 2 hours. Subsequently the gel was dried and exposed to Fuji Storage Phosphor Image Plates for 4 hours. The image plates were subsequently scanned in Storage Phosphor Imaging workstation.

7.2 Results

7.2.1 IdeR from various *Mycobacterium* species has identical DNA binding domain

IdeR orthrologs were aligned with each other (Figure 7.1). Alignment of the DNA binding domain show very high sequence identity which suggest that the target DNA motifs in various genomes can be recognized based on sequence recognition profile generated from experimentally defined IdeR target motifs from *M. tuberculosis*.

7.2.2 In-silico prediction of IdeR binding sites and target operons

A recognition profile of experimentally defined IdeR binding sites (Table 7.1) from *M. tuberculosis* was used to identify the potential IdeR binding sites and downstream

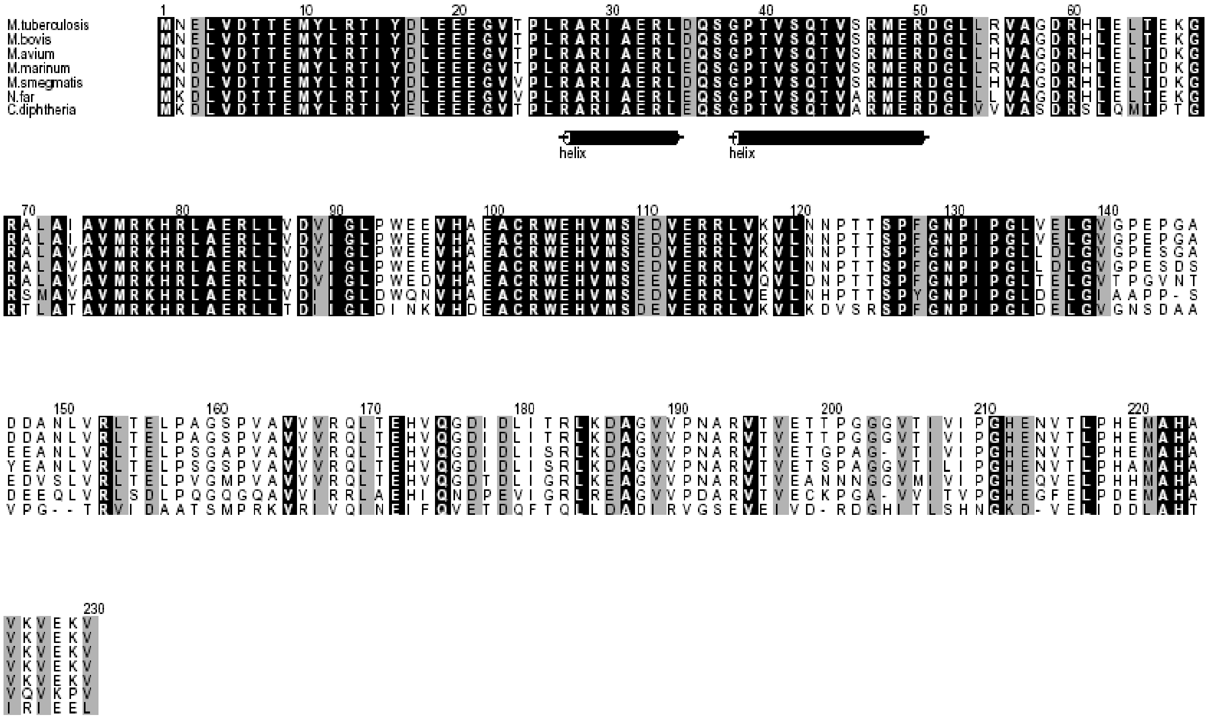


Figure 7.1: Alignment of IdeR orthologues from different species of Actinobacteria suggests highly conserved DNA binding domain

The arial black shadow show identity and the gray show similarity. Two helices are part of helix turn helix that binds to the IdeR box.

Table 7.1. Known IdeR binding sites from *M. tuberculosis*

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

operons/genes in genomes of *M. bovis* (Table 7.2 and Table 7.3), *M. avium sub sp paratuberculosis* (Table 7.4 and Table 7.5), *M. marinum* (Table 7.6 and Table 7.7), *M. smegamtis* (Table 7.8 and Table 7.9), *M. leprae* (Table 7.10 and Table 7.11) and *N. farcinia* (Table 7.12 and Table 7.13). Function for the proteins encoded by these genes was predicted by Reversed Position Specific-Basic Local Alignment Search Tool (RPS-BLAST) search against conserved domain database (Marchler-Bauer *et al*., 2003).

7.2.3 Experimental validation of predicted binding sites

A sample of predicted regulator binding motifs in (Table 7.8) upstream sequences of the *M. smegmatis* genes that code for predicted Fe2+-dicitrate sensor (FecR), periplasmic component of ABC-type Fe3+-hydroxamate transport system (FepB), Siderophore-interacting protein (ViuB) and a predicted motif in intergenic sequence of the divergently transcribed genes that were orthologous to the Rv1846 and Rv1847 were experimentally verified by EMSA using recombinant IdeR from *M. tuberculosis.* Double stranded 19-mer synthetic oligonucleotides corresponding to the predicted DNA-binding sites were labeled with ^{32}PγATP and mixed with purified IdeR in presence of Nickel ions and was assayed for the formation of DNA-protein complex using EMSA. Nickel was used as the divalent metal in the binding reactions on account of its redox stability compared with ferrous ion. IdeR is able to retard the electrophoretic mobility of the four double stranded oligonucleotides (Figure 7.2) out of the five tested. A synthetic motif- ds (5'-TTTTCATGACGTCTTCTAA-3') which was used as a negative control, did not show any complex formation. These results indicate that the predicted IdeR-binding sites can indeed bind to IdeR though the level of affinity may vary.

Table 7.2. Predicted IdeR binding sites in *M. bovis*

Score	Position	Binding site	Synonym	Product
6.15515	-151	ATAGGCAAGGCTGCCCTAA	Mb1877c	Predicted transcriptional regulator
6.12131	-51	ATAGGTTAGGCTACCCTAG	Mb2147	PPE-repeat proteins
6.11144	-85	TTAGGCAAGGCTAGCCTTG	Mb1378c	Glucitol operon activator
6.08177	-226	TTAGTGGAGTCTAACCTAA	Mb1907	Bacterioferritin
6.04733	-86	TTAGCACAGGCTGCCCTAA	Mb2405	Peptide arylation enzymes
6.04498	-73	CTAGGAAAGCCTTTCCTGA	Mb3871	Ferritin-like protein
6.02144	-345	CAAGGTAAGGCTAGCCTTA	Mb1547	Glycosyltransferases involved in cell wall biogenesis
6.02144	-50	CAAGGTAAGGCTAGCCTTA	Mb1546	pyridoxal phosphate-dependent enzyme, cell wall biogenesis
6.01356	-379	CTAGGGTAGCCTAACCTAT	Mb2145c	ATP phosphoribosyltransferase
6.01356	-95	CTAGGGTAGCCTAACCTAT	Mb2146c	Phosphoribosyl-ATP pyrophosphohydrolase
5.97952	-32	TTAGGGCAGCCTGTGCTAA	Mb2404c	Non-ribosomal peptide synthetase modules
5.95887	-2	TTATGTTAGCCTTCCCTTA	Mb3437c	Uncharacterized protein conserved in
5.95145	-79	TTAGGTAAGCCTAAGTTAA	Mb3868c	PheA, Prephenate dehydratase
5.94295	-36	TTAACTTAGGCTTACCTAA	Mb3869	CobH, Precorrin isomerase
5.89908	-292	CAAGGCTAGCCTTGCCTAA	Mb1380	Acyl-CoA synthetases (AMP-forming)/AMP-acid ligases II
5.89908	21	CAAGGCTAGCCTTGCCTAA	Mb1379	AcpP, Acyl carrier protein
5.89068	-145	TTAGGGCAGCCTTGCCTAT	Mb1878	possibly involved in aromatic compounds catabolism
5.86021	-140	AGAGGTAAGGCTAACCTCA	Mb3436c	pyridoxal phosphate-dependent enzyme, cell wall biogenesis
5.81511	-26	GCAGGTCAGGCTACCCTTA	Mb0492	MurB, UDP-N-acetylmuramate dehydrogenase
5.80946	-25	GTAGGTTAGGCTACATTTA	Mb2407c	Anthranilate/para-aminobenzoate synthases component
5.551	-36	ATAGGAAAGCCGATCCTTA	Mb0118	HisB, Histidinol phosphatase and related phosphatases
5.48236	-20	TAAGGGTAGCCTGACCTGC	Mb0491c	Penicillin V acylase and related amidases
5.46618	-302	GTAGACCAGGCTCCCCTTG	Mb3070	ABC-type Fe3+-hydroxamate transport system, periplasmic
5.46363	-112	TTAGGCTAGGCTTAGTTGC	Mb0459c	Predicted transcriptional regulators CopG/Arc/MetJ
5.39538	-139	GCAACTAAGCCTAGCCTAA	Mb0460	AcrR, Transcriptional regulator
5.36878	-50	TTAGCTTATGCAATGCTAA	Mb0290	SpoVK, ATPases of the AAA+ class
5.35233	-213	TTCGGTAAGGCAACCCTTA	Mb1383	ABC-type multidrug transport system, ATPase and permease

Table 7.3. Predicted IdeR regulated operons in *M. bovis*

Synonym	Gene	COG No.	Product
Mb1876c	-	-	Conserved Hypothetical Transmembrane Protein
Mb1877c	-	-	Possible Transcriptional Regulatory Protein
Mb1878	-	-	4-HBA-Coa Thioesterase
Mb1879	*ureA*	-	Urease Gamma Subunit Urea (Urea Amidohydrolase)
Mb1880	*ureB*	-	Urease Beta Subunit Ureb
Mb1881	*ureC*	-	Urease Alpha Subunit Urec (Urea Amidohydrolase)
Mb1882	*ureF*	-	Urease Accessory Protein Uref
Mb1883	*ureG*	-	Urease Accessory Protein Urge
Mb1884	*ureD*	-	Probable Urease Accessory Protein Ured
Mb2145c	*hisG*	-	Probable Atp Phosphoribosyltransferase Hisg
Mb2146c	*hisE*	-	Probable Phosphoribosyl-Amp Pyrophosphatase Hise
Mb2147	*PPE37*	-	Conserved Hypothetical Protein, Ppe
Mb1377c	-	-	Conserved Membrane Protein
Mb1378c	*lprD*	-	Probable Conserved Lipoprotein Lprd
Mb1379	-	-	Probable Acyl Carrier Protein (Acp)
Mb1380	*fadD33*	-	Possible Polyketide Synthase Fadd33
Mb1381	*fadE14*	-	Possible Acyl-Coa Dehydrogenase Fade14
Mb1907	*bfrA*	-	Probable Bacterioferritin Bfra
Mb2398c	*mbtH*	-	Putative Conserved Protein Mbth
Mb2399c	*mbtG*	-	Lysine-N-Oxygenase Mbtg (L-Lysine 6-Monooxygenase) (Lysine N6-Hydroxylase)
Mb2400c	*mbtF*	-	Peptide Synthetase Mbtf (Peptide Synthase)
Mb2401c	*mbtE*	-	Peptide Synthetase Mbte (Peptide Synthase)
Mb2402c	*mbtD*	-	Polyketide Synthetase Mbtd (Polyketide Synthase)
Mb2403c	*mbtC*	-	Polyketide Synthetase Mbtc (Polyketide Synthase)
Mb2404c	*mbtB*	-	Phenyloxazoline Synthase Mbtb (Phenyloxazoline Synthetase)
Mb2405	*mbtA*	-	Bifunctional Enzyme Mbta: Salicyl-Amp Ligase (Sal-Amp Ligase) + Salicyl-S-Arcp Synthetase
Mb2406	*mbtJ*	-	Putative Acetyl Hydrolase Mbtj
Mb3871	*bfrB*	-	Possible Bacterioferritin Bfrb
Mb3872c	*glpQ1*	-	Probable Glycerophosphoryl Diester Phosphodiesterase Glpq1 (Glycerophosphodiester Phosphodiesterase)
Mb3873c	-	-	Probable Conserved Transmembrane Protein
Mb1547	-	-	Probable Sugar Transferase
Mb1548	*fadD25*	-	Probable Fatty-Acid-Coa Ligase Fadd25 (Fatty-Acid-Coa Synthetase) (Fatty-Acid-Coa Synthase)
Mb3437c	-	-	Hypothetical Protein
Mb3867c	-	-	Probable Phosphoglycerate Mutase (Phosphoglyceromutase) (Phosphoglycerate Phosphomutase)
Mb3868c	*pheA*	-	Possible Prephenate Dehydratase Phea
Mb3869	-	-	Conserved Hypothetical Protein
Mb3870	-	-	Possible Transcriptional Regulatory Protein
Mb3436c	-	-	Conserved Hypothetical Protein
Mb0489c	-	-	Probable Conserved Membrane Protein
Mb0490c	-	-	Conserved Hypothetical Protein
Mb0491c	-	-	Hypothetical Protein
Mb0492	*murB*	-	Probable Udp-N-Acetylenolpyruvoylglucosamine Reductase Murb (Udp-N-Acetylmuramate Dehydrogenase)
Mb0493	*lprQ*	-	Probable Conserved Lipoprotein Lprq
Mb0118	-	-	Possible Dehydratase

Table 7.3. Contnd.

Synonym	Gene	COG No.	Product
Mb3062c	*TB22.2*	-	Probable Conserved Secreted Protein Tb22.2
Mb3063c	-	-	Conserved Hypothetical Protein
Mb3064c	-	-	Conserved Hypothetical Protein
Mb3065c	*echA17*	-	Probable Enoyl-Coa Hydratase Echa17 (Crotonase) (Unsatured Acyl-Coa Hydratase)
Mb3066c	-	-	Conserved Hypothetical Protein
Mb3067c	-	-	Probable Conserved Atp-Binding Protein Abc Transporter
Mb3068c	*serB2*	-	Probable Phosphoserine Phosphatase Serb2 (Psp) (O-Phosphoserine Phosphohydrolase) (Pspase)
Mb3069c	*ctaD*	-	Probable Cytochrome C Oxidase Polypeptide Ctad (Cytochrome Aa3 Subunit 1)
Mb3070	*fecB*	-	Probable Feiii-Dicitrate-Binding Periplasmic Lipoprotein Fecb
Mb3071	*adhC*	-	Probable Nadp-Dependent Alcohol Dehydrogenase Adhc
Mb0458c	*mmpL4*	-	Probable Conserved Transmembrane Transport Protein Mmpl4
Mb0459c	*mmpS4*	-	Probable Conserved Membrane Protein Mmps4
Mb0460	-	-	Possible Transcriptional Regulatory Protein
Mb1382c	-	COG1670	Riml, Acetyltransferases, Including N-Acetylases Of Ribosomal Proteins
Mb1383	-	-	Probable Drugs-Transport Transmembrane Atp-Binding Protein Abc Transporter
Mb1384	-	-	Probable Drugs-Transport Transmembrane Atp-Binding Protein Abc Transporter

Note: Genes that are part of an operon are together

Table 7.4 Predicted IdeR binding sites in *M avium sub sp. paratuberculosis*

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

Table 7.5. Predicted IdeR regulated operons in *M. avium sub sp paratuberculosis*

Synonym	Gene	COG No.	Product
MAP1594c	-	-	Bacterioferritin-associated ferredoxin
MAP1595	*bfrA*	COG2193	BfrA
MAP1558c	-	COG0501	Zn-dependent protease
MAP1559c	-	COG3682	Transcription regulator
MAP1560	-	COG2050	Possibly involved in aromatic compounds catabolism
MAP0191c	-	COG1316	hypothetical protein
MAP0192c	-	COG4175	hypothetical protein
MAP0193	*pheA*	COG0077	PheA
MAP0194	-	COG0406	Fructose-2,6-bisphosphatase
MAP2169c	*mbtH_3*	COG3251	MbtH_3
MAP2170c	*mbtG*	COG3486	MbtG
MAP2171c	*mbtF*	COG1020	MbtF
MAP2172c	-	COG1020	putative non-ribosomal peptide synthetase
MAP2173c	*mbtE*	COG1020	MbtE
MAP2174c	*mbtD*	COG3321	MbtD
MAP2175c	*mbtC*	COG3321	MbtC
MAP2176c	-	COG3208	Thio esterase (similar to mbtB)
MAP2177c	*mbtB*	COG1020	MbtB
MAP2178	*mbtA*	COG1021	MbtA
MAP2179	-	-	hypothetical protein
MAP2205c	*trpE2*	COG0147	TrpE2
MAP2206	-	COG3329	Predicted permease
MAP2051c	-	COG2124	Cytochrome P450 monooxygenase
MAP2052c	-	-	Bacterial regulatory proteins, tetR family
MAP2053	-	-	Hypothetical protein
MAP0791c	-	COG2226	hypothetical protein
MAP0792c	-	COG2141	F420-dependent N5,N10-methylene tetrahydromethanopterin reductase
MAP0793c	-	COG0654	monooxygenase, FAD-binding
MAP0794	-	COG1309	Bacterial regulatory proteins, tetR family
MAP0795	-	COG2141	Luciferase-like monooxygenase
MAP0024c	-	COG5651	PPE-repeat proteins
MAP0025	-	COG0236	Acyl carrier protein
MAP0026	*fadD33_1*	COG0318	FadD33_1
MAP4064c	-	COG3315	O-Methyltransferase involved in polyketide biosynthesis
MAP4065	-	COG1914	Nramp
MAP1760c	-	COG2837	Predicted_iron-dependent_peroxidase
MAP1761c	-	COG2822	Predicted periplasmic lipoprotein involved in iron transport
MAP1762c	-	COG0672	FTR1, High-affinity Fe2+/Pb2+ permease
MAP1553c	*fadE14*	COG1960	FadE14
MAP1554c	*fadD33_2*	COG0318	FadD33_2
MAP1555c	-	COG0236	Acyl carrier protein
MAP3777	-	COG3315	O-Methyltransferase involved in polyketide biosynthesis
MAP3778	-	COG0464	hypothetical protein
MAP3779	-		
MAP3780	-		
MAP3781	-		

Table 7.5. Contnd.

Synonym	Gene	COG No.	Product
MAP0677c	-	COG2159	hypothetical protein
MAP0678c	-	COG2329	enzyme involved in biosynthesis of extracellular polysaccharides
MAP0679c	*fdxB*	COG0633	FdxB
MAP0680	-	COG0318	Acyl-CoA synthetases (AMP-forming)/AMP-acid ligases II
MAP0681	-	COG1960	acyl-CoA dehydrogenase
MAP0682	-	COG1960	Putative acyl-CoA dehydrogenase
MAP0683	-	COG1024	Enoyl-CoA hydratase/isomerase family
MAP3973c	-	COG0388	Predicted amidohydrolase
MAP3974c	-	COG3832	Predicted lactoylglutathione lyase
MAP3975	*murB*	COG0812	MurB
MAP3976	-	COG1376	putative lipoprotein
MAP2412c	-	COG3173	Predicted aminoglycoside phosphotransferase
MAP2413c	-	COG1132	ABC-type multidrug/protein/lipid transport system
MAP2414c	-	COG1132	ABC-type multidrug/protein/lipid transport system
MAP2109c	-	COG2516	Predicted Fe-S oxidoreductases
MAP2110c	-	COG1575	1,4-dihydroxy-2-naphthoate octaprenyltransferase
MAP2111c	-	COG1463	ABC-type transport system , resistance to organic solvents, periplasmic
MAP2958c	*xerC*	COG4974	XerC
MAP2959c	-	COG1304	L-lactate dehydrogenase
MAP2960c	*viuB*	COG2375	ViuB

Note: Genes that are part of an operon are together

Table 7.6. Predicted IdeR binding sites in *M. marinum*

Score	Position	Binding site	Gene	Product
6.15305	-151	ATAGGCAAGGCTGCCCTAA	MM0626	Predicted Transcriptional Regulator
6.12261	-175	TTAGTTGAGTCTAACCTAA	*bfr*	Bfr, Bacterioferritin (Cytochrome B1)
6.05597	-38	TTAGCCCAGGCTGTCCTAA	*entE*	Ente, Peptide Arylation Enzymes
5.98846	-50	TTAGGTTAGACTCAACTAA	*bfd*	Bfd, Bacterioferritin-Associated Ferredoxin
5.97169	-28	TTAGGACAGCCTGGGCTAA	*lucE*	Lysine/Ornithine N-Monooxygenase
5.97169	-35	TTAGGACAGCCTGGGCTAA	-	Homolog Of Phage Mu Protein Gp30
5.94848	-29	TTAGGTAAGCCTAAGTTAA	*pheA*	Prephenate Dehydratase
5.94016	-23	TTAACTTAGGCTTACCTAA	MM3688c	Mu-Like Prophage Protein
5.93817	-42	ATAGGTTAGCCTAACTTTA	*ppe*	PPE-Repeat Proteins
5.88869	-220	TTAGGGCAGCCTTGCCTAT	*paaI*	Involved In Aromatic Compounds Catabolism
5.81087	-63	TTAGGCAAACCTGACCTTA	*ftn*	Ftn, Ferritin-Like Protein
5.80679	-25	GTAGGTTAGGCTACATTTA	*TrpE2*	Anthranilate/Para-Aminobenzoate Synthases 1
5.5772	-355	TAAAGTTAGGCTAACCTAT	MM0189	Large Exoproteins Involved In Heme Utilization Or Adhesion
5.45817	-250	TTAGGCTAGGCTTGGTTGC	MM2728	Ribosomal Protein L1
5.36613	-79	TTAGCTTATGCAATGCTAA	MM2989c	Atpases of The AAA+ Class
5.36533	-135	TTAGCCAAGACTTCTGTGA	MM0037	Periplasmic Protein Tonb, Links Inner And Outer Membranes
5.36509	14	GTAGTCCAGGCTGACGTCA	MM0542	Phosphatase
5.35819	-85	GCAACCAAGCCTAGCCTAA	MM2727c	Transcriptional Regulator
5.3317	-389	GTAGGTAAATGTAGCCTAA	MM5651	ABC-Type Uncharacterized Transport Systems
5.27839	-126	TTCGGCTACTCTGCCCTTA	MM5993c	Translation Initiation Factor 2, Gamma Subunit
5.26644	-135	CTAGAGTAGGCAACCGTAA	MM2830c	PPE-Repeat Proteins
5.26541	-23	TTCGGTGACGCTAGACTGA	MM2433	Nucleic-Acid-Binding Protein Implicated In Transcription Termination
5.24786	-207	ACAGGAGAGCCTGAACTCA	MM3528c	Signal Transduction Protein Containing Sensor And EAL Domains
5.24671	-360	TAAAGTAAGGCAACCCTTA	MM6014c	ABC-Type Multidrug Transport System, Atpase And Permease
5.22898	-231	ATTGAAAAGTCTTACCTGA	MM4392	Universal Stress Protein Uspa And Nucleotide-Binding Proteins
5.22898	-27	ATTGAAAAGTCTTACCTGA	MM4391	Phenylpropionate
5.1982	-336	CTACCGCAGCCTTACCTGG	MM0467	Acyl-Coa Dehydrogenases
5.19361	-143	TGAGTTCAGGCTCTCCTGT	MM3531	Tfp Pilus Assembly Protein
5.18451	-128	TTAGGCAACCCACGCCTGA	MM2462	FAD Synthase
5.17596	-111	CGAGCGGATGCTGGCCTTA	MM2805c	Tetrahydromethanopterin Reductase
5.17098	-54	TTCGGTAAGGCTAACATGG	MM4663	Transcriptional Regulator
5.16056	-38	CAAGACGAGGCTTGTCTAG	MM2256	Esterase/Lipase
5.14345	-330	CTACGGCAGGCTCTGCTGG	MM1490	Methylase Involved In Ubiquinone/Menaquinone Biosynthesis
5.13822	-153	ATAGGGAATCCTGGACTGC	MM3560	Uncharacterized Protein Conserved In Bacteria
5.13563	-17	TAAGGTCAGGCTCTCGTTG	MM0100	Predicted Integral Membrane Protein
5.13435	-147	ATCGATTAGGCTCTGCTCA	MM5168	Large Exoproteins Involved In Heme Utilization Or Adhesion
5.10298	-29	ATAGGGAAACCTGAAATTA	MM3095	Guanine Nucleotide Exchange Factor For Rho/Rac/Cdc42-Like Gtpases
5.09919	15	CGAAGTCAGCCTGGGCTGA	MM6019c	Rnase PH
5.09904	28	CGAGGTCACGCTTTCCTCG	MM4063	Predicted Glutamine Amidotransferase
5.09482	-273	CTTGGATAGACTGACCTGC	MM5805	Namn:DMB Phosphoribosyltransferase
5.08746	-365	CTAGCCCAGGCGACCCTGC	MM1832c	Predicted Unusual Protein Kinase
5.08247	-186	TTAGCGAAGGCTAACTAAA	MM5633c	Non-Ribosomal Peptide Synthetase Modules And Related Proteins
5.07084	-82	TCAGGAAATTCTCAACTGA	MM4422c	ABC-Type Dipeptide/Oligopeptide/Nickel Transport System, Atpase

Table 7.7. Predicted IdeR regulated operons in *M. marinum*

Gene	COG No.	Product
MM0619c	COG0829	UreH, Urease accessory protein UreH
MM0620c	COG0378	HypB, Ni2+-binding GTPase involved in regulation of expression and maturation of urease and hydrogenase
MM0621c	COG0830	UreF, Urease accessory protein UreF
MM0622c	COG0804	UreC, Urea amidohydrolase (urease) alpha subunit
MM0623c	COG0832	UreB, Urea amidohydrolase (urease) beta subunit
MM0624c	COG0831	UreA, Urea amidohydrolase (urease) gamma subunit
MM0625c	COG2050	PaaI, Uncharacterized protein, possibly involved in aromatic compounds catabolism
MM0626	COG3682	Predicted transcriptional regulator
MM0627	COG0501	HtpX, Zn-dependent protease with chaperone function
MM0578c	COG2193	Bfr, Bacterioferritin (cytochrome b1)
MM0579	COG2906	Bfd, Bacterioferritin-associated ferredoxin
MM5641c	COG2369	Uncharacterized protein, homolog of phage Mu protein gp30
MM5642	COG1021	EntE, Peptide arylation enzymes (mbtA)
MM5643	COG0657	Aes, Esterase/lipase (mbtJ)
MM5644	COG1028	FabG, Dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases)
MM5645	COG1543	Uncharacterized conserved protein
MM5646	COG2072	TrkA, Predicted flavoprotein involved in K+ transport
MM5647	COG0789	SoxR, Predicted transcriptional regulators
MM5640c	COG3486	IucD, Lysine/ornithine N-monooxygenase (mbtH)
MM3687c	COG1316	LytR, Transcriptional regulator
MM3688c	COG3941	Mu-like prophage protein
MM3689	COG0077	PheA, Prephenate dehydratase
MM3690	COG0406	GpmB, Fructose-2,6-bisphosphatase
MM0188c	COG5651	PPE-repeat proteins
MM0189	COG3210	PEPGRS
MM0190	COG0140	HisI, Phosphoribosyl-ATP pyrophosphohydrolase
MM0191	COG0040	HisG, ATP phosphoribosyltransferase
MM3685c	COG1528	Ftn, Ferritin-like protein
MM5650c	COG0147	TrpE, Anthranilate/para-aminobenzoate synthases component I
MM5651	COG3845	ABC-type uncharacterized transport systems, ATPase components
MM2725c	COG2409	Predicted drug exporters of the RND superfamily
MM2726c	COG4034	Uncharacterized protein conserved in archaea
MM2727c	COG1309	AcrR, Transcriptional regulator
MM2728	COG0081	mmpL4
MM2729	COG2409	mmpS4
MM2730	COG2907	Predicted NAD/FAD-binding protein
MM2731	COG3496	Uncharacterized conserved protein
MM2732	COG2230	Cfa, Cyclopropane fatty acid synthase and related methyltransferases
MM2733	COG3752	Predicted membrane protein
MM2734	COG1595	RpoE, DNA-directed RNA polymerase specialized sigma subunit, sigma24 homolog
MM2735	COG5343	Uncharacterized protein conserved in bacteria
MM0036c	COG0810	TonB, Periplasmic protein TonB, links inner and outer membranes
MM0037	COG0810	TonB, Periplasmic protein TonB, links inner and outer membranes
MM0038	COG3127	Predicted ABC-type transport system involved in lysophospholipase L1 biosynthesis, permease component
MM6012c	COG3173	Predicted aminoglycoside phosphotransferase
MM6013c	COG1132	MdlB, ABC-type multidrug transport system, ATPase and permease components
MM6014c	COG1132	MdlB, ABC-type multidrug transport system, ATPase and permease components

Table 7.7. Contnd.

Gene	COG No.	Product
MM0467	COG1960	CaiA, Acyl-CoA dehydrogenases
MM4391	COG4638	Phenylpropionate dioxygenase and related ring-hydroxylating dioxygenases, large terminal subunit
MM5631c	COG1020	EntF, Non-ribosomal peptide synthetase modules and related proteins (mbtB)
MM5632c	COG1020	EntF, Non-ribosomal peptide synthetase modules and related proteins (mbtG)
MM5633c	COG1020	EntF, Non-ribosomal peptide synthetase modules and related proteins (mbtF)
MM5634	COG1670	RimL, Acetyltransferases, including N-acetylases of ribosomal proteins
MM5635	COG3208	GrsT, Predicted thioesterase involved in non-ribosomal peptide biosynthesis
MM5636	COG3321	Polyketide synthase modules and related proteins (mbtC)
MM5637	COG3321	Polyketide synthase modules and related proteins (mbtE)

Note: Genes that are part of an operon are together

Table 7.8. Predicted IdeR binding sites in *M. smegmatis*

Score	Position	Binding site	Gene	Product
6.15382	-43	TTAGCGGAGTCTAACCTTA	ms3189c	Bfr, Bacterioferritin (cytochrome b1)
6.1384	-80	TTAGCACAGGCTGTCCTAA	ms4331	EntE, Peptide arylation enzymes
6.132	-64	TTAGGCAACGCTAAGCTAA	ms6168c	TolA, Membrane protein involved in colicin uptake
6.09661	-37	ATAGGCAAGGCTGGCCTCA	ms6169	Conserved protein/domain typically associated with flavoprotein oxygenases,
6.07218	-37	TTAGGACAGCCTGTGCTAA	ms4330c	EntF, Non-ribosomal peptide synthetase modules and related proteins
6.06438	-88	TTAAGTTAGGCTTACCTCA	ms6653	FecR, Fe2+-dicitrate sensor, membrane component
6.06438	-44	***TTAAGTTAGGCTTACCTCA***	ms6652	FecR, Fe2+-dicitrate sensor, membrane component
6.04683	-155	***TTAGGGAAGCCTTGCCTAT***	ms3260c	Possibly involved in aromatic compounds catabolism
5.97456	-25	CTAGGTTAGGCTACATTTA	ms4344c	TrpE, Anthranilate/para-aminobenzoate synthases component I
5.97182	-49	TTAGGTAACGCTGACCTCA	ms6656	Ftn, Ferritin-like protein
5.95731	-185	***ATAGCGAAGGCTAACCTAT***	ms7326c	FepB, ABC-type Fe3+-hydroxamate transport system, periplasmic component
5.95618	-67	TTAACGAAGGCTAGCCTCA	ms7417	Dehydrogenases with different specificities
5.84774	-40	ATAGGTTAGCCTTCGCTAT	ms7328	PanD, Aspartate 1-decarboxylase
5.8265	-23	TGAGGTAAGCCTAACTTAA	ms6650c	PheA, Prephenate dehydratase
5.80201	-54	TAAGGTTAGACTCCGCTAA	ms3190	Uncharacterized FAD-dependent dehydrogenases
5.77513	-46	***TAAGGGTACGCTTACCTTA***	ms4962c	ViuB, Siderophore-interacting protein
5.74283	-43	TAAGCCTAGCCTACCTTAA	ms1406c	AcpP, Acyl carrier protein
5.71356	-95	ATAGGTAAGCCTAACTTTG	ms0832c	SdhC, Succinate dehydrogenase/fumarate reductase, cytochrome b subunit
5.69516	-57	CAAAGTTAGGCTTTCCTTA	ms1556c	AraC-type DNA-binding domain-containing proteins
5.66395	-46	GAAGGTAAAGCTACCCTCA	ms1402	RimL, Acetyltransferases, including N-acetylases of ribosomal proteins
5.55147	-36	TGAGGCTAGCCTTCGTTAA	ms7416c	RPL15A, Ribosomal protein L15E
5.54238	-357	GTCGGCAAGCCTTTCCTGA	ms6851	AmpC, Beta-lactamase class C and other penicillin binding proteins
5.47365	-296	CCAGGAAAGGCTCAACTGA	ms7223c	CaiD, Enoyl-CoA hydratase/carnithine racemase [Lipid metabolism]
5.47365	-21	CCAGGAAAGGCTCAACTGA	ms7224c	CaiD, Enoyl-CoA hydratase/carnithine racemase
5.46104	-155	TAAGGAAAGCCTAACTTTG	ms1557	Uncharacterized conserved protein
5.44418	-25	CAAAGTTAGGCTTACCTAT	ms0833	Cdd, Cytidine deaminase
5.43697	-57	TTAGCTTAGGCATACATAA	ms8050	SpoVK, ATPases of the AAA+ class
5.42991	-20	***TTAGGTTACCCTCAGCTGT***	ms7314	ViuB, Siderophore-interacting protein
5.41897	-331	GTAGGTCAATCTCAGCTCA	ms1223c	TypA, Predicted membrane GTPase involved in stress response
5.40657	-47	TATAGTAAGGCTAACCTAA	ms3261c	Uncharacterized conserved protein
5.40177	-182	GTAGTGAAGTCTGTCATCA	ms5673	CaiA, Acyl-CoA dehydrogenases
5.37463	-257	TTAGCCTTGGCTAGCCTTG	ms5426c	MltB, Membrane-bound lytic murein transglycosylase B
5.32742	-52	ATTGGTAAGCCTTACCTTT	ms7321	Uncharacterized protein conserved in bacteria

Note: Binding sites with bold and italics were verified by EMSA.

Table 7.9. Predicted IdeR regulated operons in *M. smegmatis*

Gene	COG No.	Product
ms3189c	COG2193	Bfr, Bacterioferritin (cytochrome b1)
ms4331	COG1021	EntE, Peptide arylation enzymes (mbtA)
ms4332	COG1021	EntE, Peptide arylation enzymes
ms6168c	COG3064	TolA, Membrane protein involved in colicin uptake
ms6169	COG1853	Conserved protein/domain typically associated with flavoprotein oxygenases, DIM6/NTAB family
ms4330c	COG1020	EntF, Non-ribosomal peptide synthetase modules and related (mbtB)proteins
ms6652	COG3712	FecR, Fe2+ dicitrate sensor, membrane component
ms6653	COG3712	FecR, Fe2+ dicitrate sensor, membrane component
ms6654	COG1266	Predicted metal-dependent membrane protease
ms6655	COG1316	LytR, Transcriptional regulator
ms3258c	COG0832	UreB, Urea amidohydrolase (urease) beta subunit
ms3259c	COG0831	UreA, Urea amidohydrolase (urease) gamma subunit
ms3260c	COG2050	PaaI, Uncharacterized protein, possibly involved in aromatic compounds catabolism
Ms3262	COG3682	Predicted transcriptional regulator
ms3263	COG0501	HtpX, Zn-dependent protease with chaperone function
ms4344c	COG0147	TrpE, Anthranilate/para-aminobenzoate synthases component I
ms6656	COG1528	Ftn, Ferritin-like protein
ms7326c	COG0614	FepB, ABC-type Fe3+-hydroxamate transport system, periplasmic component
ms7417	COG1028	FabG, Dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases)
ms7328	COG0853	PanD, Aspartate 1-decarboxylase
ms7329	COG3486	IucD, Lysine/ornithine N-monooxygenase
ms6648c	COG1571	Predicted DNA-binding protein containing a Zn-ribbon domain
ms6649c	COG0406	GpmB, Fructose-2,6-bisphosphatase
ms6650c	COG0077	PheA, Prephenate dehydratase
ms3190	COG2509	Uncharacterized FAD-dependent dehydrogenases
ms4962c	COG2375	ViuB, Siderophore-interacting protein
ms1404c	COG1960	CaiA, Acyl-CoA dehydrogenases
ms1405c	COG0318	CaiC, Acyl-CoA synthetases (AMP-forming)/AMP-acid ligases II
ms1406c	COG0236	AcpP, Acyl carrier protein
ms1402	COG1670	RimL, Acetyltransferases, including N-acetylases of ribosomal proteins
ms0832c	COG2009	SdhC, Succinate dehydrogenase/fumarate reductase, cytochromes b subunit
ms1554c	COG2879	Uncharacterized small protein
ms1555c	COG1966	CstA, Carbon starvation protein, predicted membrane protein
ms1556c	COG2207	AraC-type DNA-binding domain-containing proteins
ms7314	COG2375	ViuB, Siderophore-interacting protein
ms7315	COG0609	FepD, ABC-type Fe3+-siderophore transport system, permease component
ms5673	COG1960	CaiA, Acyl-CoA dehydrogenases
ms5426c	COG2951	MltB, Membrane-bound lytic murein transglycosylase B
ms5427	COG0672	FTR1, High-affinity Fe2+/Pb2+ permease
ms5428	COG2822	Predicted periplasmic lipoprotein involved in iron transport
ms5429	COG2837	Predicted iron-dependent peroxidase

Table 7.9. Contnd.

Gene	COG No.	Product
ms7321	COG3251	Uncharacterized protein conserved in bacteria
ms7322	COG1132	MdlB, ABC-type multidrug transport system, ATPase and permease
ms7323	COG1132	MdlB, ABC-type multidrug transport system, ATPase and permease
ms7324	COG1020	EntF, Non-ribosomal peptide synthetase modules
ms7325	COG1020	EntF, Non-ribosomal peptide synthetase modules

Note: Genes that are part of an operon are together

Table 7.10. Predicted IdeR binding sites in *M. leprae*

Position	Score	Binding sie	Gene	Synonym	Product
-213	4.91319	ATAGGCAAGGCTGCCCTAA	-	ML2063	Possible Regulator
-269	4.8888	TTAGTGGAGTCTAACCTAA	bfrA	ML2038	Bacterioferritin
-208	4.57039	CGAGGTTAGACTAAGCTAA	hisE	ML130	Phosphoribosyl-ATP Pyrophosphatase
-6	4.49503	GTAGGCCAGTCTATCGTTA	murB	ML2447	UDP-N-Acetylenolpyruvoylglucosamine Reductase
-243	4.292	GTATCCTAGGCTAGCCTGG	fdxA	ML1489	Ferredoxin (Fe-S Co-Factor)
-69	4.25015	CCAGACCAGGCTACCCTAG	-	ML0453	Conserved Hypothetical Protein
-69	4.22436	GGATGACAGGCTGACCTGA	glpK	ML2314	Glycerol Kinase
-78	4.19852	TTACGCTAGTCTCAAGTAA	-	ML1689	Possible Hydrolase
-361	4.14559	TTATACAAGTCTTTGCTTT	ilvG	ML2083	Acetolactate Synthase II
-130	4.13935	CTAGGGAAGGGTACCCTCG	-	ML0591	Putative Membrane Protein
-158	4.12623	CTCGCGGAGCCTTCGCTGA	-	ML2158	Hypothetical Protein
7	4.12616	TTAGCTTACGCAATGCTAA	-	ML2537	Conserved Hypothetical Protein

Table 7.11. Predicted IdeR regulated operons in *M. leprae*

Gene	Synonym	COG No.	Product
-	ML2063	COG3682	Possible Regulator
-	ML2064	COG0501	Integral Membrane Protein
-	ML2035	-	Amycolatopsis Mediterranei U32 Nacd Nitrite Reductase
bfrA	ML2038	COG2193	Bacterioferritin
hisE	ML1309	COG0140	Phosphoribosyl-ATP Pyrophosphatase
hisG	ML1310	COG0040	ATP Phosphoribosyltransferase
-	ML2446	COG1376	Possible Lipoprotein
murB	ML2447	COG0812	UDP-N-Acetylenolpyruvoylglucosamine Reductase
	ML1488	COG0436	Putative Aspartate Aminotransferase [EC:2.6.1.1]
fdxA	ML1489	COG1146	Ferredoxin
-	ML0450	COG0214	Putative Pyridoxine Biosynthesis Protein
-	ML0451	COG0494	NTP Pyrophosphohydrolases Including
-	ML0452	COG0438	Putative Glycosyltransferase
-	ML0453	COG1560	Phosphatidylinositol Synthase Pgsa
glpK	ML2314	COG0554	Glycerol Kinase
gltS	ML1688	COG0008	Glutamyl-Trna Synthase
-	ML1689	COG0179	Possible Hydrolase
ilvG	ML2083	COG0028	Acetolactate Synthase II
-	ML0589	COG0842	ABC-Type Multidrug Transport System
-	ML0590	COG1131	ABC-Type Multidrug Transport System,
-	ML0591	-	Putative Membrane Protein
-	ML2534	-	PE-Family Protein
-	ML2535	COG1674	DNA Segregation Atpase Ftsk/Spoiiie
-	ML2536	-	Conserved Membrane Protein
-	ML2537	COG0464	Atpase, AAA Family

Note: Genes that are part of an operon are together

Table 7.12. Predicted IdeR binding sites in *N. farcinia*

Score	Position	Binding site	Gene	Product
6.59557	-79	TTAGTATAGGCTAGCCTTA	nfa7620	Putative N6-Hydroxylysine Acetyltransferase
6.41217	-97	TTAGGTAAGGCTTGCTTAA	nfa48610	Hypothetical Protein
6.36829	-107	TTAGGTAAACCTAAGCTAA	nfa1320	Hypothetical Protein
6.3235	-169	TTAAGCAAGCCTTACCTAA	nfa48600	Hypothetical Protein
6.2764	-71	ATAGGTTAGCCTTGGCTGA	nfa7720	Putative Ferric Nocobactin-Binding Protein
6.2734	-84	TAAGGCTAGCCTATACTAA	nfa7630	Putative Thioesterase
6.25985	-71	TTAGGCAATACTATCCTCA	nfa1270	Putative Ferritin Family Protein
6.22134	-187	TTAGGTAAGCCTGTCCTAT	nfa25230	Putative Thioesterase
6.17735	-32	TTAGCTTAGGCTAAGTTGA	nfa53610	Hypothetical Protein
6.09643	-48	TTAGCTTAGGTTTACCTAA	nfa1310	Hypothetical Protein
6.04411	-99	ATAGGTAAGGCTAACTTAT	nfa7500	Hypothetical Protein
5.95418	-59	CGAGGTAATGCTAACCTTA	nfa6190	Putative Hydroxybenzoate Synthase
5.89513	-90	ATAAGTTAGCCTTACCTAT	nfa7510	Putative ABC Transporter
5.88748	-83	TTTGCATAGGCTTACCTTA	nfa7600	Hypothetical Protein
5.86966	-154	AAAGGTTAAGCTGACCTAA	nfa6180	Putative Glycerol-3-Phosphate Acyltransferase
5.8606	-136	TCAGCCAAGGCTAACCTAT	nfa7730	Putative ABC Transporter ATP-Binding Protein
5.80012	-67	TCAACTTAGCCTAAGCTAA	nfa53620	Putative Ferredoxin Reductase
5.73381	-46	CGAGGTGACCCTAACCTGA	nfa49080	Putative Transcriptional Regulator
5.71038	-50	TTAGGTGACCCTGACCTCG	nfa31410	Putative Transcriptional Regulator
5.68354	-68	CCATGTTAGCCTCCCCTAA	nfa31810	Hypothetical Protein
5.67388	-82	CTGGGTTAGCCTTCGCTGA	nfa9830	Putative Transcriptional Regulator
5.67135	-119	GTAGACCAGTCTACGCTCG	nfa34270	Hypothetical Protein
5.59445	-129	CTATGTCAGCCTGCGCTAC	nfa44830	Putative Helicase
5.56573	-91	TTAATCAAGATTAACCTGA	nfa50370	Putative Acetyl-Coa Carboxylase Beta Subunit
5.55997	-111	GAAGGATAGCCTGACCTGG	nfa21170	Hypothetical Protein
5.55918	-152	ATTGGTTAGCGTAACCTAA	nfa47590	Putative GTP-Binding Elongation Factor
5.55147	-87	AGAGGGAACGCTGTCCTCA	nfa26280	Hypothetical Protein
5.50272	-63	ATAGGTTAGGCTTACCAGA	nfa25210	Putative Iron Transporter
5.49537	-144	AGAAGTTAGGCTGAGCTCA	nfa11330	Putative Cation Transporter
5.4625	-24	TGTCGTTAGGCTAACCTTA	nfa7590	Putative Siderophore-Interacting Protein
5.45336	-160	GTAGCAAAGGTTAGCCTGC	nfa9670	Putative Dihydrolipoamide Dehydrogenase
5.44706	-68	GTTGGTTAGGCAACCCTTA	nfa23710	Hypothetical Protein
5.44295	17	TTACATCAGGCTGGGTTCA	nfa38220	Hypothetical Protein
5.42779	-60	CGAGGTCAGGGTCACCTAA	nfa31420	Putative ABC Transporter
5.42257	-24	ATACCAAACGCTATCCTGG	nfa44690	Putative Deaminase
5.41275	-47	TCAGGTTAATCTTGATTAA	nfa50360	Putative Transcriptional Regulator
5.40381	0	TTGGCTAACACTTTCCTGA	nfa39240	Hypothetical Protein
5.39807	-188	CTACCAGAGCCTGTCCTTC	nfa21880	Putative Ethanolamine Ammonia-Lyase Small Subunit
5.37852	-64	TCTGGTAAGCCTAACCTAT	nfa25220	Hypothetical Protein
5.36485	-282	CTATTGAATTCTAGCTTCA	nfa55980	Hypothetical Protein
5.36485	-75	CTATTGAATTCTAGCTTCA	nfa55990	Hypothetical Protein
5.36121	-43	TTAGGTCAAGATGACGTAA	nfa44200	Hypothetical Protein
5.3576	-91	TTACGTCATCTTGACCTAA	nfa44190	Hypothetical Protein
5.33958	-64	TTTGGTTAGGCAACCCTAT	nfa6210	Putative Short Chain Dehydrogenase
5.3297	-102	TAAGGTAAGCCTATGCAAA	nfa7610	Putative Lysine-N-Oxygenase

Table 7.13. Predicted IdeR regulated operons in *N. farcinia*

Synonym	Gene	Product
nfa7620	*nbtH*	Putative N6-Hydroxylysine Acetyltransferase
nfa48610	-	Hypothetical Protein
nfa48620	-	Hypothetical Protein
nfa1320	-	Hypothetical Protein
nfa1330	-	Putative Prephenate Dehydratase
nfa1340	-	Putative Phosphoglycerate Mutase
nfa48600	-	Hypothetical Protein
nfa7720	-	Putative Ferric Nocobactin-Binding Protein
nfa7630	*nbtA*	Putative Thioesterase
nfa7640	*nbtB*	Putative Polyketide Synthase
nfa7650	*nbtC*	Putative Polyketide Synthase
nfa7660	*nbtD*	Putative Non-Ribosomal Peptide Synthetase
nfa7670	*nbtE*	Putative Non-Ribosomal Peptide Synthetase
nfa7680	*nbtF*	Putative Non-Ribosomal Peptide Synthetase
nfa12700	-	Hypothetical Protein
nfa25230	-	Putative Thioesterase
nfa25240	*ureA*	Putative Urease Gamma Subunit
nfa25250	*ureB*	Putative Urease Beta Subunit
nfa25260	*ureC*	Putative Urease Alpha Subunit
nfa53610	-	Hypothetical Protein
nfa1270	-	Putative Ferritin Family Protein
nfa1280	-	Putative Ferritin Family Protein
nfa1290	-	Putative Transcriptional Regulator
nfa1300	-	Hypothetical Protein
nfa1310	-	Hypothetical Protein
nfa7490	-	Putative RNA Pseudouridylate Synthase
nfa7500	-	Hypothetical Protein
nfa6190	-	Putative Hydroxybenzoate Synthase
nfa7510	-	Putative ABC Transporter
nfa7520	-	Putative ABC Transporter
nfa7590	-	Putative Siderophore-Interacting Protein
nfa7600	-	Hypothetical Protein
nfa6170	-	Putative 1-Acylglycerol-3-Phosphate O-Acyltransferase
nfa6180	*plsB*	Putative Glycerol-3-Phosphate Acyltransferase
nfa7730	-	Putative ABC Transporter ATP-Binding Protein
nfa53620	-	Putative Ferredoxin Reductase
nfa53630	-	Hypothetical Protein
nfa49070	-	Hypothetical Protein
nfa49080	-	Putative Transcriptional Regulator
nfa31410	-	Putative Transcriptional Regulator
nfa31810	-	Hypothetical Protein

Table 7.13. Contnd.

Synonym	Gene	Product
nfa9830	-	Putative Transcriptional Regulator
nfa9840	-	Putative Aminotransferase
nfa34270	-	Hypothetical Protein
nfa50370	-	Putative Acetyl-Coa Carboxylase Beta Subunit
nfa50380	-	Putative Acetyl-Coa Carboxylase Alpha Subunit
nfa50390	*fadE43*	Putative Acyl-Coa Dehydrogenase
nfa50400	-	Hypothetical Protein
nfa50410	-	Putative Citrate Lyase Beta Subunit
nfa50420	-	Putative Acyl-Coa Synthetase
nfa21170	-	Hypothetical Protein
nfa26280	-	Hypothetical Protein
nfa26290	-	Hypothetical Protein
nfa25190	-	Putative Iron Transporter ATP-Binding Protein
nfa25200	-	Putative Iron Transporter
nfa25210	-	Putative Iron Transporter
nfa11310	-	Hypothetical Protein
nfa11320	-	Putative Transcriptional Regulator
nfa11330	-	Putative Cation Transporter
nfa23710	-	Hypothetical Protein
nfa38220	-	Hypothetical Protein
nfa31420	-	Putative ABC Transporter
nfa31430	-	Putative ABC Transporter
nfa31440	-	Putative Transporter
nfa44690	-	Putative Deaminase
nfa50360	-	Putative Transcriptional Regulator
nfa25220	-	Hypothetical Protein
nfa55980	-	Hypothetical Protein
nfa55990	-	Hypothetical Protein
nfa44200	-	Hypothetical Protein
nfa44210	-	Hypothetical Protein
nfa44180	-	Hypothetical Protein
nfa44190	-	Hypothetical Protein
nfa6210	-	Putative Short Chain Dehydrogenase
nfa7610	*nbtG*	Putative Lysine-N-Oxygenase

Note: Genes that are part of an operon are together

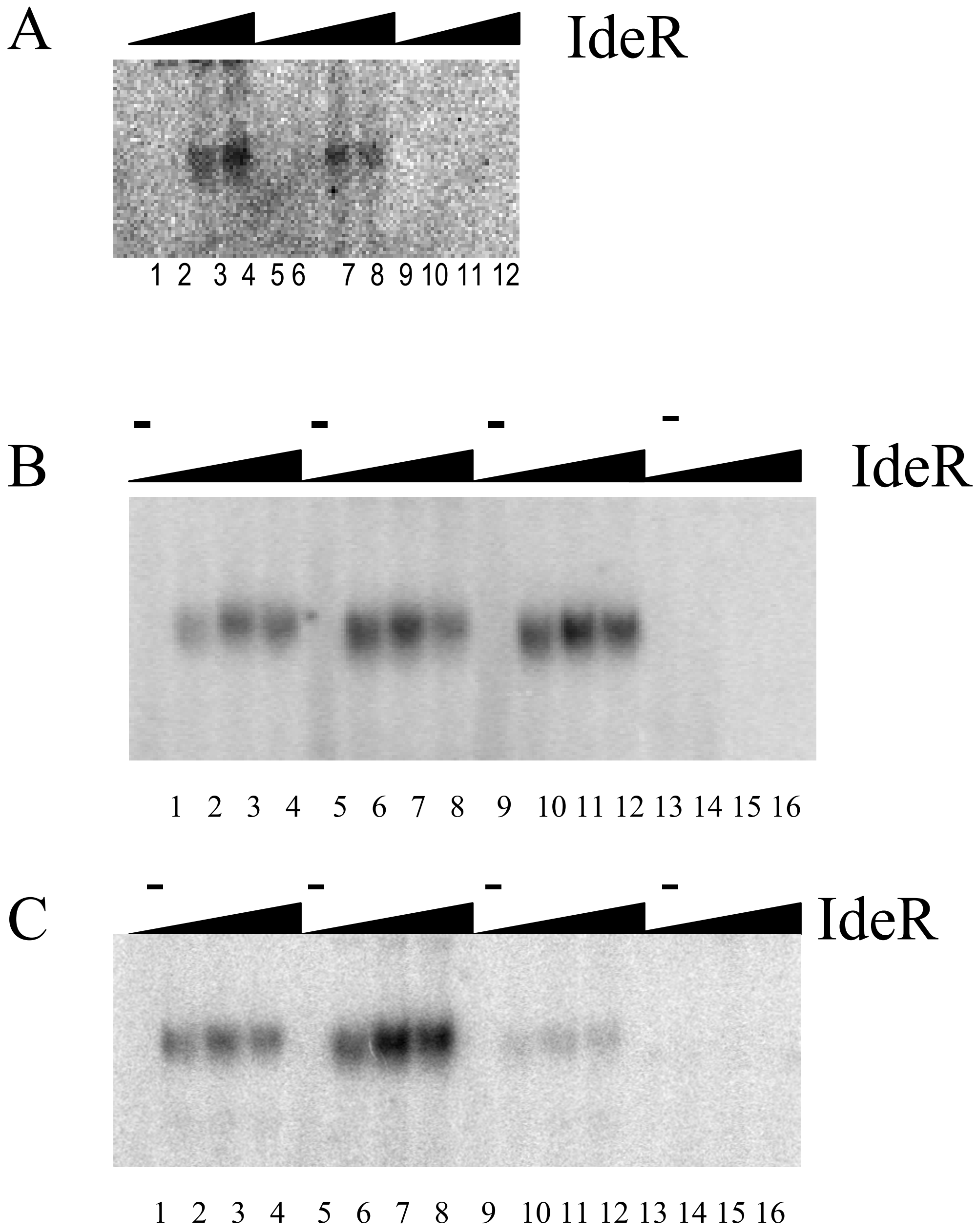


Figure 7.2: IdeR binds to the predicted IdeR binding regulatory motifs in *M. smegmatis*

The lanes indicated by (-) have the probe alone without IdeR. Increasing concentration of IdeR was added to 32P-labelled DNA probes in the presence of 200 μM Ni+ and complexes were resolved on a 7% Tris-borate polyacrylamide gel. Binding conditions and gel electrophoresis are described in Materials and Methods.

1. Radiolabeled DtxR binding motif (lane 1-4), Radiolabeled 7326 motif (lane 5-8), Radiolabeled motif without IdeR binding site (lane 9-12), IdeR was added in increasing concentration from 0 to 10 picomoles. No binding was shown till 1 picomole
2. Radiolabeled 4962 motif (lane 1-4), Radiolabeled DtxR motif (lane 5-8), Radiolabeled 3260 motif (lane 9-12), Radiolabeled motif without IdeR binding site (lane13-16)
3. Radiolabeled 7314 motif (lane 1-4), Radiolabeled DtxR binding motif (lane 5-8), Radiolabeled 6652 motif (lane 9-12), Radiolabeled motif without IdeR binding site(lane13-16). IdeR was added in increasing concentration from 5 picomoles to 20 picomoles (B and C)

7.3 Discussion

7.3.1 Conserved IdeR dependent genes in *Mycobacterium* species

Table 7.14 shows the distribution of orthlogues genes of IdeR regulated genes belonging to different functional category in across the *Mycobacterium* species. Here I discuss the most frequent genes across the IdeR regulons of *Mycobacterium* species that could play an important role in adaptation to the iron levels in different environments.

Orthlogues of the *trpE2* (Rv2386c), *pheA* (Rv3838c) and Rv3837c in other *Mycobacterium* species are predicted to regulate by IdeR. Presence of these genes across the IdeR regulon of *Mycobacterium* species suggests an important role of their cognate gene products in iron metabolism. The gene *trpE2* has been predicted to code for an isochorismate synthase that can catalyze the conversion of chorismate to isochorismate, the precursor for salicylate (Quadri *et al*., 1998). Later its orthologue ybtS in *Yersinia enterocolitica* has been suggested to catalyze formation of salicylate from chorismate (Pelludat *et al*., 2003). The gene *pheA* codes for a predicted prephenate dehydratase, which catalyzes a committed, step in the biosynthesis of the aromatic amino acid phenylalanine. The gene *pheA* and other gene Rv3837c, which is predicted to encode 2, 3-PDG dependent phosphoglycerate mutase, belong to the same operon and are likely to be involved in similar function.

The genes that code for an iron storage protein (BfrA), siderophore biosynthesis protein (MbtB, MbtA) and siderophore transport system are also conserved across the IdeR regulon of *Mycobacterium* species. The mycobactin biosynthesis operon is conserved across the IdeR regulon of pathogenic *Mycobacterium* species but not in nonpathogenic *M. smegamtis*. In *M. smegmatis* exochelin biosynthesis locus shows strong predicted IdeR box, but another locus, which is equivalent to the Mycobactin locus of *M. tuberculosis*, did not has IdeR binding site.

Table 7.14. Distribution of orthologues of IdeR regulated genes across the actinobacteria

Gene	*Mtub*	*Mbov*	*Mavi*	*Mmar*	*Mlep*	*Msme*	*Nfar*
Aromatic amino acid metabolism							
pheA	Rv3838c	Mb3868c	MAP0193	MM3689	*ML0078	ms6650c	*nfa1330
fbp	Rv3837c	Mb3867c	MAP0194	MM3690	*ML0079	ms6649c	*nfa1340
hisE	Rv2122c	Mb2146c	*MAP1847c	MM0190	ML1309	*ms3924c	*nfa31860
hisG	Rv2121c	Mb2145c	*MAP1846c	MM0191	ML1310	*ms3923c	*nfa31850
trpE2	Rv2386c	Mb2407c	MAP2205c	MM5650c	-	ms4344c	nfa6190
paaI	Rv1847	Mb1878	MAP1560	MM0625c	-	ms3260c	nfa25230
Urease							
ureA	Rv1848	Mb1879	-	MM0624c	-	ms3259c	nfa25240
ureB	Rv1849	Mb1880	-	MM0623c	-	ms3258c	nfa25250
ureC	Rv1850	Mb1881	-	MM0622c	-	*ms3256c	nfa25260
ureF	Rv1851	Mb1882	-	MM0621c	-	*ms3255c	*nfa25390
uerG	Rv1852	Mb1883	-	MM0620c	-	*ms3254c	*nfa25400
ureD	Rv1853	Mb1884	-	MM0619c	-	*ms3252c	*nfa25410
Fatty acid metabolism							
fadD	Rv1344	Mb1379	MAP1555c	*MM3394c	-	ms1406c	-
fadE	Rv1345	Mb1380	MAP1554c	-	-	ms1405c	-
fadB	Rv1346	Mb1381	MAP1553c	-	-	ms1404c	-
-	Rv1347	Mb1382c	*MAP3149c	*MM2072	-	ms1402	nfa7620
Cell wall biosynthesis							
murB	Rv0482	Mb0492	MAP3975	*MM3892c	ML2447	-	*nfa51970
Siderophore biosynthesis							
mbtJ	Rv2385	Mb2406	MAP2197	MM5643	-	-	-
mbtA	Rv2384	Mb2405	MAP2178	MM5642	-	ms4331	*nfa6200
mbtB	Rv2383c	Mb2404c	MAP2177c	MM5631c	-	ms4330c	nfa7680
mbtC	Rv2382c	Mb2403c	MAP2175c	MM5636	-	*ms4326c	nfa7640
mbtD	Rv2381c	Mb2402c	MAP2174c	MM5637	-	*ms4325c	nfa7650
mbtE	Rv2380c	Mb2401c	MAP2173c	MM5633c	-	*ms4324c	nfa7660
mbtF	Rv2379c	Mb2400c	MAP2171c	MM5632c	-	*ms4323c	nfa7670
mbtG	Rv2378c	Mb2399c	MAP2170c	MM5640c	-	*ms4321c	nfa7610
mbtH	Rv2377c	Mb2398c	MAP1872c	*MM0115	-	*ms4320c	*nfa5500
Siderophore transport							
-	Rv1348	Mb1383	MAP2414c	MM6014c	-	ms6824c	nfa7510
-	Rv1349	Mb1384	MAP2413c	MM6013c	-	ms6822c	nfa7520
Iron storage							
bfrA	Rv1876	Mb1907	MAP1595	MM0578c	ML2038	ms3189c	-
bfrB	Rv3841	Mb3871	-	MM3685c	-	ms6656	nfa1270

The operon containing the genes Rv0282, Rv0283, and Rv0284 is also conserved across the predicted IdeR regulon of *Mycobacterium* species. The gene Rv0282 predicted to code FtsK, a protein implicated to have role in cell division and peptidoglycan synthesis or modification (Begg *et al*., 1995; Daniel *et al*., 1996). The gene Rv0282 codes for a hypothetical protein. The gene Rv0284 code for the protein belonging to the AAA-superfamily of ATPases associated with a wide variety of cellular activities, including membrane fusion, proteolysis, and DNA replication (Frickey *et al*., 2004).

In addition to the above genes that are conserved across the predicted regulon of mycobacterium species are Rv1847 and Rv1846, which were not detected by previous studies (Schmitt *et al*., 1995; Rodriguez *et al*., 1999; Gold *et al*., 2001; Rodriguez et al., 2002). The two genes are divergently transcribed and their cognate orthologues in other Mycobacterium species shows strong predicted IdeR binding site. The gene Rv1876c code for a predicted 4-hydroxy benzoyl coA thioesterase (*PaaI*) and downstream genes to the Rv1846 code for subunits of urease. The urease gene is reported as iron regulated and virulence gene in other bacteria (Badruzzaman *et al*., 2004; Olszewski *et al*., 2004). The genes Rv1846 and Rv1845 are divergently transcribed to the Rv1847 and their cognate orthlogues in other *Mycobacterium* species belongs to same operon. The gene Rv1846 codes for a BlaI family of transcription regulator and the other gene Rv1848 code for BlaR1 family of protein. The two families of proteins together confer resistance to variety of β-lactum antibiotics and widely distributed in pathogenic bacteria. In *Staphylococcus aureas*, BlaR1 family of protein MecR1, present in the cytoplasmic membrane, detects the presence of the β-lactum by means of an extracellular penicillin binding-domain and transmits the signal via a second intracellular zinc metaloprotease signaling domain. Binding of a β-lactum to MecR1 stimulates the autocatalytic conversion of intracellular Zinc metaloprotease signaling domain of MecR1 from an inactive proenzyme to an active protease. The activated form of MecR1 cleaves BlaI family of transcription regulator, MecI and de-represses the transcription of β-lactamase (Hanique *et al*., 2004; Wilke *et al*., 2004).

7.3.2 IdeR regulated genes not present in *M. tuberculosis*, but present in other *Mycobacterium* species

Analysis of genes that are under the control of IdeR in *M. avium, M. bovis, M. marinum and M. smegmatis* reveals novel genes, predicted to be involved in iron transport. The genes, which code for a predicted iron permease, iron transporter and iron dependent peroxidase belong to an operon and well represented in sequenced bacterial genomes. The mycobacterial genes could play a similar role in the oxidase dependent iron transport system in *Candida albicans* and *Saccaromyces cervacia* (Robert *et al*., 1996), but former is a peroxidase dependent iron transport system. The peroxidase dependent iron transport system could have role in peroxide stress defense as well as control of intracellular iron levels.

In addition to the peroxidase dependent transport system, IdeR can also regulate the genes that code for predicted citrate dependent iron transport system (FecR, FecB) siderophore interacting protein (ViuB) and *Mycobacterium* natural-resistance-associated macrophage protein (Mramp). In *Vibrio cholerae*, ViuB is suggested to be a cytoplasmic protein involved in ferric vibriobactin uptake and processing. The protein, Mramp is an orthologue of natural-resistance-associated macrophage protein (Nramp) and competes with later for the same divalent-cations, for intracellular survival of mycobacteria (Agranoff *et al*., 1999).

7.4 Conclusion

Analysis of IdeR regulated genes across the *Mycobacterium* and related organism *N. farcinica* has lead to the identification of conserved iron regulated genes. Genes that code for predicted anthranilate synthase, prephenate dehydratase, 2, 3-PDG dependent phosphoglycerate mutase, 4-hydroxy benzoyl coA thioesterase and antibiotic regulatory system are conserved across predicted IdeR regulons of *Mycobacterium* species, but their role in iron metabolism is yet to be identified. The siderophore, mycobactin a virulent determinant in *M. tuberculosis* is present in all the predicted IdeR regulons of pathogenic

Mycobacterium species but not in nonpathogenic *M. smegamtis*. Analysis of predicted IdeR regulons has also identified several genes that could be involved in iron homeostasis in mycobacteria. A peroxidase dependent iron transport system could be involved in peroxide stress defense as well as control of intracellular iron levels. Citrate dependent iron transport system and siderophore interacting protein could be involved in transport of iron and release of iron from siderophores respectively. *Mycobacterium* natural resistance associated macrophage protein has a role in survival of mycobacteria in phagosome by competing with mammalian natural resistance-associated macrophage for the same divalent-cations.

Chapter 8

Prediction of Regulons in *M. tuberculosis* Genome

Mycobacterium tuberculosis is the causative agent of tuberculosis in humans. It experience wide ranging environmental conditions during the course of infection process. The pathogen enters the alveoli by airborne transmission and is taken up into the resident alveolar macrophages. By escaping phagosome-lysosome fusion, the intracellular bacilli are able to avoid killing and survive under low pH, low nutrients, nitrogen and oxygen stress and general stresses. Inside the macrophages, the pathogen can enter into a dormant stage, where it encounters hypoxia and starvation.

Reprogramming of the complex transcription regulatory network is known to be responsible for adaptation of the pathogen to these diverse environments (Kendall *et al*., 2004). According to COG functional category, *M. tuberculosis* has 180 transcription regulators, 18 two-component systems and 20 sigma factors, which could be the components of transcription regulatory network.

Identification of target genes of these regulators or identifying regulons could be useful to understand the modular nature of transcription regulatory network. In spite of several experimental studies including micro arrays, only few regulons are known in *M. tuberculosis*. The regulons IdeR (Schmitt *et al*., 1995; Rodriguez *et al*., 1999; Gold *et al*., 2001; Rodriguez *et al*., 2002), lexA (Durbach et al., 1997; Brooks *et al*., 2001; Dullaghan *et al*., 2002; Boshoff *et al*., 2003), and DevR (Park *et al*., 2003) are well studied in *M. tuberculosis.*

The availability of genome sequence data for many organisms has lead to the development of a comparative genomics tool, called phylogenetic footprinting to predict the transcription factor binding sites by finding unusually well conserved regions in Orthologous upstream sequences (Bailey and Elkan, 1995; Sandelin *et al*., 2004). The basis for this tool is that the orthologous genes could have similar regulatory signals and the signals will be conserved during the evolution. McCue and coworkers (McCue *et al*., 2002) showed that the selection of upstream sequences from three species is optimal for phylogenetic footprinting. They also showed that number of orthologues, phylogenetic distance, and similarity of habitat are important factors in the selection of species for phylogenetic footprinting.

The orthologous upstream sequences can be completely identical, not identical but show identical regulatory signals and not identical. The first and latter types are not suitable for phylogenetic footprinting. To address this issue optimal similarity between the upstream sequences was computed to select the upstream sequences for phylogenetic footprinting irrespective to phylogenetic relationship of the species.

Two orthologous upstream sequences with optimal similarity to each of the *M. tuberculosis* upstream sequences were selected from other actinobacteria to identify possible regulatory signals in upstream to the transcription units of *M. tuberculosis.* The approach could identify 84% of the known regulatory sites in *M. tuberculosis.* Further, clustering of transcription units by predicted regulatory sites lead to the identification of novel genes clustered along with genes that are part of known regulons.

8.1 Method

Complete genome sequences of *M. tuberculosis H37Rv, M. leprae TN, M. bovis AF2122/97, M. avium subsp. paratuberculosis str. k10, N. farcinica IFM 10152* and *C. diphtheriae* were downloaded from NCBI (National Centre for Biotechnolgy Infomration) ftp site (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/).

Complete genome sequence of *M. marinum* was obtained from the Sanger institute (http://www.sanger.ac.uk/Projects/Microbes/). Unfinished sequence of *M. smegmatis* was obtained from The Institute for Genomic Research, TIGR (http://www.tigr.org/tdb/mdb/mdbinprogress.html). The genome sequences of *M. marinum, M. microti and M. smegmatis* were annotated by GLIMMER software (Delcher *et al*., 1999). Orthologues of *M. tuberculosis* in other genomes was identified by bi-directional best-hit using BLASTP software (Altschul *et al*., 1997).

8.1.1 Determination of cut-off score

In *M. tuberculosis*, there were total of 44 genes known to contain regulatory protein binding sites. Corresponding 44 upstream sequences were extracted from *M. tuberculosis* genome. Matcher from EMBOSS (Rice *et al*., 2000) was used to align each upstream sequence with its orthologous upstream sequences in 10 other actinobacteria. Alignment score was calculated as percent length of the locally aligned segment with respect to the length of the smallest upstream sequence.

Example:

M.tuberculosis atgtgctgctgctgtg**C-CTGCTGCTGCT**gctcgtcgcg
M.marinum actgtatatcgtagca**CGCTGCTGCTGCT**taactacgtag

Length of the conserved segment (LCS) = 13
Length of the smallest upstream sequence (LSU) = 38
Alignment score (S) = LCS*100/LSU = (13*100)/38 = 34.21

Two orthologous upstream sequences with the scores nearest to 10 were selected for phylogentic foot printing. Mean score was calculated by sum of the scores in all 44 orthologous sets divided by total number of score, 88. Similarly other datasets were prepared with the scores nearest to 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90 and 95.

The software MEME was used to identify the conserved sites (Timothy *et al*., ISMB, 1994) in 18 datasets. Both strands of the DNA were searched for the motifs with lengths ranging from 16 to 34. Motifs with palindrome nature and which could be repeated any numbers of times were searched.

The dataset with mean scores 70.6718, 76.2364, 80.5919, 84.0123, 86.4602 and 88.1612 shows highest number of predicted sites matching to the known sites. Mean of 70.6718, 76.2364, 80.5919, 84.0123, 86.4602 and 88.1612, which is 80.59, is considered as the optimal score to select orthologous upstream sequences for phylogenetic footprinting.

8.1.2 Prediction of cis-regulatory elements by Phylogenetic footprinting

M. tuberculosis transcription units were predicted using the method described in Chapter 2. There were total of 2255 transcription units where 900 are poly cistronic units encoding 2571 genes. Orthologues of the first gene in each transcription unit were identified in other Actinoabcteria. Among 2255 *M. tuberculosis genes,* 1855 contain two or more orthologues in other species. Corresponding 1855 upstream sequences up to the length of 300 were extracted from *M. tuberculosis* genome. For each upstream sequence, two orthologous upstream sequences (with 80 % score) were selected from other actinobacteria. The software MEME was used to identify the conserved sites. Among these, sites with length less than 24 or completely identical sequences among the orthologues were excluded.

8.1.3 Clustering of transcription units by cis-regulatory elements

The conserved elements predicted in 1855 different sets of orthologous upstream sequences were used for clustering of 1855 corresponding transcription units from *M. tuberculosis.* The software PROCSE (Erik *et al*., 2002) was used for clustering of *M. tuberculosis* according to the conserved motifs.

8.2 Results

In *M. tuberculosis*, there were total of 44 genes known to contain regulatory protein binding sites (Table 8.1). To determine the optimal alignment score to select orthologous upstream sequence irrespective of phylogenetic relationship, orthologous upstream sequence sets corresponding to the above 44 genes were prepared with each set

Table 8.1: Known regulatory protein binding sites in *M. tuberculosis*

Regulator	Binding site	Gene	Gene product
lexA	CGAACATACTTTCG	Rv0335c	PE
lexA	CGAAAGTATGTTCG	Rv0336	Hypothetical Protein Rv0336
lexA	CGAACATACTTTCG	Rv0515	Hypothetical Protein Rv0515
lexA	AGAACGGTTGTTCG	Rv2578c	DNA Repair Photolyase
lexA	CGAACGATTGTTCG	Rv2594c	RuvC
lexA	CGAACAATCGTTCG	Rv2595	Hypothetical Protein Rv2595
lexA	CAAACATGTGTTCG	Rv2719c	Hypothetical Protein Rv2719c
lexA	CAAACATGTGTTCG	Rv2720	SOS-Response Transcriptional Repressors
lexA	CGAACAGGTGTTCG	Rv2737c	RecA
lexA	CGAACAATTGTTCG	Rv3370c	DNA Polymerase III
lexA	CGAACAATTGTTCG	Rv3371	Hypothetical Protein Rv3371
IdeR	TAAGGCTAGCCTTACCTTG	Rv1519	Hypothetical Protein Rv1519
IdeR	ATAGGCAAGGCTGCCCTAA	Rv1846c	Predicted Transcriptional Regulator
IdeR	ATAGGCAAGGCTGCCCTAA	Rv1847	Hypothetical Protein Rv1847
IdeR	TTAGTGGAGTCTAACCTAA	Rv1876	Bacterioferritin (Cytochrome B1)
IdeR	CTAGGGTAGCCTAACCTAT	Rv2122c	HisI
IdeR	CTAGGGTAGCCTAACCTAT	Rv2123	PPE
IdeR	TTAGCACAGGCTGCCCTAA	Rv2383c	Non-Ribosomal Peptide Synthetase Modules
IdeR	TTAGCACAGGCTGCCCTAA	Rv2384	Peptide Arylation Enzymes
IdeR	TAAATGTAGCCTAACCTAC	Rv2386c	Anthranilate/Para-Aminobenzoate Synthases Component I
IdeR	TTAACTTAGGCTTACCTAA	Rv3838c	PheA
IdeR	TTAACTTAGGCTTACCTAA	Rv3839	Hypothetical Protein Rv3839
IdeR	CTAGGAAAGCCTTTCCTGA	Rv3841	BfrB
devR	GTGGGGCCGAAGGTCCTCAA	Rv0574c	Putative Enzyme Of Poly-Gamma-Glutamate Biosynthesis
devR	TAAGGGACTTTCGCCCCTTC	Rv1733c	Hypothetical Protein Rv1733c
devR	TTAGGGCCGGAAGTCCCCAA	Rv1738	Hypothetical Protein Rv1738
devR	GCCGGGACTTCAGGCCCTAT	Rv1738	Hypothetical Protein Rv1738
devR	GTAGGGCATAAAGTCTCTAA	Rv1813c	Hypothetical Protein Rv1813c
devR	CATGAGGCTTTAGTCCCCAA	Rv2005c	Hypothetical Protein Rv2005c
devR	CATGAGGCTTTAGTCCCCAA	Rv2006	Trehalose And Maltose Hydrolases
devR	TCGGGGACTTCTGTCCCTAG	Rv2031c	HspX
devR	TCGGGGACTTCTGTCCCTAG	Rv2032	Hypothetical Protein Rv2032
devR	CACGGGTCAAACGACCCTAG	Rv2626c	Hypothetical Protein Rv2626c
devR	GGCGGGACGTAAGTCCCTAA	Rv2627c	Hypothetical Protein Rv2627c
devR	GGCGGGACGTAAGTCCCTAA	Rv2628	Hypothetical Protein Rv2628
devR	GTGGGGACCAACGCCCCTGG	Rv3134c	Hypothetical Protein Rv3134c
devR	GTAGGGCCCAGTGCCCCAGT	Rv3825c	Pks2

containing three sequences, of which one was from *M. tuberculosis* and rest were selected from other actinobacteria. The software MEME is used to identify the conserved sites in 18 orthologous upstream sequence datasets.

As shown in Table 8.2 the dataset with mean scores 70.6718, 76.2364, 80.5919, 84.0123, 86.4602 and 88.1612 shows highest number of predicted sites (37), matching to the known sites. Mean of 70.6718, 76.2364, 80.5919, 84.0123, 86.4602 and 88.1612, which is 80.59, was considered as the optimal score to select orthologous upstream sequences for phylogentic footprinting. Table 8.3 shows the predicted sites in upstream sequences of *M. tuberculosis* genes, which are known to contain binding sites of DevR, IdeR and LexA.

There are 2255 predicted transcription units in *M. tuberculosis* where 900 are poly cistronic units encoding 2571 genes. Since the sequence, upstream to the first gene of the transcription unit contains the regulatory sequence, its orthologous upstream sequences were used for prediction of regulatory elements by phylogenetic footprinting. Among 2255 genes, which could have regulatory signals in its upstream sequence, 1855 contain two or more orthologues in other *Mycobacterium* species.

The software MEME was used to identify the conserved sites. Conserved motifs along with the flanking sequences with length less than 24 nucleotides and the motifs along with flanking sequences having 100% similarity were excluded. Finally, phylogenetically conserved motifs were selected in upstream sequences of 1803 transcription units. A searchable database of these motifs has been generated which could be used in identifying cognate transcription regulators in future. Phylogenetially conserved motifs in all the upstream sequences of *M. tuberculosis* can be accessed at the CDFD website (http://www.cdfd.org.in/predictregulon/mtubregulon/motifs/).

The conserved elements predicted in 1855 different sets of orthologous upstream sequences were used for clustering of 1855 corresponding genes from *M. tuberculosis.* The conserved elements either include or substantially overlap a set of

Table 8.2: Statistics of sites predicted in *M. tuberculosis*

Score	Mean score	No. of sites predicted	Percent of total sites
10	58.1627	32	72%
15	58.6794	32	72%
20	58.6794	32	72%
25	58.1627	32	72%
30	58.6794	32	72%
35	59.4555	33	72%
40	60.3382	33	72%
45	61.7047	34	72%
50	63.2552	34	75%
55	65.6979	35	75%
60	**70.6718**	**37**	**84%**
65	**76.2364**	**37**	**84%**
70	**80.5919**	**37**	**84%**
75	**84.0123**	**37**	**84%**
80	**86.4602**	**37**	**84%**
85	**88.1612**	**37**	**84%**
90	89.6342	35	75%
95	90.4864	35	75%

Table 8.3: Conserved elements matching to the known regulatory sites of *M. tuberculosis*

Regulator	Known binding site	Gene	Conserved element
lexA	CGAACATACTTTCG	Rv0335c	cccgcacctgat**CGAACATACTTTCG**atactacca
lexA	CGAAAGTATGTTCG	Rv0336	cgatggtagtat**CGAAAGTATGTTCG**atcaggtgcggg
lexA	CGAACATACTTTCG	Rv0515	cccgcacctgat**CGAACATACTTTCG**atactaccagcc
lexA	AGAACGGTTGTTCG	Rv2578c	tgacaaagtat**AGAACGGTTGTTCG**aataatgg
lexA	CGAACGATTGTTCG	Rv2594c	cgctagcgtat**CGAACGATTGTTCG**gaaatggctga
lexA	CGAACAATCGTTCG	Rv2595	cctcagccatttc**CGAACAATCGTTCG**atacgctagcgga
lexA	CAAACATGTGTTCG	Rv2719c	gcaccaagaat**CAAACATGTGTTCG**acaggcgtgtt
lexA	CAAACATGTGTTCG	Rv2720	gcaccaagaat**CAAACATGTGTTCG**acaggcgtgtt
lexA	CGAACAGGTGTTCG	Rv2737c	gtcacacttgaat**CGAACAGGTGTTCG**gctactgtggtga
lexA	CGAACAATTGTTCG	Rv3370c	aactgcgctgtat**CGAACAATTGTTCG**atatactgtggaa
lexA	CGAACAATTGTTCG	Rv3371	aactgcgctgtat**CGAACAATTGTTCG**atatactgtggaa
IdeR	TAAGGCTAGCCTTACCTTG	Rv1519	acgcggacgct**TAAGGCTAGCCTTACCTTG**taaaaa
IdeR	ATAGGCAAGGCTGCCCTAA	Rv1846c	cgacgaagtaatg**ATAGGCAAGGCTGCCCTAA**tttagcaagcgtt
IdeR	ATAGGCAAGGCTGCCCTAA	Rv1847	gaagtaatg**ATAGGCAAGGCTGCCCTAA**tttagcaag
IdeR	TTAGTGGAGTCTAACCTAA	Rv1876	cctaagctga**TTAGTGGAGTCTAACCTAA**caatgacccg
IdeR	ATAGGTTAGGCTACCCTAG	Rv2122c	cctaat**ATAGGTTAGGCTACCCTAG**ttattcctgtg
IdeR	CTAGGGTAGCCTAACCTAT	Rv2123	cacaggaataa**CTAGGGTAGCCTAACCTAT**attagg
IdeR	TTAGCACAGGCTGCCCTAA	Rv2383c	ccctcccctg**TTAGCACAGGCTGCCCTAA**ttttagtggt
IdeR	TTAGCACAGGCTGCCCTAA	Rv2384	ccctcccctg**TTAGCACAGGCTGCCCTAA**ttttagtggt
IdeR	GTAGGTTAGGCTACATTTA	Rv2386c	acccattaa**GTAGGTTAGGCTACATTTA**ctagc
IdeR	TTAACTTAGGCTTACCTAA	Rv3838c	ccagaccgtgcattag**TTAACTTAGGCTTACCTAA**a
IdeR	TTAGGTAAGCCTAAGTTAA	Rv3839	tccacgacctcctgtgt**TTAGGTAAGCCTAAGTTAA**
IdeR	TTAACTTAGGCTTACCTAA	Rv3841	cgtgcattag**TTAACTTAGGCTTACCTAA**acacaggagg
IdeR	TCAGGAAAGGCTTTCCTAG	Rv1519	gaaggcaatacttac**TCAGGAAAGGCTTTCCTAG**ttaccaca
devR	GTGGGGCCGAAGGTCCTCAA	Rv0574c	tcgtgg**GTGGGGCCGAAGGTCCTCAA**gaccgcgcccaaaggtcac
devR	TAAGGGACTTTCGCCCCTTC	Rv1733c	ttgtcgga**TAAGGGACTTTCGCCCCTTC**ccgcctgc
devR	TTAGGGCCGGAAGTCCCCAA	Rv1738	ccggctcag**TTAGGGCCGGAAGTCCCCAA**tgtggcaga
devR	GCCGGGACTTCAGGCCCTAT	Rv1738	accccagtg**GCCGGGACTTCAGGCCCTAT**cggagggct
devR	GTAGGGCATAAAGTCTCTAA	Rv1813c	tatacctgacccgg**GTAGGGCATAAAGTCTCTAA**cag
devR	CATGAGGCTTTAGTCCCCAA	Rv2005c	agtcaccggt**CATGAGGCTTTAGTCCCCAA**tcggacggccaa
devR	TTGGGGACTAAAGCCTCATG	Rv2006	gccgtccga**TTGGGGACTAAAGCCTCATG**accggtgactgtcccg
devR	TCGGGGACTTCTGTCCCTAG	Rv2031c	cccgcgct**TCGGGGACTTCTGTCCCTAG**ccctggcc
devR	TCGGGGACTTCTGTCCCTAG	Rv2032	cccgcgct**TCGGGGACTTCTGTCCCTAG**ccctggcc
devR	CACGGGTCAAACGACCCTAG	Rv2626c	ccgcggcc**CACGGGTCAAACGACCCTAG**tgttcgct
devR	TTAGGGACTTACGTCCCGCC	Rv2627c	accgcgtgcggaacgacgcg**TTAGGGACTTACGTCCCGCC**ggaagtc
devR	GGCGGGACGTAAGTCCCTAA	Rv2628	tgacttcc**GGCGGGACGTAAGTCCCTAA**cgcgtcgt
devR	GTAGGGCCCAGTGCCCCAGT	Rv3134c	acaaaccgaa**GTAGGGCCCAGTGCCCCAGT**agcacagccgcttagaa

Note: Subsequence of conserved element matching to the known site was shown in uppercase and bold.

regulatory protein binding sites with mean length of 24. The software PROCSE is used for clustering of *M. tuberculosis* transcription units according to the phylogenetically conserved elements. There were 593 clusters with the cluster size ranging from two to seven. Members of these clusters are likely to be part of same regulon. The data is accessible at CDFD (Center for DNA Fingerprinting and Diagnostics) web site (http://www.cdfd.org.in/predictregulon/mtubregulon/clusters/. Table 8.4 shows the clusters containing the genes that are part of IdeR, LexA and DevR regulons in *M. tuberculosis.* These clusters also contain the genes that are not part of any of the known regulon. The genes Rv1846 and Rv1847, which are clustered along with the genes of IdeR regulon, are shown to contain IdeR binding site (Prakash *et al*., 2005).

Table 8.5 shows a cluster containing a transcription regulator and its target genes. The cluster contains a transcription regulator, which is homologous to the PhoU of *E. coli* and a phosphate transport system (Aguena *et al*., 2002). PhoU in *E. coli* is a repressor of high affinity phosphate uptake and under phosphate excess PhoU down-regulates the *pho* regulon.

Table 8.4. Clusters containing the genes of known regulons

Regulator	Synonym	Gene	COG No.	Product
LexA	Rv0335c	*PE*	-	PE
LexA	Rv2720	*lexA*	COG1974	SOS-response transcriptional repressors (RecA-mediated autopeptidases)
LexA	Rv2737c	*recA*	COG1372	RecA
LexA	Rv3371	-	-	Hypothetical Protein Rv3371
	Rv0699	-	-	Hypothetical Protein Rv0699
LexA	Rv3370c	*dnaE2*	COG0587	DNA Polymerase III
LexA	Rv2719c	-	-	Hypothetical Protein Rv2719c
	Rv1219c	-	COG1309	Hypothetical Protein Rv1219c
LexA	Rv2578c	*SplB*	COG1533	DNA Repair Photolyase
LexA	Rv2595	-	COG2002	Hypothetical Protein Rv2595
	Rv1271c	-	-	Hypothetical Protein Rv1271c
	Rv1745c	*Idi*	COG1443	Isopentenyldiphosphate Isomerase
LexA	Rv2594c	*ruvC*	COG0817	RuvC
	Rv2695	-	COG0596	Hypothetical Protein Rv2695
LexA	Rv0515	-	-	Hypothetical Protein Rv0515
	Rv1002c	*PMT1*	COG1928	Dolichyl-Phosphate-Mannose--Protein O-Mannosyl Transferase
	Rv1011	*IspE*	COG1947	4-Diphosphocytidyl-2C-Methyl-D-Erythritol 2-Phosphate Synthase
LexA	Rv0336	-	-	Hypothetical Protein Rv0336
	Rv1364c	*rsbU*	COG2208	Serine Phosphatase Rsbu
	Rv1927	-	COG3361	Uncharacterized Conserved Protein
	Rv2962c	-	COG1819	Glycosyl Transferases
	Rv0070c	*glyA2*	COG0112	Glycine/Serine Hydroxymethyltransferase
	Rv0332	-	-	Hypothetical Protein Rv0332
	Rv2771c	-	COG0655	Hypothetical Protein Rv2771c
	Rv0092	*ctpA*	COG2217	Cation Transporter, Atpase
IdeR*	Rv1347c	-	COG1670	Hypothetical Protein Rv1347c
IdeR	Rv1519	-	-	Hypothetical Protein Rv1519
IdeR	Rv1847	-	COG2050	Hypothetical Protein Rv1847
IdeR	Rv2122c	*hisI*	COG0140	HisI
IdeR	Rv1846c	-	COG3682	Predicted transcriptional regulator
IdeR	Rv1876	*bfrA*	COG2193	Bacterioferritin (cytochrome b1)
IdeR	Rv2123	*PPE*	-	PPE
	Rv3230c	*Hmp*	COG1018	Flavodoxin reductases (ferredoxin-NADPH reductases) family 1
	Rv0216	*MaoC*	COG2030	Acyl dehydratase
IdeR	Rv2383c	*mbtB*	COG1020	Non-ribosomal peptide synthetase modules and related proteins
	Rv2269c	-	-	Hypothetical protein Rv2269c
IdeR	Rv2384	*mbtA*	COG1021	Peptide arylation enzymes
IdeR*	Rv0451c	*mmpS4*	-	mmpS4
IdeR	Rv2386c	*trpE2*	COG0147	Anthranilate/para-aminobenzoate synthases component I
	Rv2709	-	-	Hypothetical Protein Rv2709
IdeR	Rv3841	*bfrB*	COG1528	BfrB

Table 8.4. Contnd.

Regulator	Synonym	Gene	COG No.	Product
	Rv1918c	*PPE*	-	Ppe
	Rv2590	*fadD9*	COG3320	Putative Dehydrogenase Domain Of Multifunctional Non-Ribosomal Peptide Synthetases And Related Enzymes
IdeR	Rv3838c	*pheA*	COG0077	PheA
	Rv2559c	-	COG2256	Atpase Related To The Helicase Subunit Of The Holliday Junction Resolvase
	Rv2560	-	COG5473	Predicted Integral Membrane Protein
IdeR	Rv3839	-	-	Hypothetical Protein Rv3839
	devR			
	Rv0307c	-	-	Hypothetical Protein Rv0307c
DevR	Rv0574c	*PgsA*	COG2843	Putative Enzyme Of Poly-Gamma-Glutamate Biosynthesis (Capsule Formation)
	Rv2101	*helZ*	COG0553	HelZ
	Rv2988c	*leuC*	COG0065	3-Isopropylmalate Dehydratase Large Subunit
DevR	Rv1733c	-	-	Hypothetical Protein Rv1733c
DevR	Rv1738	-	-	Hypothetical Protein Rv1738
DevR	Rv2627c	-	-	Hypothetical Protein Rv2627c
	Rv0195	-	COG2197	Hypothetical Protein Rv0195
DevR	Rv1813c	-	-	Hypothetical Protein Rv1813c
DevR	Rv2005c	-	COG0589	Hypothetical Protein Rv2005c
DevR	Rv2006	*otsB*	COG1554	Trehalose And Maltose Hydrolases (Possible Phosphorylases)
DevR	Rv2031c	*hspX*	COG0071	HspX
DevR	Rv2032	-	-	Hypothetical Protein Rv2032
	Rv3074	-	-	Hypothetical Protein Rv3074
DevR	Rv2626c	-	COG0517	Hypothetical Protein Rv2626c
	Rv3121	*cypX*	COG2124	Cytochrome P450
	Rv3315c	*cdd*	COG0295	Cytidine Deaminase
	Rv1805c	-	-	Hypothetical Protein Rv1805c
DevR	Rv2628	-	-	Hypothetical Protein Rv2628
DevR	Rv3134c	-	COG0589	Hypothetical Protein Rv3134c
	Rv0393	-	-	Hypothetical Protein Rv0393
DevR	Rv3825c	*pks2*	COG3321	Pks2

Table 8.5. Cluster containing auto-regulatory protein and its target genes

Synonym	Gene	Gene product
Rv0928	*pstS3*	Periplasmic phosphate-binding lipoprotein psts3
Rv0929	*pstC2*	Phosphate-transport integral membrane abc transporter pstc2
Rv0930	*pstA1*	Probable phosphate-transport integral membrane abc transporter
Rv3298c	*lpqC*	Possible esterase lipoprotein lpqc
Rv3299c	*atsB*	Probable arylsulfatase atsb (sulfatase)
Rv3300c	-	Hypothetical protein
Rv3301c	*phoY1*	Probable phosphate-transport system transcriptional regulatory protein
Rv1772	-	Hypothetical protein
Rv3033	-	Hypothetical protein
Rv3035	-	Hypothetical protein
Rv0231	*fadE4*	Probable acyl-coA dehydrogenase fadE4

Note: Genes that are together belong to same operon

Summary

The work described in this thesis reports a systematic approach to predict regulons in bacterial genomes. Initially the operons were predicted and grouped by *cis*-regulatory elements, which were predicted by two approaches 1) Shannon relative entropy, and 2) Phylogenetic footprinting.

Using the first approach, the binding sites of iron responsive DtxR family of transcription regulators and their target genes were identified in species of *Mycobacterium* and *Corynebacterium*. In *C. diptheriae*, novel iron-regulated genes that code for starvation inducible DNA-binding protein, Formamidopyrimidine-DNA glycosylase, sortases and proteins of secretary system were identified. Furthermore conserved iron-regulated genes that could have important role in adaptation to the intracellular iron levels were identified across the genomes of mycobacteria and related organism *N. farcinica*. Using this approach, novel iron regulated genes that code for predicted 4-hydroxy benzoyl co-A thioesterage and an antibiotic resistance regulatory system were identified in *M. tuberculosis* H37Rv.

Using the second approach, *cis*-regulatory elements were predicted upstream to the 1803 predicted transcription units in *M. tuberculosis* H37Rv. The 1803 predicted transcription units were clustered by predicted *cis*-regulatory elements. The genes within the clusters are likely to be part of same regulon.

To conclude, the system developed and described in this thesis will have a far-reaching impact in the post genome era when more and more genome sequences would be made available in literature.

References

Agranoff D, Monahan IM, Mangan JA, Butcher PD, Krishna S. ***Mycobacterium tuberculosis* expresses a novel pH-dependent divalent cation transporter belonging to the Nramp family.** *J Exp Med.*, 1999, **5**:717-724.

Aguena M, Yagil E, Spira B: **Transcriptional analysis of the pst operon of *Escherichia coli.*** *Mol Genet Genomics* 2002, **268:**518-524.

Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, David J. Lipman: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.

Andrews SC: **Iron storage in bacteria.** *Adv Microb Physiol* 1998, **40:**281-351.

Babitzke P, Schaak J, Yakhnin AV, Bevilacqua PC: **Role of RNA structure in transcription attenuation in *Bacillus subtilis*: the trpEDCFBA operon as a model system.** *Methods Enzymol* 2003, 371:392-404.

Badruzzaman M, Matsui H, Fazle Akbar SM, Matsuura B, Onji M. **Mechanism of action of low recurrence of gastritis caused by *Helicobacter pylori* with the type II urease B gene**. *Helicobacter*, 2004, **2**:173-180.

Bailey TL, Elkan C: **The value of prior knowledge in discovering motifs with MEME.** *Proc Int Conf Intell Syst Mol Biol.*1995, **3:**21-29.

Begg KJ, Dewar SJ, Donachie WD. **A new *Escherichia coli* cell division gene, ftsK.** *J Bacteriol.,* 1995, **21**:6211-6222.

Benos PV, Bulyk, ML, Stormo GD: **Additivity in protein-DNA interactions: how good an approximation is it?.** *Nucleic Acids Res* 2002, **30:** 4442-4451.

Billington SJ, Esmay PA, Songer JG, Jost BH: **Identification and role in virulence of putative iron acquisition genes from *Corynebacterium pseudotuberculosis.*** *FEMS Microbiol Lett* 2002, **208:**41-45.

Bockhorst J, Qiu Y, Glasner J, Liu M, Blattner F, Craven M: **Predicting bacterial transcription units using sequence and expression data.** *Bioinformatics* 2003, **19:**i34-i43.

Boshoff HI, Reed MB, Barry, CE, Mizrahi V: **DnaE2 polymerase contributes to in vivo survival and the emergence of drug resistance in *Mycobacterium tuberculosis.*** *Cell*, 2003, **113:**183-193.

Brooks PC, Movahedzadeh F, Davis EO: **Identification of some DNA damage-inducible genes of *Mycobacterium tuberculosis*: apparent lack of correlation with LexA binding.** *J Bacteriol* 2001, **15:**4459-4467.

Butterton JR, Calderwood SB. **Identification, cloning, and sequencing of a gene required for ferric vibriobactin utilization by *Vibrio cholerae*.** *J Bacteriol.,* 1994, **18**:5631-5638.

Castagnetto JM., Hennessy SW, Roberts VA., Getzoff ED, Tainer, JA, Pique ME: **MDB: the Metalloprotein Database and Browser at The Scripps Research Institute.** *Nucleic Acids Res* 2002**, 30:**379-382.

Cerdeno-Tarraga AM, Efstratiou A, Dover LG, Holden MT, Pallen M, Bentley SD, Besra GS, Churcher C, James KD, De Zoysa A, Chillingworth T, Cronin A, Dowd L, Feltwell T, Hamlin N, Holroyd S, Jagels K, Moule S, Quail MA, Rabbinowitsch E, Rutherford KM, Thomson NR, Unwin L, Whitehead S, Barrell BG, Parkhill J: **The complete genome sequence and analysis of *Corynebacterium diphtheriae* NCTC13129.** *Nucleic Acids Res* 2003, **31:**6516-6523.

Chen X, Su Z, Xu Y, Jiang T: **Computational Prediction of Operons in *Synechococcus* sp. WH8102.** *Genome Inform Ser Workshop Genome Inform* 2004,**15:**211-222.

Daniel RA, Williams AM, Errington J. **A complex four-gene operon containing essential cell division gene pbpB in *Bacillus subtilis*.** *J Bacteriol*., 1996, **8**:2343-2450.

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

De Lorenzo V, Neilands JB: **Characterization of iucA and iucC genes of the aerobactin system of plasmid ColV-K30 in *Escherichia coli*.** *J Bacteriol,* 1986*,* **167:**350-355.

Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. **Improved microbial gene identification with GLIMMER.** *Nucleic Acids Res* 1999, **27:**4636-4641.

Drazek ES, Hammack CA, Schmitt MP: ***Corynebacterium diphtheriae* genes required for acquisition of iron from haemin and haemoglobin are homologous to ABC hemin transporters.** *Mol Microbiol* 2000, **36:**68-84.

Dullaghan EM, Brooks PC, Davis EO: **The role of multiple SOS boxes upstream of the *Mycobacterium tuberculosis* lexA gene--identification of a novel DNA-damage-inducible gene.** *Microbiology*, 2002, **148:**3609-3615.

Durbach SI, Andersen SJ, Mizrahi V: **SOS induction in mycobacteria: analysis of the DNA-binding activity of a LexA-like repressor and its role in DNA damage induction of the recA gene from *Mycobacterium smegmatis*.** *Mol Microbiol* 1997, **26:**643-653.

Erik VN, Mihaela Z, Nikolaus R, Eric DS: Probabilistic clustering of sequences: Inferring new bacterial regulons by comparative genomics. *Proceedings of the National Academy of Science* USA, 2002, **99:**7323-7328.

Feese MD, Ingason BP, Goranson-Siekierke J, Holmes RK, Hol WG: **Crystal structure of the iron-dependent regulator from *Mycobacterium tuberculosis* at 2.0-A resolution reveals the Src homology domain 3-like fold and metal binding function of the third domain.** *J Biol Chem* 2001, **276:**5959-66.

Frickey T, Lupas AN. **Phylogenetic analysis of AAA proteins.** *J Struct Biol*., 2004, **1-2:**2-10.

Gaudu P, Weiss B: **Flavodoxin mutants of *Escherichia coli K-12.*** *J Bacteriol* 2000, **182:**1788-1793.

Gold B, Rodriguez GM, Marras SA, Pentecost M, Smith I. **The *Mycobacterium tuberculosis* IdeR is a dual functional regulator that controls transcription of genes involved in iron acquisition, iron storage and survival in macrophages.** *Mol Microbiol.,* 2001, **3:**851-865.

Grundy FJ, Henkin TM: **The S box regulon: a new global transcription termination control system for methionine and cystein biosynthesis genes in grampositive bacteria.** *Mol Microbiol* 1998, **30:**737-749.

Hanique S, Colombo ML, Goormaghtigh E, Soumillion P, Frere JM, Joris B. **Evidence of an intramolecular interaction between the two domains of the BlaR1 penicillin receptor during the signal transduction.** *J Biol Chem*., 2004, **14**:14264-14272.

Hawley DK, McClure WR: **Compilation and analysis of *Escherichia coli* promoter DNA sequences.** *Nucleic Acids Res* 1983, **11:**2237-2255.

Hawley DK, McClure WR: **Mechanism of activation of transcription initiation from the lambda PRM promoter.** *J Mol Biol* 1982, **157**:493-525.

Henkin TM: **Control of transcription termination in prokaryotes.** *Annu Rev Genet* 1996, **30:**35-57.

Ho WL, Yu RC, Chou CC. **Effect of iron limitation on the growth and cytotoxin production of *Salmonella choleraesuis* SC-5.** *Int J Food Microbiol*., 2004, **3:**295-302.

Hori M, Yonei S, Sugiyama H, Kino K, Yamamoto K, Zhang QM: **Identification of high excision capacity for 5-hydroxymethyluracil mispaired with guanine in DNA of *Escherichia coli* MutM, Nei and Nth DNA glycosylases.** *Nucleic Acids Res* 2003, **31:**1191-1196.

Huerta AM, Collado-Vides J: **Sigma70 promoters in *Escherichia coli*: specific transcription in dense regions of overlapping promoter-like signals.** *J Mol Biol*. 2003, **2:**261-78.

Jacob F, Monad J: **Genetic Regulatory Mechanisms in the Synthesis of Proteins.** *J Mol Biol* 1961, **3**:318-356.

Jannick DB, Henrik N, Gunnar VH, Søren B: **Improved prediction of signal peptides: SignalP 3.0.** *J. Mol. Biol* 2004, **340:**783-795.

Kendall SL, Rison SC, Movahedzadeh F, Frita R, Stoker NG: **What do microarrays really tell us about *Mycobacterium tuberculosis*?** *Trends Microbiol* 2004, **12:**537-544.

Kleibl K: **Molecular mechanisms of adaptive response to alkylating agents in *Escherichia coli* and some remarks on O(6)-methylguanine DNA-methyltransferase in other organisms.** *Mutat Res* 2002, **512:**67-84.

Kunkle CA, Schmitt MP: **Analysis of the *Corynebacterium diphtheriae* DtxR Regulon: Identification of a putative siderophore synthesis and transport system that is similar to the Yersinia high-pathogenicity island-encoded yersiniabactin synthesis and uptake system.** *J Bacteriol*, 2003, **185:**6826-6840.

Lee JH, Wang T, Ault K, Liu J, Schmitt MP, Holmes RK: **Identification and characterization of three new promoter/operators from *Corynebacterium diphtheriae* that are regulated by the diphtheria toxin repressor (DtxR) and iron.** *Infect Immun* 1997, **65**:4273-4280.

Lesnik EA, Sampath R, Levene HB, Henderson TJ, McNeil JA, Ecker DJ: **Prediction of rho-independent transcriptional terminators in *Escherichia coli*.** *Nucleic Acids Res* 2001, **29:**3583-3594.

Marchler-Bauer A, Anderson JB, DeWeese-Scott C, Fedorova ND, Geer LY, He S, Hurwitz DI, Jackson JD, Jacobs AR, Lanczycki CJ, Liebert CA, Liu C, Madej T, Marchler GH, Mazumder R, Nikolskaya AN, Panchenko AR, Rao BS, Shoemaker BA, Simonyan V, Song JS, Thiessen PA, Vasudevan S, Wang Y, Yamashita RA, Yin JJ, Bryant SH: CDD: **a curated Entrez database of conserved domain alignments.** *Nucleic Acids Res,* 2003, **31**:383-387

Maria DE, Hanif GK, Owen We, Hamilton OS, Salzberg L: **Prediction of Transcription Terminators in Bacterial Genomes.** *J Mol Biol* 2000, **301:** 27-33.

Martinez A, Kolter R: **Protection of DNA during oxidative stress by the non specific DNA-binding protein Dps.** *J Bacteriol* 1997*,* **179:**5188-5194.

McCue LA, Thompson W, Carmack CS, Lawrence CE: **Factors influencing the identification of transcription factor binding sites by cross-species comparison.** *Genome Res* 2002, **12:**1523-1532.

Moreno-Hagelsieb G, Collado-Vides J: **A powerful non-homology method for the prediction of operons in prokaryotes.** *Bioinformatics* 2002, **18:**S329-S336.

Narendrakumar R, Yue W. **A High-Affinity Iron Permease Essential for *Candida albicans* Virulence.** *Science* 2000, **288**:1062-1064.

Olszewski MA, Noverr MC, Chen GH, Toews GB, Cox GM, Perfect JR, Huffnagle GB. **Urease expression by *Cryptococcus neoformans* promotes microvascular sequestration, thereby enhancing central nervous system invasion.** *Am J Pathol*., 2004, **5**:1761-1771..

Oram DM, Avdalovic A, Holmes RK: **Construction and characterization of transposon insertion mutations in *Corynebacterium diphtheriae* that affect expression of the diphtheria toxin repressor (DtxR).** *J Bacteriol*, 2002, **184:**5723-5732.

Outten FW, Wood MJ, Munoz FM, Storz G: **The SufE protein and the SufBCD complex enhance SufS cysteine desulfurase activity as part of a sulfur transfer pathway for Fe-S cluster assembly in *Escherichia coli.*** *J Biol Chem* 2003, **278:**45713-45719

Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N: **The use of gene clusters to infer functional coupling.** *Proc Natl Acad Sci* USA 1999, **96:**2896-2901.

Palyada K, Threadgill D, Stintzi A. **Iron acquisition and regulation in *Campylobacter jejuni.*** *J Bacteriol*., 2004, **14:**4714-4729.

Park HD, Guinn KM, Harrell MI, Liao R, Voskuil MI, Tompa M, Schoolnik GK, Sherman DR: **Rv3133c/dosR is a transcription factor that mediates the hypoxic response of *Mycobacterium tuberculosis.*** *Mol Microbiol* 2003, **48:**833-843.

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

Perez-Rueda E, Collado-Vides J: **The repertoire of DNA-binding transcriptional regulators in *Escherichia coli* K-12.** *Nucleic Acids Res* 2000, **28:**1838-47.

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

Qian Y, Lee JH, Holmes RK: **Identification of a DtxR-regulated operon that is essential for siderophore-dependent iron uptake in *Corynebacterium diphtheriae*.** *J Bacteriol* 2002, **184:**4846-4856.

Quadri LE, Sello J, Keating TA, Weinreb PH, Walsh CT. **Identification of a *Mycobacterium tuberculosis* gene cluster encoding the biosynthetic enzymes for assembly of the virulence-conferring siderophore mycobactin.** *Chem Biol.*, 1998, **5**: 631-645.

Register KB, Ducey TF, Brockmeier SL, Dyer DW: **Reduced virulence of a *Bordetella bronchiseptica* siderophore mutant in neonatal swine.** *Infect Immun* 2001*,* **69:**2137-2143.

Rice P, Longden I, Bleasby A: EMBOSS: **the European Molecular Biology Open Software Suite.** *Trends Genet* 2000, **16:**276-277.

Richardson JP: **Rho-dependent termination and ATPases in transcript termination.** *Biochim Biophys Acta* 2002, 2:251-260.

Robert S, Daniel SY,, Yuko Y, Richard DK, Andrew D. **A Permease-Oxidase Complex Involved in High-Affinity Iron Uptake in Yeast.** *Science*, 1996, **271**: 1552-1557.

Rodriguez GM, Gold B, Gomez M, Dussurget O, Smith I. **Identification and characterization of two divergently transcribed iron regulated genes in Mycobacterium tuberculosis.** *Tuber Lung Dis.* 1999, **5:**287-298. Erratum in: *Tuber Lung Dis* 1999, **6:**382.

Rodriguez GM, Voskuil MI, Gold B, Schoolnik GK, Smith I. **IdeR, An essential gene in *Mycobacterium tuberculosis*: role of IdeR in iron-dependent gene expression, iron metabolism, and oxidative stress response**. *Infect Immun.*, 2002, **7:**3371-3381.

Russo TA, Carlino UB, Johnson JR: **Identification of a new iron-regulated virulence gene, ireA, in an extraintestinal pathogenic isolate of *Escherichia coli*.** *Infect Immun* 2001, **69:**6209-6216.

Sabatti C, Rohlin L, Oh MK, Liao JC: **Co-expression pattern from DNA microarray experiments as a tool for operon prediction.** *Nucleic Acids Res* 2002, **30:**2886-2893.

Salgado H, Gama-Castro S, Martinez-Antonio A, Diaz-Peredo E, Sanchez-Solano F, Peralta-Gil M, Garcia-Alonso D, Jimenez-Jacinto V, Santos-Zavaleta A, Bonavides-Martinez C, Collado-Vides J: **RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12.** *Nucleic Acids Res*. 2004, **32:**D303-6.

Salgado H, Moreno-Hagelsieb G, Smith TF, Collado-Vides J: **Operons in *Escherichia coli*: genomic analyses and predictions.** *Proc Natl Acad Sci* USA 2000, **97:**6652-7.

Sandelin A, Wasserman WW, Lenhard B: **ConSite: web-based prediction of regulatory elements using cross-species comparison. Nucleic Acids Res.** 2004, **32:**W249-52.

Sarsero JP, Merino E, Yanofsky C. **A *Bacillus subtilis* operon containing genes of unknown function senses tRNATrp charging and regulates expression of the genes of tryptophan biosynthesis.** *Proc Natl Acad Sci* USA 2000, **6:**2656-61.

Schmitt MP, Drazek ES: **Construction and consequences of directed mutations affecting the hemin receptor in pathogenic *Corynebacterium* species.** *J Bacteriol* 2001, **183:**1476-1481.

Schmitt MP, Holmes RK: **Cloning, sequence, and footprint analysis of two promoter/operators from *Corynebacterium diphtheriae* that are regulated by the diphtheria toxin repressor (DtxR) and iron.** *J. Bacteriol.*, 1994, **4**:1141-1149.

Schmitt MP, Predich M, Doukhan L, Smith I, Holmes RK. **Characterization of an iron-dependent regulatory protein (IdeR) of *Mycobacterium tuberculosis* as a functional homolog of the diphtheria toxin repressor (DtxR) from *Corynebacterium diphtheriae*.** *Infect Immun.*, 1995, **11:**4284-4289.

Schmitt MP: **Transcription of the *Corynebacterium diphtheriae* hmuO gene is regulated by iron and heme**. *Infect. Immun.,* 1997, **11**:4634-4641.

Schneider TD, Stephens RM: **Sequence logos: a new way to display consensus sequences.** *Nucleic Acids Res* 1990, **18:**6097-6100.

Schneider TD: **Information content of individual genetic sequences.** *J Theor Biol* 1997, **189:**427-441.

Shannon CE: **A mathematical theory of communication.** *Bell System Technical Journal* 1948, 379-423 and 623-656.

Smith JL: **The physiological role of ferritin-like compounds in bacteria.** *Crit Rev Microbiol* 2004, **30:**173-185.

Suh JW, Boylan SA, Oh SH, Price CW: **Genetic and transcriptional organization of the *Bacillus subtilis* spc-alpha region.** *Gene* 1996, **169:**17-23.

Tao X, Murphy JR: **Binding of the metalloregulatory protein DtxR to the diphtheria tox operator requires a divalent heavy metal ion and protects the palindromic sequence from DNase I digestion.** *J. Biol. Chem.,* 1992, **30**:21761-21764.

Tao X, Schiering N, Zeng HY, Ringe D, Murphy JR: **Iron, DtxR, and the regulation of diphtheria toxin expression.** *Mol Microbiol* 1994, **14:**191-197.

Tao X, Schiering N, Zeng HY, Ringe D, Murphy JR: **Iron, DtxR, and the regulation of diphtheria toxin expression.** *Mol Microbiol.,* 1994, **14:**191-197.

ter Huurne AA, Muir S, van Houten M, Koopman MB, Kusters JG, van der Zeijst BA, Gaastra W:**The role of hemolysin(s) in the pathogenesis of *Serpulina hyodysenteriae*.** *Zentralbl Bakteriol* 1993, **278:**316-325.

Thieffry D, Huerta AM, Perez-Rueda E, Collado-Vides J: **From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in *Escherichia coli*.** *Bioessays* 1998, 20:433-440.

Timothy L. Bailey, Charles Elkan: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers.** *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, 1994, 28-36.

Ton-That H, Schneewind O: **Assembly of pili on the surface of *Corynebacterium diphtheriae*.** *Mol Microbiol* 2003, **50:**1429-1438.

Unniraman S, Prakash R, Nagaraja V. **Conserved economics of transcription termination in eubacteria.** *Nucleic Acids Res* 2002, **30:** 675-684.

Unniraman S, Prakash R, Nagaraja V: **Alternate paradigm for intrinsic transcription termination in eubacteria.** *J Biol Chem* 2001, **45:**41850-41855.

Urbanski NK, Beresewicz A. **Generation of *OH initiated by interaction of Fe2+ and Cu+ with dioxygen; comparison with the Fenton chemistry.** *Acta Biochim Pol.*, 2000, **4:**951-962.

Wang L, Trawick JD, Yamamoto R, Zamudio C: **Genome-wide operon prediction in *Staphylococcus aureus*.** *Nucleic Acids Res* 2004, **32:**3689-3702.

Wilke MS, Hills TL, Zhang HZ, Chambers HF, Strynadka NC. **Crystal Structures of the Apo and Penicillin-acylated Forms of the BlaR1 {beta}-Lactam Sensor of *Staphylococcus aureus*.** *J Biol Chem.*, 2004, **45**:47278-47287.

Wilson KS, von Hippel PH: **Transcription termination at intrinsic terminators: The role of the RNA hairpin.** *Proc Natl Acad Sc* USA 1995, **92:**8793-8797.

Yanofsky C, Konan KV, Sarsero JP: **Some novel transcription attenuation mechanisms used by bacteria.** *Biochimie* 1996, **78:**1017-1024.

Yanofsky C: **Attenuation in the control of expression of bacterial genomes.** *Nature*1981, **289:**751-758.

Yanofsky C: **Transcription attenuation: once viewed as a novel regulatory strategy.** *J Bacteriol* 2000, **182:**1-8.

Yarnell WS, Roberts JW: **Mechanism of intrinsic transcription termination and antitermination.** *Science* 1999, **284:** 611-615.

Zaika EI, Perlow RA, Matz E, Broyde S, Gilboa R, Grollman AP, Zharkov, DO: **Substrate discrimination by formamidopyrimidine-DNA glycosylase: a mutational analysis.** *J Biol Chem* 2004*,* **279:**4849-4861.

Zhao G, Ceci P, Ilari A, Giangiacomo L, Laue TM, Chiancone E, Chasteen ND: **Iron and hydrogen peroxide detoxification properties of DNA-binding protein from starved cells. A ferritin-like DNA-binding protein of *Escherichia coli*.** *J Biol Chem* 2002, **277:**27689-27696.

Zheng Y, Szustakowski JD, Fortnow L, Roberts RJ, Kasif S: **Computational identification of operons in microbial genomes**. *Genome Res* 2002, **12:**1221-1230.

Appendix I

Publications

Bioinformatics Advance Access published March 3, 2005

Computational prediction and experimental verification of novel IdeR binding sites in the upstream sequences of *Mycobacterium tuberculosis* ORFs

Prachee Prakash[1], Sailu Yellaboina[2], Akash Ranjan[2] and Seyed E. Hasnain[1,3,*]
[1]Laboratory of Molecular and Cellular Biology, Centre for DNA Fingerprinting and Diagnostics, Hyderabad, 500 076, India
[2]Laboratory of Computational and Functional Genomics, Centre for DNA Fingerprinting and Diagnostics, Hyderabad 500 076, India
[3]Jawaharlal Nehru Centre for Advanced Scientific Research, Jakkur, Bangalore, 560064, India

Running Head: IdeR regulated genes of *M. tuberculosis*

*To whom Correspondence should be addressed
Contact: Seyed E Hasnain, Laboratory of Molecular and Cellular Biology, CDFD, Nacharam, Hyderabad, 500 076, India; Tel: 7155604-05; Fax: 91-040-7155479, 7155610; E-Mail: hasnain@cdfd.org.in Tel: 7155604-05, Fax: 91-040-7155479, 7155610

ABSTRACT

IdeR (Iron dependent regulator) is a key regulator of virulence factors and iron acquisition systems in *M. tuberculosis*. Despite the wealth of information available on IdeR regulated genes of *Mtb*, there is still an underlying possibility that there are novel genes/pathways that have gone undetected, the identification of which could give new insights into understanding the pathogenesis of *Mtb*. We describe an *in silico* approach employing positional relative entropy method to identify potential IdeR binding sites in the upstream sequences of all the 3919 ORFs of *Mtb*. While many of the predictions made by this approach overlapped with the ones already identified by microarray experiments and binding assays, pointing to the accuracy of our method, few genes for which there has been no evidence for IdeR regulation were additionally identified. Our results have implications on iron dependent regulatory mechanism of *Mtb* vis-a-vis the activity of urease operon and novel transcription regulators and transporters.

INTRODUCTION

In pathogenic bacteria, many virulence factors and iron acquisition systems are regulated by iron dependent transcription regulators [Litwin and Calderwood,1993]. One of the key regulators of such systems in *Mycobacterium tuberculosis* is IdeR [Iron dependent regulator], first identified as a homologue of the DtxR [Diphtheria toxin Repressor] protein *of* *Corynebacterium diphtheriae* [Schmitt *et al.*,1995]. IdeR has been known to govern the expression of a wide variety of genes ranging from those involved in iron acquisition and oxidative stress response to ones that code for enzymes involved in aromatic amino acid biosynthesis [Gold *et al.*, 2001; Rodriguez and Smith, 2003]. The success of *M. tuberculosis* in the establishment of an infection is also dependent upon its ability to acquire iron from its neighbouring environment. While low iron is a limiting factor for the pathogen growth and survival, even high iron is detrimental as it leads to the formation of highly reactive hydroxyl radicals *via* the Fenton reaction. Hence, acquisition of iron by pathogenic bacteria has to be tightly regulated.

IdeR was first identified as the mycobacterial equivalent of DtxR of *Corynebacterium diphtheriae* [Schmitt *et al.*, 1995]. DtxR serves as a repressor of *tox* gene, the structural gene for diphtheria toxin. Apart from this function, DtxR also behaves as a regulatory protein involved in the iron metabolism of the bacterium [Boyd *et al.*, 1990; Schmitt and Holmes, 1991]. The function of DtxR was found to be similar to the very well known Fur protein of gram-negative bacteria. Under iron sufficient conditions, DtxR causes repression of genes involved in iron metabolism by binding to their operator sequences, with a high specificity, thereby blocking transcription [Tao *et al.*, 1992]. IdeR of *M. tuberculosis* shares 59% overall amino acid identity within a 230 aa stretch to DtxR. Initial experiments were carried out to determine if IdeR represses transcription of DtxR regulated promoters. DNA mobility shift assays and DNA footprinting analysis showed that IdeR binds to the same promoter sequences to which the Corynebacterial DtxR protein binds. Binding was observed to be metal dependent [Schmitt *et al.*, 1995].

In *M. tuberculosis*, under iron sufficient conditions, IdeR binds to the upstream sequences of genes required for growth in low iron conditions thereby preventing their transcription. Under iron limiting conditions, IdeR no longer binds to these promoter regions which are free to allow the binding of RNA polymerase and subsequent transcription of the downstream gene. *ide*R is also an essential gene of *M. tuberculosis* and the encoded protein regulates the expression of genes involved in the metabolism of iron and oxidative stress response [Rodriguez *et al.*, 2002]. Though *ide*R null mutant of *M. tuberculosis* can not be generated without the incorporation of a second copy of the gene, Rodriguez and coworkers obtained a rare mutant of *ide*R in which the lethal effects of *ide*R inactivation were alleviated by a suppressor mutation. This mutant showed a restricted iron assimilation capacity. The authors also studied the transcription profiles of wild-type, *ide*R mutant, and *ide*R-complemented mutant *Mtb* strains using DNA microarray. This resulted in the identification of genes regulated by iron and IdeR. These genes encode proteins involved in siderophore biosynthesis and iron storage, enzymes of aromatic amino acid biosynthesis, putative transporters, members of the PE/PPE family, transcriptional regulators, and enzymes involved in lipid metabolism.

IdeR of *M. tuberculosis* in association with ferrous ions binds to a 19 bp inverted repeat consensus sequence or iron box [TTAGGTTAGGCTAACCTAA] in the upstream sequences of several genes [Schmitt *et al.*,1995]. Gel mobility shift assays, DNA footprinting, Reporter gene assays and DNA microarray are four techniques that have been exploited by a multitude of workers to determine the genes expressed/repressed in *M. tuberculosis* as a function of iron availability [Gold *et al.*, 2001; Camacho *et al.*, 1999]. Genes involved in iron acquisition and storage have been shown to be IdeR regulated [Dussurget *et al.*, 1996; Gold *et al.*, 2001]. Several other genes not directly involved in siderophore biosynthesis have also been shown to be expressed or repressed as a function of iron stress [Rodriguez and Smith, 2003; Dussurget *et al.*, 1999]. These reports suggest that IdeR is a global regulator that controls several genes involved in iron metabolism and processes related to iron metabolism. Experimental evidence for iron mediated regulation of quite a few mycobacterial genes exists. Two divergently transcribed genes, *his*E [a part of the histidine operon] and a PPE gene [Rv2123] have been shown to be IdeR regulated [Rodriguez *et al.*, 1999]. Gel shifts and footprinting assays have revealed that IdeR regulates the expression of these genes by binding to the iron boxes in the regulatory region and binding was dependent upon the concentration of iron in the reaction mix. Few other genes involved in the biosynthesis of siderophores [*mbt*A, *mbt*B, *mbt*I], biosynthesis of aromatic amino acids [*trp*E2, *phe*A, *his*E, *his*B] and others like iron storage proteins [bacterioferritins, *bfr*A, *bfr*B] have also been experimentally shown to be part of the IdeR regulon [Gold *et al.*, 2001]. Functional characterization of genes not apparently involved in iron metabolism would lead to further insights into the relation between iron metabolism and various aspects of mycobacterial physiology. This report describes the use of a computational approach to identify novel genes under the regulatory control of IdeR followed by its experimental verification. Our results while confirming already known IdeR regulated genes, have additionally identified new genes.

METHODS

Computational prediction of IdeR binding sites

The complete genome sequence of *Mycobacterium tuberculosis* H37Rv was downloaded from NCBI ftp site [*ftp://ftp.ncbi.nlm.gov*] and IdeR binding sites were collected from literature [Table 1]. Profiles for recognition of IdeR binding sites were calculated by positional relative entropy method assuming that each position is independent sites [Yellaboina *et al.*, 2004a, 2004b]. A matrix was developed for the purpose, which was used to scan the upstream sequences of all the genes from -400 to +20 of the translation start site to identify potential IdeR binding. Consensus IdeR binding sites were used to compute probability distributions of four different nucleotides within the binding sites of known sequences as well as throughout the genome. The probability distributions of nucleotides within and outside regulatory region were used to compute the relative entropy of segments [length 19bp] along the +20 to -400 regions of all the genes *M. tuberculosis*. Finally segments were sorted according to the relative entropy and segments with high relative entropy were considered as probable Iron dependent repressor binding sites.

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

pRSETa expression vector [Promega] with an N-terminal 6X His tag was used to clone the ORF Rv2711 of *M. tuberculosis* that encodes IdeR. Briefly, Rv2711 was amplified from *M. tuberculosis* H37Rv DNA using primers with specific restriction enzyme sites. [Forward primer: ATT***GGATCC***ATGAACGAGTTGGTTGATA and Reverse primer: TGT***AAGCTT***GACTTTCTCGACCTTGACC] and the amplicon was cloned into the corresponding sites of pRSETa expression vector. *E coli* BL21DE3 cells transformed with the 6xHis tagged chimeric construct were grown in 1L of LB medium supplemented with 100μg/ml of ampicillin and 10%glycerol. IPTG [0.1mM] was added to a mid log phase culture. The cells were kept in an incubator shaker for another five hours at 27^0C and 150rpm to allow protein expression. After induction, cells were harvested by centrifugation and resuspended in 20ml of lysis buffer [10mM Tris HCl, 100mM NaCl and 10% glycerol, pH7.5] with 0.1mM PMSF and disrupted using a sonicator. After a second round of centrifugation for 10 minutes at 10,000xg, the supernatant was applied to a Ni-NTA affinity column [Qiagen, USA].

Affinity chromatography: The supernatant was allowed to bind to Ni-NTA column [Qiagen] packed in a polypropylene column. The recombinant protein was purified after washing the column with 5 bed volumes of lysis buffer containing 30mM imidazole and eluting with 250mM imidazole. The eluates were analyzed by SDS PAGE and dialyzed against Tris buffer to remove salts and imidazole. The purity of the eluted protein was checked on SDS PAGE followed by Coomassie Blue staining

Gel retardation assay: Binding of IdeR to the predicted iron box was carried out in a 20μl reaction consisting of 1X buffer [10mM Tris HCl, 50mMNaCl, 10% glycerol, 5μg/ml acetylated BSA, 1mM DTT, 1mM PMSF and 50mM $MgCl_2$], 1% NP40, 1μg/ml poly dIdC, purified IdeR [1μM] and 3-5fmol of ^{32}P labeled probe. The probe consisted of the annealed 19bp oligo corresponding to the predicted IdeR box end labeled with ^{32}P using T4 polynucleotide kinase enzyme. Reaction was performed in the presence and absence of $CoCl_2$ (200μM). Unlabeled oligo was used for specific competition. After the addition of labeled probe, the reaction mixture was incubated for 15 min at 25^0C followed by loading on a 4% polyacrylamide gel in 1XTBE buffer. Electrophoresis was carried out at 200V for 30 minutes at 4^0C. After electrophoresis, the gel was dried and analyzed by autoradiography.

South western assay: The bacterial extract overexpressing *M tuberculosis* IdeR was separated on an SDS/PAGE gel and the proteins were electrophoretically transferred to a nitrocellulose membrane in a buffer containing 25mM Tris, 190mM glycine and 20% methanol for 16hrs at 30mA. The protein on the membrane was renatured by incubating in blocking buffer [2% non fat dry milk, 1% BSA, 10mM Hepes NaOH, 0.1mM EDTA, 200mM NaCl, 50mM $MgCl_2$ and 16μg/ml sonicated sperm DNA]. After renaturation, the membrane was placed in a hybridization bag with binding buffer [blocking buffer with 0.2% non-fat dry milk and 10^6 cpm/ml labeled oligo. Hybridization was performed with constant shaking for 16 hours. The membrane was briefly rinsed in blocking buffer without skimmed milk or BSA, dried, covered with plastic wrap and subjected to autoradiography.

RESULTS

Novel IdeR binding sites are present upstream of *fec*B, a periplasmic lipoprotein coding gene and Rv1404, a putative transcriptional regulator

The consensus IdeR binding site collected from published literature [Table 1] was used to identify similar IdeR binding regions in the –400 to +20 regions of all the 3919 ORFs of *Mycobacterium tuberculosis*. A complete list of IdeR binding sites with the highest scores as calculated by the positional relative entropy method is shown in Table 2.

To date, the most detailed study on prediction of IdeR binding sites along with experimental verification has been carried out by Gold *et al.* [2001]. Additionally, microarray analysis of genes induced by low iron and in an IdeR mutant strain have also shed light on the iron-dependent regulation of mycobacterial genes [Rodriguez *et al.*, 2002]. While our method indeed identified novel IdeR binding sites, the possibility of occurrence of additional such sites cannot be ruled out. As a first step towards the analysis of our predictions, the results were compared with the available information on IdeR regulated genes. Though most of these genes were earlier known to be IdeR regulated, the present study identified for the first time that a part of the ferric dicitrate type transporter complex, FecB, a periplasmic lipoprotein and Rv1404, a putative transcriptional regulator are possibly regulated by IdeR [Table 2]. The upstream sequence of *fec*B shows the presence of an IdeR box at –302 position. On account of absence of reports on the details of the iron transport system of *M. tuberculosis*, the ferric dicitrate transporter system does seem to be an important candidate. A new transcription regulator [Rv1404] and a hypothetical protein Rv2663 that were not earlier predicted to be part of the IdeR regulon could also be identified in this study.

IdeR binds to the IdeR box present in the intergenic region between the ORFs Rv1846c and Rv1847c

While many of the IdeR binding sites predicted by this study overlapped with ones predicted by earlier workers, experimental evidence demonstrated by *in vitro* binding experiments and reporter gene assays is available for only a few. These include *his*E, Rv2123 [Rodriguez *et al.*, 1999], Rv3402, *mbt*I *his*G, *mbt*A, *mbt*B, *mbt*I, Rv3402 and *bfr*A [Gold *et al.*, 2001] etc. As per our prediction, the IdeR box upstream of the ORF Rv1846c shows one of the highest similarity score to the IdeR consensus sequence. However, experimental evidence for the same does not exist. Moreover, Rv1846c does not figure in the list of genes induced in an IdeR mutant strain [Rodriguez and Smith, 2003]. The binding site between Rv1846-Rv1847 was also observed to be conserved in other mycobacteria. Hence, it was decided to determine if IdeR binds to this predicted iron box. The ORF, Rv2711 that encodes IdeR was cloned in the *Bam*HI and *Hind*III sites of pRSETa vector with an N-terminal Histidine tag and expressed and purified as a recombinant protein in *E. coli* BL21 cells [Figure 1]. Purified recombinant IdeR was used in gel retardation and south-western assays to test if it binds to the predicted IdeR box in the intergenic region between Rv1846c and Rv1847 [Figure 2A]. As evident from the gel shift assay [Figure 2B], IdeR does bind to the 19bp IdeR binding site present in the intergenic region between Rv1846c and Rv1847. The binding could be competed out using cold oligos indicating the specificity of binding.

To convincingly demonstrate binding of IdeR to the above mentioned probable iron box, south-western analysis was carried out. *E. coli* BL21 strain transformed with recombinant plasmid carrying *M. tuberculosis ide*R was grown to mid log phase and fractionated by electrophoresis on a polyacrylamide gel. The gel was probed with radiolabeled oligonucleotide corresponding to the predicted iron box. While the vector control lysate lane [Figure 2C, Lane 1] did not show any binding, the induced cultures showed a positive binding. These data conclusively demonstrate that IdeR indeed binds to the predicted iron box element present in the divergently transcribed ORFs, Rv1846c and Rv1847 of *M. tuberculosis*.

DISCUSSION

*N*on-availability of soluble form of iron is an important form of nutritional stress presented by the host to the bacterium, it is therefore logical to assume that genes responsible for the acquisition of iron are essential for full virulence and establishment of a successful infection. *M. tuberculosis has* an elaborate network of genes for the biosynthesis of siderophores, the iron acquisition systems [Qadri *et al.*, 1998]. Recent experiments have shown that these genes are regulated by iron-dependent regulatory proteins [Gold *et al.*, 2001]. Transcriptional control plays a key role in regulating gene expression in response to various environmental conditions. Apart from the production of siderophores as a function of low iron availability, *M. tuberculosis* also produces many other iron regulated proteins, which are the probable virulence factors of the bacterium [Rodriguez and Smith, 2003].

The ferric dicitrate type transporter complex of *M. tuberculosis* as a probable IdeR regulated system

While a number of transporter proteins like Rv1463 (an ABC transporter), Rv2459 (a probable drug efflux pump), Rv1348 (membrane protein similar to Yersiniabactin uptake system) etc have been earlier predicted to be IdeR regulated, the present work suggests for the first time IdeR dependent regulation of *fec*B of *M. tuberculosis*. FecB has been annotated as a probable Fe[III] dicitrate binding periplasmic lipoprotein. The *fec* operon is very well characterized in *E. coli* and a dyad repeat sequence GAAAATAATTCTTATTTCG present upstream to *fec*A has been proposed to serve as the binding site of the Fur iron repressor protein in *E. coli* [Zimmermann *et al.*, 1984, Pressler *et al.*, 1988]. It is thus likely that FecB of *M. tuberculosis* could also be part of the iron transport complex of the bacterium and the regulation of the gene is brought about by IdeR, the Fur equivalent of *M. tuberculosis*. Additionally, as predicted by the method described above as well as the NCBI pattern search by Gold *et al.* [2001], another membrane protein coded by Rv1348c [similar to the yersiniabactin uptake system] shows an IdeR box in its upstream sequence. This protein also appears to be an important candidate in the uptake of siderophore like compounds.

Regulation of a probable MarR equivalent transcriptional regulator, Rv1404 by IdeR

Quite a few transcription regulators are known to be under the regulatory control of IdeR [Rodriguez *et al.*, 2002]. Results presented above could also identify Rv1404, a novel transcriptional regulator that shares some similarity to the Multiple antibiotic resistance regulator [MarR] protein from *E. coli*, as a probable IdeR regulated gene. If the antibiotic resistance regulator function of Rv1404 is proven, this could provide a clue to iron dependent regulation of antibiotic resistance in *M. tuberculosis*. Here, it would be worth mentioning that the ORF Rv1846c that is also predicted to have an IdeR box in its upstream sequence also shows some similarity to the penicillase repressor protein of *Bacillus licheniformis*. These findings suggest that IdeR could be a global regulator that activates even other regulatory proteins that take care of the iron dependent regulation of a broader network of *M. tuberculosis* genes.

Regulation of the urease operon by IdeR

The ORFs Rv1846 and Rv1847 have interesting predicted functions that are important in the context of the pathology of *M. tuberculosis*. While Rv1847 is a hypothetical protein probably a thioesterase involved in the biosynthesis of aromatic compounds, Rv1846c codes for a transcription regulator with some similarity to the penicillase repressor protein of *Bacillus licheniformis*. Interestingly, Rv1847 also appears to be part of the same operon that codes for genes involved in the biosynthesis of the urease enzyme. It is known that *Mycobacterium tuberculosis* survives in the acidic, toxic and hostile environment of the macrophage phagolysosome. One mechanism of survival is to somehow increase the pH of the phagolysosome. In this respect, the activity of the urease operon assumes importance as it could possibly help in neutralization of the acidic pH [Clemens *et al.*, 1995]. However, the mechanism of regulation of *M. tuberculosis* urease operon has not yet been described anywhere. As an iron box exists upstream to the urease operon (directly upstream of ORF, Rv1847), it was tempting to speculate that urease could also be regulated by IdeR. Additional evidence springs from the fact that in many pathogenic bacteria like *H. pylori*, the urease operon is regulated by ferric uptake regulatory [Fur] proteins [Bijlsma *et al.*, 2002]

Along with the prediction of a high score, experimental evidence for binding of IdeR of *M. tuberculosis* to iron box element upstream of the urease operon has been provided in this work. Urease has been implicated in the virulence of several other pathogenic micro orgamisms. In *H.pylori*, *Salmonlla typhimurium* and *Escherichia coli*, urease is regulated by Ferric uptake regulator in response to pH [Bijlsma *et al.*, 2002; Heimer *et al.*, 2002]. In case of *M. tuberculosis*, ammonia generated by the action of urease may be of considerable importance in alkalinizing the microenvironment of the organism and preventing phagosome-lysosome fusion and phagosome acidification. In addition ammonia generated by the action of urease should be available to the organism for assimilation of nitrogen into biomolecules.

In summary, this study enhances the current understanding of the complex network of *M. tuberculosis* genes expressed/repressed as a consequence of iron stress. The study also adds considerably to the understanding of the various mechanism of survival adopted by the bacterium to survive inside in the iron deficient environment presented by the host.

Acknowledgements
This research in Hasnain laboratory was supported by grants from the Department of Biotechnology, Council of Scientific & Industrial Research (CSIR) Govt. of India. PP and SY are recipients of Senior Research Fellowship from the CSIR.

REFERENCES

Bijlsma,J.J., Waidner,B., Vliet,A.H., Hughes,N.J., Hag,S., Bereswill,S., Kelly,D.J., Vandenbroucke-Grauls,C.M., Kist,M. and Kusters,J.G. (2002) The *Helicobacter pylori* homologue of the ferric uptake regulator is involved in acid resistance. *Infect Immun*, **70,**606-11.

Boyd,J., Oza,M.N. and Murphy,J.R. (1990) Molecular cloning and DNA sequence analysis of a diphtheria tox iron-dependent regulatory element (dtxR) from *Corynebacterium diphtheriae. Proc Natl Acad Sci* USA, **87**,5968-72.

Camacho,L.R., Ensergueix, D., Perez, E., Gicquel, B., and Guilhot, C. (1999) Identification of a virulence gene cluster of *Mycobacterium tuberculosis* by signature-tagged transposon mutagenesis. *Mol Microbiol*, **34**,257–267.

Clemens,D.L., Lee,B.Y. and Horwitz,M.A. (1995) Purification, characterization, and genetic analysis of *Mycobacterium tuberculosis* urease, a potentially critical determinant of host-pathogen interaction. *J Bacteriol.*, **177**,5644-52.

Dussurget,O., Rodriguez,M. and Smith,I. (1996) An ideR mutant of Mycobacterium smegmatis has derepressed siderophore production and an altered oxidative-stress response. *Mol Microbiol.*, **22**, 535-44.

Dussurget,O., Timm,J., Gomez,M., Gold,B., Yu,S., Sabol,S.Z., Holmes,R.K., Jacobs,W.R. Jr, Smith,I. (1999) Transcriptional control of the iron-responsive *fxb*A gene by the mycobacterial regulator IdeR. *J Bacteriol.*,**181**:3402-8.

Gold,B., Rodriguez,G.M., Marras,S.A., Pentecost,M. and Smith,I. (2001) The *Mycobacterium tuberculosis* IdeR is a dual functional regulator that controls transcription of genes involved in iron acquisition, iron storage and survival in macrophages. *Mol Microbiol.*,**42**,851-65.

Heimer,S.R., Welch,R.A., Perna,N.T., Posfai,G., Evans,P.S., Kaper,J.B., Blattner,F.R. and Mobley,H.L. (2002).Urease of enterohemorrhagic *Escherichia col*i: evidence for regulation by fur and a trans-acting factor. *Infect Immun.*,**70**,1027-31.

Litwin,C.M. and Calderwood,S.B. (1993). Role of iron in regulation of virulence genes. *Clin Microbiol Rev.* ,**6**,137-49.

Pressler,U., Staudenmaier,H., Zimmermann,L. and Braun,V. (1988) Genetics of the iron dicitrate transport system of *Escherichia coli.,J Bacteriol.*,**170**,2716-24.

Quadri,L.E., Sello,J., Keating,T.A., Weinreb,P.H. and Walsh,C.T. (1998) Identification of a *Mycobacterium tuberculosis* gene cluster encoding the biosynthetic enzymes for assembly of the virulence-conferring siderophore mycobactin. *Chem Biol.*,**5**,631-45.

Rodriguez,G.M., Gold,B., Gomez,M., Dussurget,O. and Smith,I. (1999) Identification and characterization of two divergently transcribed iron regulated genes in *Mycobacterium tuberculosis*. *Tuber Lung Dis*.,**79**, 287-98.

Rodriguez,G.M., Voskuil,M.I., Gold,B., Schoolnik,G.K. and Smith,I. (2002) *ide*R, An essential gene in *Mycobacterium tuberculosis*: role of IdeR in iron-dependent gene expression, iron metabolism, and oxidative stress response. *Infect Immun*.,**70**,3371-81.

Rodriguez,G.M. and Smith,I. (2003) Mechanisms of iron regulation in mycobacteria: role in physiology and virulence. *Mol Microbiol.* ,**47**,1485-94.

Schmitt,M.P. and Holmes,R.K. (1991) Iron-dependent regulation of diphtheria toxin and siderophore expression by the cloned *Corynebacterium diphtheriae* repressor gene *dtx*R in *C. diphtheriae* C7 strains. *Infect Immun*.,**59**,1899-904.

Schmitt,M.P., Predich,M., Doukhan,L., Smith,I. And Holmes,R.K . (1995) Characterization of an iron-dependent regulatory protein (IdeR) of *Mycobacterium tuberculosis* as a functional homolog of the diphtheria toxin repressor (DtxR) from *Corynebacterium diphtheriae*. *Infect Immun*., **63**,4284-9.

Tao,X., Boyd,J. and Murphy,J.R. (1992) Specific binding of the diphtheria tox regulatory element DtxR to the tox operator requires divalent heavy metal ions and a 9-base-pair interrupted palindromic sequence*. Proc Natl Acad Sci* USA.,**89**,5897-901.

Yellaboina,S., Ranjan,S., Chakhaiyar,P., Hasnain,S.E., and Ranjan,A. (2004). Prediction of DtxR regulon: Identification of binding sites and operons controlled by Diphtheria toxin repressor in *Corynebacterium diphtheriae*. *BMC Microbiology*. **00**:000-000 [In press]

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

Zimmermann,L., Hantke,K. and Braun,V. (1984). Exogenous induction of the iron dicitrate transport system of *Escherichia coli* K-12. *J Bacteriol*.,159,271-7.

FIGURE LEGENDS

Figure 1: Purification of the Iron-dependent Regulator (IdeR) of *Mycobacterium tuberculosis* as a recombinant protein in *E. coli*. *M. tuberculosis* IdeR (coded by ORF Rv2711) was cloned in the *Bam*HI/*Hin*dIII sites of pRSETa expression vector and purified as a 6X Histidine tagged recombinant protein using affinity chromatography procedures. Purified protein was fractionated on a 10% polyacrylamide gel and stained with Coomassie Brilliant Blue dye. M represents the protein molecular size marker (Broad range, Genei, India), E1-E7 show the successive eluted fractions of the recombinant protein. Arrowhead indicates the position of the pure eluted protein.

Figure 2: Recombinant IdeR binds to predicted iron box element. A: Schematic representation of the divergently transcribed ORFs, Rv1846c and Rv1847 with an IdeR binding site in the intergenic region. *ure* A, B and C are genes of the urease operon. B: Autoradiogram of the gel retardation assay demonstrating the binding of IdeR to the predicted iron box shown in A. Binding was specific as indicated by the disappearance of the band upon addition of cold oligo (Lanes 5 and 6). Absence of a band in Lane 7 confirms a metal dependent binding of IdeR to the predicted iron box. C: Autoradiogram of the south western assay demonstrating the binding of *Mtb* IdeR in *E. coli* BL21 cell lysates (indi\uced for 2hrs and 5hrs) to the predicted iron box shown in A. The cultures were induced for 2hrs and 5hrs, fractionated on a 10% polyacrylamide gel, transferred to a nitrocellulose membrane, renatured and hybridized with ^{32}P labeled 19bp oligo deoxyribonucleotide. Arrowhead indicates the position of the

band. Specificity of binding was confirmed by the absence of the corresponding band in vector control lane (Lane1).

Table 1: Known IdeR binding sites in the upstream sequences of *M. tuberculosis* ORFs

IdeR binding sequence	**Downstream ORF**
CAAGGTAAGGCTAGCCTTA	Rv1519
TTATGTTAGCCTTCCCTTA	Rv3403c
TTAACTTAGGCTTACCTAA	Rv3839
TTAGGCAAGGCTAGCCTTG	Rv1343c
CAAGGCTAGCCTTGCCTAA	Rv1344
TATGGCATGCCTAACCTAA	Rv1347c
TTCGGTAAGGCAACCCTTA	Rv1348
ATAGGTTAGGCTACCCTAG	Rv2122c
CTAGGGTACCCTAACCTAT	Rv2123
AGAGGTAAGGCTAACCTCA	Rv3402c
TTAGTGGAGTCTAACCTAA	Rv1876
GTAGGTTAGGCTACATTTA	Rv2386c
CTAGGAAAGCCTTTCCTGA	Rv3841
TTAGCTTATGCAATGCTAA	Rv0282
TTAGGCTAGGCTTAGTTGC	RV0451c
TTAGCACAGGCTGCCCTAA	Rv2383c
TTAGGGCAGCCTGTGCTAA	Rv2384

Table 2: Candidate IdeR binding sites in the genome of *Mycobacterium tuberculosis*

IdeR binding site	Position Of binding site relative to Translation start site	Score	Gene Anno tation	Rv number	Predicted Function
TTAGTGGAGTCTAACCTAA	-226	5.2563	*bfr*A	Rv1876	Bacterioferritin
ATAGGCAAGGCTGCCCTAA	**-151**	**5.19346**	**_**	**Rv1846c**	**Predicted transcriptional regulator**
TTAGCACAGGCTGCCCTAA	-86	5.16997	*mbt*A	Rv2384	Peptide arylation enzymes
TTAGGGCAGCCTGTGCTAA*	-32	5.15772	*mbt*B	Rv2383c	Peptide arylation enzymes
TTATGTTAGCCTTCCCTTA*	-2	5.14546	_	Rv3403c	hypothetical protein
CTAGGAAAGCCTTTCCTGA*	-73	5.12055	*bfr*B	Rv3841	Ferritin-like protein
TTAGGCAAGGCTAGCCTTG*	-85	5.09743	_	Rv1343c	hypothetical protein
TTAACTTAGGCTTACCTAA	-36	5.09181	_	Rv3839	hypothetical protein
ATAGGTTAGGCTACCCTAG*	-51	5.07767	PPE	Rv2123	PPE
TTAGGTAAGCCTAAGTTAA	-79	5.04482	*phe*A	Rv3838c	Prephenate dehydratase
CTAGGGTAGCCTAACCTAT*	-95	5.04385	*his*I	Rv2122c	Phosphoribosyl-ATP pyrophosphohydrolase
TTAGGGCAGCCTTGCCTAT	**-146**	**5.0165**	**-**	**Rv1847**	**Hypothetical protein**
CAAGGCTAGCCTTGCCTAA	-292	4.97724	*fad*D33	Rv1345	Acyl-CoA synthetases (AMP-forming)/AMP- acid ligases II
CAAGGTAAGGCTAGCCTTA*	-50	4.97077	_	Rv1519	hypothetical protein
CAAGGTAAGGCTAGCCTTA	-345	4.97077	_	Rv1520*	Glycosyltransferases involved in cell wall biogenesis
GTAGGTTAGGCTACATTTA*	-25	4.8669	*trp*E2	Rv2386c	Anthranilate/para-aminobenzoate synthases component I
GCAGGTCAGGCTACCCTTA	-26	4.82224	*mur*B	Rv0482	UDP-N-acetylmuramate dehydrogenase
ATAGGAAAGCCGATCCTTA	-36	4.64865	_	Rv0114	Histidinol phosphatase and related phosphatases
GTAGACCAGGCTCCCCTTG	**-302**	**4.62592**	***fec*B**	**Rv3044**	**ABC-type Fe3+-siderophores ransport systems**
TAAGGGTAGCCTGACCTGC	-20	4.61752	_	Rv0481c	hypothetical protein
TTAGGCTAGGCTTAGTTGC*	-112	4.59032	*mmp*54	Rv0451c	mmpS4
GCAACTAAGCCTAGCCTAA	-139	4.54925	_	Rv0452	Transcriptional regulator
CTATGTGATACTGACCTGA	-42	4.5466	*glp*Q2	Rv0317c	Glycerophosphoryl diester phosphodiesterase
AGATGCTAGACTTTCCTGA	**-77**	**4.54327**	**_**	**Rv1404**	**Transcriptional regulator**
TTACGGCAGCCTGTTGTAA	**-35**	**4.53876**	**_**	**Rv2663**	**hypothetical protein**
TTAGCTTATGCAATGCTAA*	-50	4.49914	_	Rv0282	hypothetical protein
TTCGGTAAGGCAACCCTTA*	-213	4.41965	_	Rv1348	hypothetical protein
TCACTGTAGTCTTAGCTGA	-179	4.39591	_	Rv0698	hypothetical protein
ATCCGTAAGTCTAAACTTA	-26	4.35929	_	Rv2034	Predicted transcriptional regulators
TTACTGCAATCTCCACTGA	-149	4.33623	*fad*A5	Rv3546	Acetyl-CoA acetyltransferases
TATGGCATGCCTAACCTAA	-31	4.02212	_	Rv1347c	Acetyltransferase

TTACCGCGCACTGCTCTAT	-17	3.51297	_	Rv1344	Acyl carrier protein
TATGGCATGCCTAACCTAA	-50	4.02212	_	Rv1347c	Acetyltransferase
GTAGGTTAGGACAGCCTTT	-102	3.92933	_	Rv0338c	Fe-S oxidoreductases
TAATGGCAGACTGTGATTC	-3	3.89219	*ppi*A	Rv0009	Peptidyl-prolyl cis trans isomerase

Sequences with asterisk (*) represent the experimentally confirmed IdeR binding sites. Sequences in bold represent the experimentally unverified novel IdeR binding sites predicted by the positional relative entropy method used in this study.

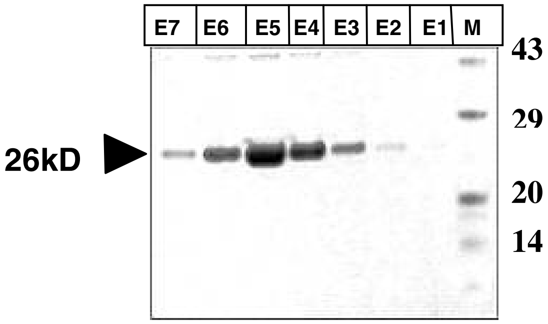


Figure 1

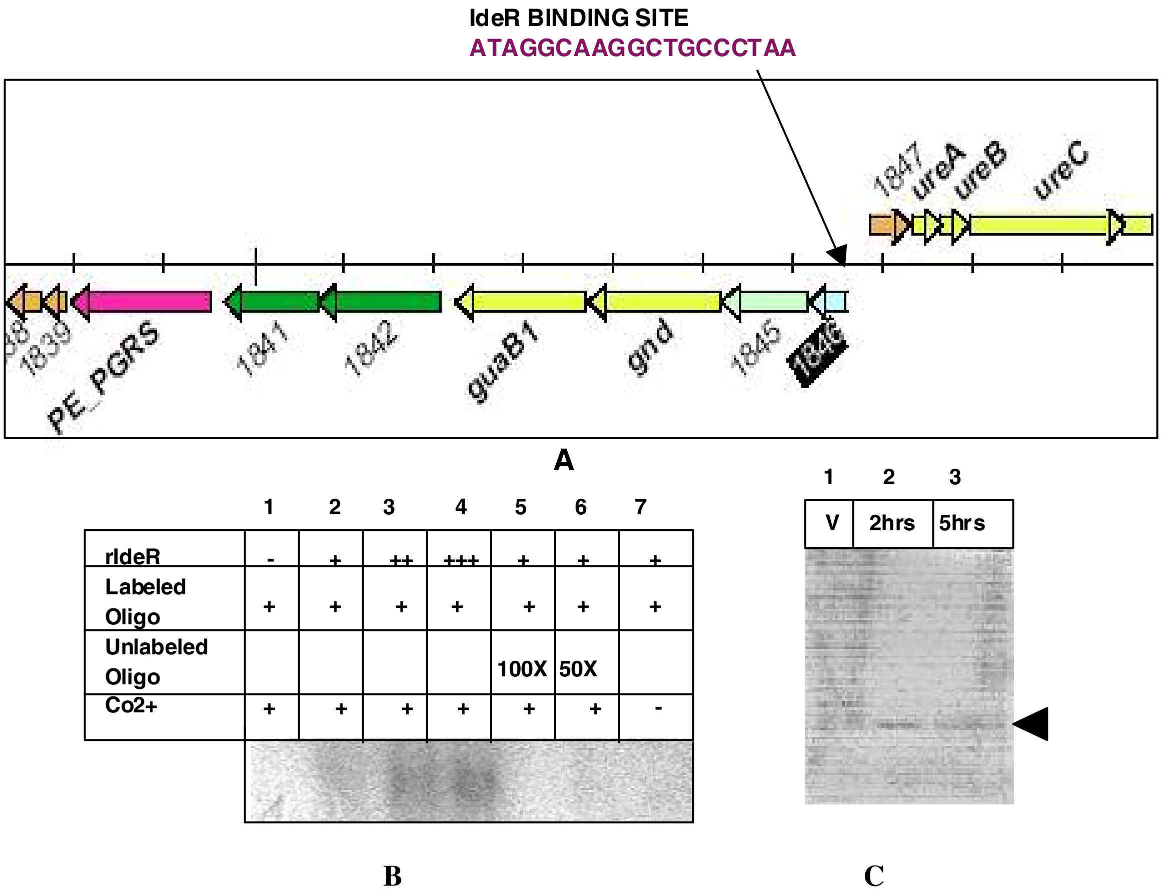


Figure 2

BMC Microbiology

Research article

Prediction of DtxR regulon: Identification of binding sites and operons controlled by Diphtheria toxin repressor in *Corynebacterium diphtheriae*

Sailu Yellaboina[1], Sarita Ranjan[1], Prachee Chakhaiyar[2], Seyed Ehtesham Hasnain[2] and Akash Ranjan*[1]

Address: [1]Computational and Functional Genomics Group, Centre for DNA Fingerprinting and Diagnostics, Hyderabad 500076, INDIA and [2]Laboratory of Cellular and Molecular Biology, Centre for DNA Fingerprinting and Diagnostics, Hyderabad 500076, INDIA

Email: Sailu Yellaboina - sailu@cdfd.org.in; Sarita Ranjan - sarita@cdfd.org.in; Prachee Chakhaiyar - prachee@cdfd.org.in; Seyed Ehtesham Hasnain - ehtesham@cdfd.org.in; Akash Ranjan* - akash@cdfd.org.in

* Corresponding author

Published: 24 September 2004

BMC Microbiology 2004, **4**:38 doi:10.1186/1471-2180-4-38

Received: 08 April 2004

Accepted: 24 September 2004

This article is available from: http://www.biomedcentral.com/1471-2180/4/38

Abstract

Background: The diphtheria toxin repressor, DtxR, of *Corynebacterium diphtheriae* has been shown to be an iron-activated transcription regulator that controls not only the expression of diphtheria toxin but also of iron uptake genes. This study aims to identify putative binding sites and operons controlled by DtxR to understand the role of DtxR in patho-physiology of *Corynebacterium diphtheriae*.

Result: Positional Shannon relative entropy method was used to build the DtxR-binding site recognition profile and the later was used to identify putative regulatory sites of DtxR within *C. diphtheriae* genome. In addition, DtxR-regulated operons were also identified taking into account the predicted DtxR regulatory sites and genome annotation. Few of the predicted motifs were experimentally validated by electrophoretic mobility shift assay. The analysis identifies motifs upstream to the novel iron-regulated genes that code for Formamidopyrimidine-DNA glycosylase (FpG), an enzyme involved in DNA-repair and starvation inducible DNA-binding protein (Dps) which is involved in iron storage and oxidative stress defense. In addition, we have found the DtxR motifs upstream to the genes that code for sortase which catalyzes anchoring of host-interacting proteins to the cell wall of pathogenic bacteria and the proteins of secretory system which could be involved in translocation of various iron-regulated virulence factors including diphtheria toxin.

Conclusions: We have used an *in silico* approach to identify the putative binding sites and genes controlled by DtxR in *Corynebacterium diphtheriae*. Our analysis shows that DtxR could provide a molecular link between Fe^{+2}-induced Fenton's reaction and protection of DNA from oxidative damage. DtxR-regulated Dps prevents lethal combination of Fe^{+2} and H_2O_2 and also protects DNA by nonspecific DNA-binding. In addition DtxR could play an important role in host interaction and virulence by regulating the levels of sortase, a potential vaccine candidate and proteins of secretory system.

Background

Iron is an important inorganic component of a cell. Iron is required as co-factor for various essential enzymes and proteins some of which are involved in electron transport (Cytochromes), redox reactions (oxidoreductases) and regulation of gene expression (fumarate-nitrate reduction regulatory protein, iron-binding protein) [1]. However a higher level of intracellular iron can catalyze formation of hydroxyl radicals and reactive oxygen species through Fenton's reaction which could be lethal to the cell [2]. Hence, a careful regulation of iron-requiring enzymes/ proteins and iron uptake proteins/enzymes is required for the survival of bacteria.

Inorganic iron is also known to influence virulence in many pathogenic bacteria such as *Corynebacterium diphtheriae*, *Escherichia coli*, and *Bordetella bronchiseptica* [3-5]. The diphtheria toxin repressor DtxR is known as an iron-activated global transcription regulator that represses the transcription of various iron-dependent genes in *C. diphtheriae* [6,7]. Eight DtxR-binding sites in upstream sequences of operons/genes named as *tox, hmuO, irp1, irp2, irp3, irp4, irp5 and irp6* have been identified by DNA footprinting methods [6]. The product of *tox* gene is diphtheria toxin which catalyzes the NAD-dependent ADP ribosylation of eukaryotic aminoacyl-transferase-II, thereby causing inhibition of protein synthesis and subsequent death of the host. The *hmuO* gene, which encodes a haem oxygenase, oxidizes the haem to release free iron. The operons *irp1* and *irp6* encode the products with homology to ABC-type ferric-siderophore transport systems. The gene *irp3* encodes a homologue of AraC-type transcriptional activators. The products of *irp2, irp4 and irp5* do not show any homology to the other known proteins. In addition, *C. diphtheriae* with inactive DtxR has been shown to be sensitive to killing by exposure to high iron conditions or hydrogen peroxide than the wild type [8].

This work uses an *in silico* method to identify additional DtxR-binding sites and target genes to understand the role of DtxR in virulence and patho-physiology of *C. diphtheriae*.

Results

In silico *identification of putative DtxR-binding sites*

Experimentally characterized DtxR-binding motifs were collected from the literature (Table 1). These binding sites were used to identify additional putative DtxR-binding sites along with associated operons in C. *diphtheriae* NCTC13129 genome (see materials and methods). Table 2 shows the predicted DtxR-binding sites with score 3.7438 or more. We could identify five (tox, irp4, irp5, irp6 and hmuO) of the eight known DtxR-binding sites, in sequenced *C. diphtheriae* NCTC13129 genome. We could not find irp1 and irp2 motifs as the corresponding genes (*irp1, irp2*) are not present in the sequenced strain NCTC13129 [9]. The regulator binding sites of *irp3*, *irp4* and *irp6* genes in the strain NCTC13129 shows one base change from the binding sites reported in strain C7 [6]. Binding site of *irp3* gene (TTAGGTGAGACGCACCCAT) although exists in strain NCTC13129, but not there in the predicted sites, because it is located within the coding region of *irp3* ORF. The predicted ORF of *irp3* in the sequenced strain NCTC13129 has different start position and is larger than what was previously reported in strain C7 [9,10].

In addition, we have identified binding sites in upstream sequences of eight genes recently reported to be regulated by DtxR [7]. However, our prediction differs from the previous report for five (secY, deoR, chtA, frgA, sidA) of the seven sites which were identified by BLAST search (Table 2). Our prediction agreed with the previous report that the genes such as *recA* (DIP1450) and *ywjA* (DIP1735) are not under a direct DtxR regulation as we could not detect any motif upstream to these gene with scores above the cutoff value [7].

Table 1: Known DtxR-binding sites from *C. diptheriae*

Binding site	Gene	Product	Reference
TTAGGATAGCTTTACCTAA	*tox*	Diphtheria toxin	[25]
TTAGGTTAGCCAAACCTTT	*Irp1*	Periplasmic protein of siderophore transport system	[26]
GCAGGGTAGCCTAACCTAA	*Irp2*	Hypothetical protein	[26]
TTAGGTGAGACGCACCCAT	*Irp3*	AraC-type transcription regulator	[10]
ATTACTAACGCTAACCTAA	*Irp4*	Hypothetical protein	[10]
CTAGGATTGCCTACACTTA	*Irp5*	Hypothetical protein	[10]
TTTCCTTTGCCTAGCCTAA	*Irp6*	Periplasmic protein of siderophore transport system	[6]
TGAGGGGAACCTAACCTAA	*hmuO*	Haem oxygenase	[27]

Table 2: Predicted DtxR-binding sites in *C. diphtheriae*

Score	Position	Site	Gene	Synonym	Product
4.45904	-80	TGAGGGGAACCTAACCTAA	*hmuO*	DIP1669**	heme oxygenase
4.39003	-52	TTAGGATAGCTTTACCTAA	*Tox*	DIP0222**	Diphtheria toxin precursor
4.25877	-60	ATAGGCTACACTTACCTAA	-	DIP0624	Putative membrane protein
4.21068	-168	TTGGATTAGCCTACCCTAA	-	DIP2162**	ABC-type peptide transport system periplasmic component
4.2033	-21	*TTAGGGTAGCTTCGCCTAA*	*iucA*	DIP0586	Putative siderophore biosynthesis related protein
4.17632	-78	ATAGGCATGCCTAACCTCA	-	DIP2330	Putative membrane protein
4.07921	-130	TTAGGTCAGGGTACCCTAA	-	DIP0370	Putative succinate dehydrogenease cytochrome B subunit
4.03559	-30	TTAGCTTAACCTTGCCTAT	*arsR*	DIP0415	Putative ArsR family regulatory protein
4.01967	-239	*TTAGGGTAGGCTAATCCAA*	*sidA**	DIP2161	nonribosomal peptide synthase
3.99985	-74	TTTTCTTTGCCTAGCCTAA	*irp6A*	DIP0108**	Ferrisiderophore receptor Irp6A
3.99195	-241	TTAGGCACCCCTAACCTAG	-	DIP0539	Putative sugar ABC transport syste ATP-binding protein
3.98554	-72	TTAGCTTAGCCCTAGCTAA	-	DIP0169	Putative secreted protein
3.9296	-26	CTAGGATTGCCTACACTTA	*Irp5*	DIP0894**	Conserved hypothetical protein
3.9073	-93	GTTGGGTTGCCCAACCTAC	-	DIP2106	Putative ABC transport system, ATP-binding subunit
3.89763	-86	ATAGGTTAGGTTAACCTTG	*chtA**	DIP1520	Putative membrane protein
3.89676	-130	*TTGTGTTAGCCTAGGCTAA*	*secA*	DIP0699	Translocase protein
3.89169	-26	*TTGGGGTGGCCTATCCTTA*	-	DIP2304	Putative DNA-repair glycosylase
3.88042	-172	TTAGGTAAGTGTAGCCTAT	*htaA**	DIP0625	Putative membrane protein
3.86534	-69	ATTACTAATGCTAACCTAA	*Irp4*	DIP2356**	Putative conserved membrane protein
3.85539	-173	TTAGGGTGGGCTAACCTGC	*deoR**	DIP1296	Putative DNA-binding protein
3.84889	-75	TTAGGGAACTCTTGCCTTA	*piuB**	DIP0124	Putative membrane protein
3.83816	-121	TTAGCTAGGGCTAAGCTAA	-	DIP0168	Putative glycosyl transferase
3.83576	-219	GTAACAAAGGCAAGCCTAA	*xerD*	DIP1510	Putative integrase/recombinase
3.8224	-216	ATAGGCAAGGTTAAGCTAA	-	DIP0417	Putative membrane protein
3.81905	-47	GTTGGACAGGTTACCCTAA	*frgA**	DIP1061	Putative iron-siderophore uptake system permease
3.8148	-37	*TGTGGGCACACCAACCTAA*	-	DIP2272	possible sortase-like protein
3.76235	-136	TTGGGGTTGCCCTTCCTAA	-	DIP0142	Hypothetical protein
3.76233	-268	CTAGGTTAGGGGTGCCTAA	*secY**	DIP0540	preprotein translocase SecY subunit
3.74673	-110	TAAACATAGCCAAACCAAA	*nrdF1*	DIP1865	ribonucleotide reductase beta-chain 1
3.7438	-81	TAAGGATAGGCCACCCCAA	*Dps*	DIP2303	Starvation inducible DNA-binding protein

Note: **Indicate the gene synonym with experimentally identified binding site in *C. diphtheriae* [6]. * Indicates the genes known to be regulated by DtxR [7]. The binding sites in Italics were verified by EMSA. The gene pairs, DIP0624-DIP0625, DIP2161-DIP2162, DIP0168-DIP0169, DIP0539-DIP0540 and DIP2303-DIP2304 are divergently transcribed and contain common regulatory regions.

Experimental validation of predicted binding sites

Since our approach to identify DtxR-regulated genes is purely computational in nature, we decided to test the validity of our predictions. A sample of predicted regulator binding motifs (Table 2) (upstream to ORFs: DIP2161, DIP0699, DIP0586, DIP2304, DIP2272) were experimentally verified by EMSA using IdeR, an orthologue of DtxR from *M. tuberculosis*. DtxR and IdeR are iron-dependent regulators. A pair wise sequence comparison of the two proteins shows a high (58%) overall sequence identity (similarity 72%) which increases further to 92% identity and 100% similarity in DNA recognition domain. In addition, the structural comparison of two regulators also shows a very similar 3D organization, suggesting that the IdeR regulator would be able to recognize the DtxR motif [11].

Synthetic double stranded oligonucleotides corresponding to DNA-binding sites were labeled with ^{32}P and mixed with purified IdeR in presence of manganese ions and was assayed for the formation of DNA-protein complex using EMSA. Manganese was used as the divalent metal in the binding reactions on account of its redox stability compared with ferrous ion. Electrophoretic mobility of all five double stranded oligonucleotides tested was retarded by IdeR (Figure 1). However a synthetic motif (TTTTCATGACGTCTTCTAA) used as a negative control did not show any complex formation. These results indicate that the predicted DtxR-binding sites can indeed bind to DtxR.

Identification and annotation of DtxR-regulated genes C. diphtheriae genome

In addition to the binding site prediction, we have also identified co-regulated genes (operons) downstream to the predicted DtxR-binding site (Table 3). Function of the proteins encoded by the putative genes in Table 2 and Table 3 was predicted by RPS-BLAST search against conserved domain database [12].

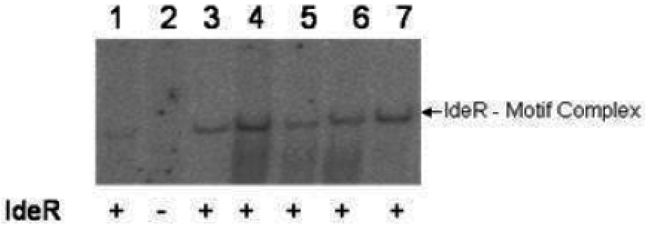


Figure 1
IdeR binds the predicted DtxR-binding DNA fragments. 30 pmoles of IdeR was added to ^{32}P-labelled DNA probes in the presence of 200 μM Mn^{2+}, and complexes were resolved on a 7% Tris-borate polyacrylamide gel containing 150 μM Mn^{2+}. Lane 1: Control gel retardation using Radiolabeled DNA motif without DtxR-binding site. Lane 2: Radiolabeled DIP2161 motif without IdeR. Lane 3: Radiolabeled DIP2161 motif with IdeR. Lane 4: Radiolabeled DIP0699 motif with IdeR. Lane 5: Radiolabeled DIP0586 motif with IdeR. Lane 6: Radiolabeled DIP2304 motif with IdeR. Lane 7: Radiolabeled DIP2272 motif with IdeR.

Discussion

Our analysis identified putative DtxR motifs upstream to various operons/genes which could be involved in siderophore biosynthesis, ABC-type transport systems, iron storage, oxidative stress defense and iron-sulfur cluster biosynthesis. In addition, we have also identified the motifs upstream of operons that could be involved in anchoring of host-interacting proteins to the cell wall and secretion of various virulence factors. Important functions of some of these DtxR-regulated genes and their role in *C. diphtheriae* physiology are discussed here.

Regulation of siderophore biosynthesis and ABC-type transport systems

Predicted member of the DtxR regulon, the gene DIP0586, codes for the IucA/IucC family of enzymes that catalyze discrete step in the biosynthesis of the aerobactin [13]. In addition to known DtxR-regulated siderophore transport genes (irp1, irp6), DtxR could also regulate other ABC-type transport systems similar to Manganese/Zinc, peptide/Nickel and multidrug subfamilies of ABC transporters. The peptide/nickel transport system (DIP2162-DIP2165) has been suggested to be recently acquired by pathogenic *C. diphtheriae*[9].

Regulation of iron storage and oxidative stress defense

We predict that DtxR could regulate divergently transcribed genes DIP2303 and DIP2304 whose products are similar to starvation inducible DNA-binding protein (Dps) and Formamidopyrimidine-DNA glycosylase (Fpg), respectively. Dps in *Escherichia coli* is induced in response to oxidative or nutritional stress and protects DNA from oxidative stress damage by nonspecific binding [14]. Dps also catalyzes oxidation of ferrous iron to ferric iron by hydrogen peroxide ($2Fe^{2+} + H_2O_2 + 2H_2O \rightarrow 2Fe^{+3}OOH_{(core)} + 4H^+$) which in turn prevents hydroxyl radical formation by Fenton's reaction ($Fe^{2+} + H_2O_2 \rightarrow Fe^{+3} + HO^- + HO\cdot$) and thereby prevents subsequent DNA damage [15]. The enzyme, formamidopyrimidine-DNA glycosylase is a primary participant in the repair of 8-oxoguanine, an abundant oxidative DNA lesion [16]. The gene DIP1510 which codes for the site-specific recombinase XerD could also be regulated by DtxR. The *xerD* gene in *E. coli* belongs to the oxidative stress regulon [17].

Regulation of proteins involved in iron-sulfur cluster biosynthesis and iron-sulfur cluster containing proteins

We predict that the operon DIP1288-DIP1296, which is similar to the *suf* operon of *E. coli*, could be regulated by DtxR. The *suf* operon in bacteria encodes the genes for Fe-S cluster assembly machinery [18]. In addition, genes encoding the iron-sulfur containing proteins such as succinate dehydrogenase (Sdh), cytochrome oxidase (CtaD) and Ribonucleotide reductase (NrdF1) in *C. diphtheriae* also show DtxR motif in their upstream sequences.

Regulation of sortases

We predict that DtxR could regulate the recently acquired pathogenic island DIP2271-DIP2272, encoding the sortase srtA and hypothetical protein, respectively [9]. Sortases are membrane-bound trans-peptidases that catalyze the anchoring of surface proteins to the cell wall peptidoglycan [9]. Such systems are often used by gram-positive pathogens to anchor host-interacting proteins to the bacterial surface[19].

Regulation of protein translation and translocation system

DtxR could regulate two operons that contain genes DIP0699 (*secA*) and DIP0540 (*secY*) that code for the protein translocation system. The *sec*Y-containing operon, which is similar to the streptomycine operon spc from *B. subtilis* and other bacteria, involves the genes required for protein translation and translocation [20]. The operon contains additional sialidase gene (DIP0543) in comparison to non pathogenic Corynebacterium species. Activity of sialidase has been linked to virulence in several other microbial pathogens and may enhance fimbriae mediated adhesion in *Corynebacterium diphtheriae* by unmasking receptors on mammalian cells [9].

The Sec system can both translocate proteins across the cytoplasmic membrane and insert integral membrane proteins into it. The former proteins but not the latter possess N-terminal, cleavable, targeting signal sequences that are required to direct the proteins to the Sec system. Some of the DtxR-regulated genes including diphtheria toxin (Table 4) show predicted signal sequences by SignalP 3.0 [21] and hence they may play an important role in host interaction and virulence of *Corynebacterium diphtheriae* [9].

Table 3: Predicted DtxR-regulated operons in *C. diphtheriae*

Synonym	Gene	Orthologue	Product
DIP2158		COG1131	ABC-type transport system permease and ATPase component
DIP2159		COG1131	ABC-type transport system permease and ATPase component
DIP2160	-	COG3321	Polyketide synthase modules and related proteins
DIP2161*	-	COG1020	Non-ribosomal peptide synthetase modules and related proteins
DIP0586	*iucA*	Pfam04183	Catalyse discrete steps in biosynthesis of the siderophore aerobactin
DIP0587	-	-	Putative membrane protein
DIP0588	-	-	Putative membrane protein
DIP1059	*fepC*	COG1120	ABC-type cobalamin/Fe3+-siderophores transport systems
DIP1060	*fepG*	COG4779	ABC-type enterobactin transport system
DIP1061*	*fepD*	COG0609	ABC-type Fe3+-siderophore transport system
DIP2162	*ddpA*	COG0747	ABC-type peptide transport system periplasmic component
DIP2163	*ddpB*	COG0601	ABC-type peptide/nickel transport systems permease components
DIP2164	*ddpC*	COG1173	ABC-type peptide/nickel transport systems permease components
DIP2165	*dpdD*	COG0444	ABC-type peptide/nickel transport systems ATPase component
DIP0169	*lraI*	COG0803	ABC-type metal ion transport system, periplasmic component
DIP0170	*znuC*	COG1121	ABC-type Mn/Zn transport systems, ATPase component
DIP0171	*znuB*	COG1108	ABC-type Mn2+/Zn2+ transport systems, permease components
DIP0172	*znuB*	COG1108	ABC-type Mn2+/Zn2+ transport systems, permease components
DIP0173	*lraI*	COG0803	ABC-type metal ion transport system, periplasmic component
DIP2106	*mdlB*	COG1131	ABC-type multidrug transport system, ATPase and permease component
DIP2107	*mdlB*	COG1131	ABC-type multidrug transport system, ATPase and permease component
DIP0625	*htaa*	Pfam04213	Haemin transporter associated protein
DIP0626	*hmuT*	COG4558	ABC-type haemin transport system
DIP0627	*hmuU*	COG0609	ABC-type Fe3+-siderophore transport system
DIP0628	*hmuV*	COG4559	ABC-type haemin transport system
DIP0629*	*htaa*	Pfam04213	Haemin transporter associated protein
DIP1519*	*htaa*	pfam04213	Haemin transporter associated protein
DIP1520*	*htaa*	pfam04213	Haemin transporter associated protein
DIP2303	*dps*	COG0783	Starvation inducible DNA-binding protein
DIP2304	-	COG0266	Formamidopyrimidine-DNA glycosylase
DIP2305	-	COG0063	Predicted sugar kinase
DIP1510	*xerD*	COG4974	Site-specific recombinase
DIP1288	-	-	Conserved hypothetical protein
DIP1289	*uup*	COG0488	ATPase components of ABC transporters with duplicated ATPase domains
DIP1290	-	COG2151	Predicted metal-sulfur cluster biosynthetic enzyme
DIP1291	*iscU*	COG0822	NifU homolog involved in Fe-S cluster formation
DIP1292	*csd*	COG0520	Selenocysteine lyase
DIP1293	*sufC*	COG0396	ABC-type transport system involved in Fe-S cluster assembly
DIP1294	-	COG0719	ABC-type transport system involved in Fe-S cluster assembly
DIP1295	*sufB*	COG0719	ABC-type transport system involved in Fe-S cluster assembly
DIP1296*	*deoR*	COG2345	DeoR family transcriptional regulator
DIP0370	-	-	Putative succinate dehydrogenease (cytochrome b)
DIP0371	-	COG1053	Succinate dehydrogenase/fumarate reductase
DIP0372	-	COG0479	Succinate dehydrogenase/fumarate reductase
DIP0373	-	-	Putative membrane protein
DIP0374	-	-	Putative membrane protein

Table 3: Predicted DtxR-regulated operons in *C. diphtheriae* *(Continued)*

DIP0375	-	-	Putative membrane protein
DIP0376	-	-	Putative membrane protein
DIP0377	-	-	Putative membrane protein
DIP1864	*ctaD*	COG0843	Heme/copper-type cytochrome/quinol oxidases
DIP1865	*nrdF1*	COG0208	Ribonucleotide reductase
DIP2330	-	-	Putative membrane protein
DIP2331	-	COG1012	NAD-dependent aldehyde dehydrogenases
DIP0124*	-	Pfam03929	Uncharacterized iron-regulated membrane protein (DUF337)
DIP0622	-	-	Putative membrane protein
DIP0623	*metA*	COG2021	Homoserine acetyltransferase
DIP0624	-	-	Putative membrane protein
DIP0415	-	Pfam01022	Bacterial regulatory protein
DIP0539	-	COG3839	ABC-type sugar transport systems
DIP0168	-	-	Putative glycosyl transferase
DIP0417	-	-	Putative membrane protein
DIP0142	-	-	Hypothetical protein
DIP0143	-	-	-
DIP0144	*tra8*	COG2826	Transposase and inactivated derivatives
DIP2271	-	-	Putative membrane protein
DIP2272	-	COG3764	Sortase (surface protein transpeptidase)
DIP0699	*secA*	COG0653	Preprotein translocase subunit SecA (ATPase
DIP0700	-	-	Hypothetical protein
DIP0540*	*secY*	Pfam00344	Eubacterial secY protein
DIP0541	*Adk*	COG0563	Adenylate kinase and related kinases
DIP0542	*mapA*		Methionine aminopeptidase
DIP0543	-	-	Sialidases or neuraminidases;
DIP0544	*erfK*	Pfam03734	This family of proteins contains a conserved histidine and cysteine
DIP0545	*infA*	COG0361	Translation initiation factor 1 (IF-1)
DIP0546	*rpsM*	COG0099	Ribosomal protein S13
DIP0547	*rpsK*	COG0100	Ribosomal protein S11
DIP0548	*rpsD*	COG0522	Ribosomal protein S4 and related proteins
DIP0549	*rpoA*	COG0202	DNA-directed RNA polymerase
DIP0550	*rplQ*	COG0203	Ribosomal protein L17
DIP0551	*truA*	COG0101	Pseudouridylate synthase

Note: * Indicate the genes reported be regulated by DtxR. Genes listed together belongs to same operon.

Conclusions

The bioinformatics method used to predict the targets of DtxR in *C. diphtheriae* NCTC13129 genome is promising, as some of the predicted targets were experimentally verified. The approach identified novel DtxR-regulated genes, which could play an important role in physiology of *C. diphtheriae* NCTC13129. DtxR, generally known as a repressor of diphtheriae toxin and iron siderophore/transport genes, can also regulate other metal ion transport genes, iron storage, oxidative stress, DNA-repair, biosynthesis of iron-sulfur cluster, Fe-S-cluster containing proteins, and even protein sortase and translocation systems.

Methods

Source of genome sequence

The complete genome sequence of *C. diphtheriae* was downloaded from NCBI ftp site [22], and the DtxR-binding sites identified by experimental methods were collected from literature [6,10,25-27].

Table 4: DtxR-regulated genes containing the potential signal sequence

Gene	Product
DIP0222	Diphtheria toxin
DIP0109	IRP6B
DIP2356	IRP4
DIP2162	ABC-type peptide transport system periplasmic component
DIP0172	Putative membrane protein
DIP2107	Putative integral membrane transport protein
DIP0625	Haemin transporter associated protein
DIP0626	ABC-type haemin transport system
DIP0627	ABC-type haemin transport system
DIP1519	Haemin transporter associated protein
DIP0629	Haemin transporter associated protein
DIP1520	Haemin transporter associated protein
DIP2330	Putative membrane protein
DIP0543	Sialidases or neuraminidases

Prediction of DtxR-binding sites

DtxR-binding site recognition profile was calculated by positional Shannon relative entropy method [23,24]. The positional relative entropy Q_i at position *i* in a binding site is defined as

$$Q_i = \sum_{b=A,T,G,C} f_{b,i} \log_{10} \frac{f_{b,i}}{q_b}$$

where *b* refers to each of the possible base (A, T, G, C), $f_{b,i}$ is observed frequency of each base at position *i* and q_b is the frequency of base *b* in the genome sequence. The contribution of each base to the positional Shannon's relative entropy is calculated by multiplying positional frequency of each base with positional relative entropy. The binding site profile thus generated was used to scan upstream sequences of all the genes of the *Corynebacterium diphtheriae* genome. The score of each site is calculated as the sum of the respective positional Shannon relative entropy of each of the four possible bases. A maximally scoring site is selected from the upstream sequence of each gene. The lowest score among the input binding sites is considered as cut-off score. The sites scoring higher than the cut-off value are reported as potential binding sites conforming to the consensus sequence.

Prediction of operons

Co-directionally transcribed genes, downstream to the predicted binding site were selected as potential co-regulated genes (operons) according to one of the following criteria (a) Co-directionally transcribed orthologous gene pairs, conserved in at least 4 genomes; (b) genes belong to the same cluster of orthologous gene function category and the intergenic distance is less than 200 base pairs; (c) the first three letters in gene names are identical (gene names for putative genes were assigned from COG database); (d) intergenic distance is less than 90 base pairs [24].

Functional assignment of genes

The function of predicted genes was inferred using the RPS-BLAST search against conserved domain database [12]. These genes were further classified according to their function.

Expression and purification of IdeR

The iron-dependent regulator IdeR from M. *tuberculosis* was expressed from a recombinant pRSET vector containing the IdeR gene fused to a six His affinity tag (P. Chakhiyar unpublished). The expressed protein was first purified using Ni-NTA Metal Chelate Affinity chromatography; later it was desalted and concentrated using Centricon Ultra filtration device. The concentration of the recombinant protein was estimated using Bradford method.

Electrophoretic mobility shift assay

Double-stranded oligonucleotides containing the predicted binding motif (19 bp long) were end labeled with T4 polynucleotide kinase and [γ^{32}P]-ATP and were incubated with the recombinant purified IdeR protein in a binding reaction mixture. The binding reaction mixture (20-μl total volume) contain the DNA-binding buffer (20 mM Tris-HCl [pH 8.0], 2 mM DTT, 50 mM NaCl, 5 mM $MgCl_2$, 50% glycerol, 5 μg of bovine serum albumin per ml), 10 μg of poly(dI-dC) per ml (for nonspecific binding) and 200 μM $MnCl_2$. The reaction mixture was incubated at room temperature for 30 min. Approximately 2 μl of the tracking dye (50% sucrose, 0.6% bromophenol blue) was added to the reaction mixture at the end of incubation and was loaded onto 7% polyacrylamide gel containing 150 μM $MnCl_2$ in 1 × Tris-borate-EDTA buffer. The gel was electrophoresed at 200 V for 2 hours. Subsequently the gel was dried and exposed to Fuji Storage Phosphor Image Plates for 16 hours. The image plates were subsequently scanned in Fuji Storage Phosphor Imaging workstation.

List of abbreviations

DtxR – Diphtheria toxin repressor; IdeR – Iron-dependent regulator; Dps – DNA-binding protein from starved cells; RPS-BLAST – Reversed Position Specific – Basic Local Alignment Search Tool; EMSA – Electrophoretic Mobility Shift Assay

Authors' contributions

SY: carried out the computation, data analysis, and manuscript preparation. SR: Carried out the EMSA and drafted the manuscript. PC: provided the cloned IdeR construct, drafted the manuscript. SH: Manuscript preparation and coordination. AR: Design of the study and coordination. All authors read and approved the final manuscript.

Acknowledgements

This work is partially supported by CSIR NMITLI Grant to AR. SR is supported by CSIR NMITLI Grant. YS and PC is supported by CSIR Research Fellowships.

References

1. Castagnetto JM, Hennessy SW, Roberts VA, Getzoff ED, Tainer JA, Pique ME: **MDB: the Metalloprotein Database and Browser at The Scripps Research Institute.** *Nucleic Acids Res* 2002, **30:**379-382.
2. Urbanski NK, Beresewicz A: **Generation of *OH initiated by interaction of Fe2+ and Cu+ with dioxygen; comparison with the Fenton chemistry.** *Acta Biochim Pol* 2000, **47:**951-962.
3. Tao X, Schiering N, Zeng HY, Ringe D, Murphy JR: **Iron, DtxR, and the regulation of diphtheria toxin expression.** *Mol Microbiol* 1994, **14:**191-197.
4. Russo TA, Carlino UB, Johnson JR: **Identification of a new iron-regulated virulence gene, ireA, in an extraintestinal pathogenic isolate of *Escherichia coli*.** *Infect Immun* 2001, **69:**6209-6216.
5. Register KB, Ducey TF, Brockmeier SL, Dyer DW: **Reduced virulence of a Bordetella bronchiseptica siderophore mutant in neonatal swine.** *Infect Immun* 2001, **69:**2137-2143.
6. Qian Y, Lee JH, Holmes RK: **Identification of a DtxR-regulated operon that is essential for siderophore-dependent iron uptake in *Corynebacterium diphtheriae*.** *J Bacteriol* 2002, **184:**4846-4856.
7. Kunkle CA, Schmitt MP: **Analysis of the *Corynebacterium diphtheriae* DtxR Regulon: Identification of a putative siderophore synthesis and transport system that is similar to the Yersinia high-pathogenicity island-encoded yersiniabactin synthesis and uptake system.** *J Bacteriol* 2003, **185:**6826-6840.
8. Oram DM, Avdalovic A, Holmes RK: **Construction and characterization of transposon insertion mutations in *Corynebacterium diphtheriae* that affect expression of the diphtheria toxin repressor (DtxR).** *J Bacteriol* 2002, **184:**5723-5732.
9. Cerdeno-Tarraga AM, Efstratiou A, Dover LG, Holden MT, Pallen M, Bentley SD, Besra GS, Churcher C, James KD, De Zoysa A, Chillingworth T, Cronin A, Dowd L, Feltwell T, Hamlin N, Holroyd S, Jagels K, Moule S, Quail MA, Rabbinowitsch E, Rutherford KM, Thomson NR, Unwin L, Whitehead S, Barrell BG, Parkhill J: **The complete genome sequence and analysis of *Corynebacterium diphtheriae* NCTC13129.** *Nucleic Acids Res* 2003, **31:**6516-6523.
10. Lee JH, Wang T, Ault K, Liu J, Schmitt MP, Holmes RK: **Identification and characterization of three new promoter/operators from *Corynebacterium diphtheriae* that are regulated by the diphtheria toxin repressor (DtxR) and iron.** *Infect Immun* 1997, **65:**4273-4280.
11. Feese MD, Ingason BP, Goranson-Siekierke J, Holmes RK, Hol WG: **Crystal structure of the iron-dependent regulator from Mycobacterium tuberculosis at 2.0-A resolution reveals the Src homology domain 3-like fold and metal binding function of the third domain.** *J Biol Chem* 2001, **276:**5959-66.
12. Marchler-Bauer A, Anderson JB, DeWeese-Scott C, Fedorova ND, Geer LY, He S, Hurwitz DI, Jackson JD, Jacobs AR, Lanczycki CJ, Liebert CA, Liu C, Madej T, Marchler GH, Mazumder R, Nikolskaya AN, Panchenko AR, Rao BS, Shoemaker BA, Simonyan V, Song JS, Thiessen PA, Vasudevan S, Wang Y, Yamashita RA, Yin JJ, Bryant SH: **CDD: a curated Entrez database of conserved domain alignments.** *Nucleic Acids Res* 2003, **31:**383-387.
13. de Lorenzo V, Neilands JB: **Characterization of iucA and iucC genes of the aerobactin system of plasmid ColV-K30 in *Escherichia coli*.** *J Bacteriol* 1986, **167:**350-355.
14. Martinez A, Kolter R: **Protection of DNA during oxidative stress by the non specific DNA-binding protein Dps.** *J Bacteriol* 1997, **179:**5188-5194.
15. Zhao G, Ceci P, Ilari A, Giangiacomo L, Laue TM, Chiancone E, Chasteen ND: **Iron and hydrogen peroxide detoxification properties of DNA-binding protein from starved cells. A ferritin-like DNA-binding protein of *Escherichia coli*.** *J Biol Chem* 2002, **277:**27689-27696.
16. Zaika EI, Perlow RA, Matz E, Broyde S, Gilboa R, Grollman AP, Zharkov DO: **Substrate discrimination by formamidopyrimidine-DNA glycosylase: a mutational analysis.** *J Biol Chem* 2004, **279:**4849-4861.
17. Gaudu P, Weiss B: **Flavodoxin mutants of *Escherichia coli K-12*.** *J Bacteriol* 2000, **182:**1788-1793.
18. Outten FW, Wood MJ, Munoz FM, Storz G: **The SufE protein and the SufBCD complex enhance SufS cysteine desulfurase activity as part of a sulfur transfer pathway for Fe-S cluster assembly in *Escherichia coli*.** *J Biol Chem* 2003, **278:**45713-45719.
19. Ton-That H, Schneewind O: **Assembly of pili on the surface of *Corynebacterium diphtheriae*.** *Mol Microbiol* 2003, **50:**1429-1438.
20. Suh JW, Boylan SA, Oh SH, Price CW: **Genetic and transcriptional organization of the *Bacillus subtilis* spc-alpha region.** *Gene* 1996, **169:**17-23.
21. Jannick DB, Henrik N, Gunnar VH, Søren B: **Improved prediction of signal peptides: SignalP 3.0.** *J Mol Biol* 2004, **340:**783-795.
22. **NCBI FTP site** [ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Corynebacterium_diphtheriae]
23. Shannon CE: **A mathematical theory of communication.** *Bell System Technical Journal* 1948:379-423. 623–656
24. Yellaboina S, Seshadri J, Kumar MS, Ranjan A: **PredictRegulon: A webserver for the prediction of the regulatory protein binding sites and operons in prokaryote genomes.** *Nucleic Acids Res* 2004, **32:**W318-W320.
25. Tao X, Murphy JR: **Binding of the metalloregulatory protein DtxR to the diphtheria tox operator requires a divalent heavy metal ion and protects the palindromic sequence from DNase I digestion.** *J Biol Chem* 1992, **267:**21761-21764.
26. Schmitt MP, Holmes RK: **Cloning, sequence, and footprint analysis of two promoter/operators from Corynebacterium diphtheriae that are regulated by the diphtheria toxin repressor (DtxR) and iron.** *J Bacteriol* 1994, **176:**1141-1149.
27. Schmitt MP: **Transcription of the Corynebacterium diphtheriae hmuO gene is regulated by iron and heme.** *Infect Immun* 1997, **65:**4634-4641.

W318–W320 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh364

PredictRegulon: a web server for the prediction of the regulatory protein binding sites and operons in prokaryote genomes

Sailu Yellaboina[1], Jayashree Seshadri[1], M. Senthil Kumar[2] and Akash Ranjan[1,*]

[1]Computational & Functional Genomics Group and [2]Molecular Oncology Laboratory, Centre for DNA Fingerprinting and Diagnostics, EMBnet India Node, Hyderabad 500076, India

Received February 15, 2004; Revised and Accepted March 4, 2004

ABSTRACT

An interactive web server is developed for predicting the potential binding sites and its target operons for a given regulatory protein in prokaryotic genomes. The program allows users to submit known or experimentally determined binding sites of a regulatory protein as ungapped multiple sequence alignments. It analyses the upstream regions of all genes in a user-selected prokaryote genome and returns the potential binding sites along with the downstream co-regulated genes (operons). The known binding sites of a regulatory protein can also be used to identify its orthologue binding sites in phylogeneticaly related genomes where the *trans*-acting regulator protein and cognate *cis*-acting DNA sequences could be conserved. PredictRegulon can be freely accessed from a link on our world wide web server: http://www.cdfd.org.in/predictregulon/.

INTRODUCTION

With over 100 bacterial genomes sequenced, a key challenge of post-genomic research is to dissect the complex transcription regulatory network which controls the metabolic and physiological process of a cell. A first step towards this goal is to identify the genes within a genome that are controlled by a specific transcription regulatory protein. This paper describes a web server tool—PredictRegulon—for genome-wide prediction of potential binding sites and target operons of a regulatory protein for which few experimentally identified binding sites are known. This technique could utilize the available experimental data on binding sites of transcription regulatory proteins from various bacterial species (1–3) for identification of regulons in phylogenetically related species.

PREDICTREGULON METHOD

The program, PredictRegulon, first constructs the binding site recognition profile based on ungapped multiple sequence alignment of known binding sites. This profile is calculated using Shannon's positional relative entropy approach (4). The positional relative entropy Q_i at position i in a binding site is defined as

$$Q_i = \sum_{b=\mathrm{A,T,G,C}} f_{b,i} \log_{10} \frac{f_{b,i}}{q_b},$$

where b refers to each of the possible bases (A, T, G, C), $f_{b,i}$ is observed frequency of each base at position i and q_b is the frequency of base b in the genome sequence. The contribution of each base to the positional Shannon relative entropy is calculated by multiplying each base frequency by positional relative entropy as follows:

$$W_{b,i} = f_{b,i} \cdot Q_i,$$

where $W_{b,i}$ refers to the weighted Shannon relative entropy of the base b (A, T, G, C) at position i. Finally, a $4 \times L$ entropy matrix (L is the length of the binding site) is constructed representing the binding site recognition profile, where each matrix element is the weighted positional Shannon relative entropy of a base.

The profile, encoded as the matrix, is used to scan the upstream sequences of all the genes of the user-selected genome. The entropy score of each site is calculated as the sum of the respective positional nucleotide entropy ($W_{b,i}$). A maximally scoring site is selected from the upstream sequence of each gene. The score may represent the strength of interaction between regulatory protein and binding site (5). The lowest score among the input sites is considered as the cut-off score. The sites scoring higher than the the cut-off value are reported as potential binding sites conforming to the consensus profile.

*To whom correspondence should be addressed. Tel: +9140 27171454; Fax: +9140 27155610; Email: akash@cdfd.org.in

Co-directionally transcribed genes downstream of the predicted binding site were selected as potential co-regulated genes (operons) according to one of the following criteria: (i) co-directionally transcribed orthologous gene pairs conserved in at least three genomes (6); (ii) genes belong to the same cluster of orthologous gene function category and the intergenic distance is <200 bp (7); (iii) the first three letters in gene names are identical (the gene names for all the bacterial species were assigned using the COG annotation); (iv) intergenic distance is <90 bp (8).

This method has two specific requirements: a few experimentally determined regulatory protein binding sites should be available for developing the binding site recognition profile, and the profile should be applicable to the genome where the regulator or its homologue is present. In the absence of any experimental information on the regulatory sites in a given genome one may look up the known regulatory motifs from other related species from one of the four online databases which host the information about known transcription regulatory protein binding sites in prokaryote genomes (1–3).

A limitation of this approach is that it may predict a few false positive sites as candidates. However, this limitation can be overcome by experimental validations, by either *in vitro* binding studies with double strand oligonucleotides containing the binding sites (designed based on prediction) and regulatory proteins or real-time PCR analysis of candidate co-regulated genes.

Table 1. Known LexA binding sites of *Bacillus subtilis* from the PRODORIC database

Binding site	Gene
AGAACAAGTGTTCG	*din*C
AGAACTCATGTTCG	*din*B
CGAACTTTAGTTCG	*din*A
CGAATATGCGTTCG	*rec*A
CGAACGTATGTTTG	*din*C
CGAACCTATGTTTG	*din*R
CGAACAAACGTTTC	*din*R
GGAATGTTTGTTCG	*din*R

Table 2. Output of PredictRegulon web server (predicted LexA binding sites)

Score	Position	Site	Gene	Synonym	COG	Product
5.37	−8	CGAACGTATGTTCG	—	Rv3776[a]	—	Hypothetical protein Rv3776
5.32	−100	CGAACATGTGTTCG	—	Rv3073c[a]	COG3189	Uncharacterized conserved protein
5.32	−144	CGAACATGTGTTCG	pyrR	Rv1379[a]	COG2065	Pyrimidine operon attenuation protein
5.22	−8	CGAACACATGTTCG	—	Rv3074[a]	—	Hypothetical protein Rv3074
5.2	−142	CGAACAATTGTTCG	—	Rv3371[a]	—	Hypothetical protein Rv3371
5.2	−64	CGAACAATTGTTCG	dnaE2	Rv3370c[a]	COG0587	DNA polymerase III
5.19	−36	CGAACGATTGTTCG	ruvC	Rv2594c[a]	COG0817	ruvC
5.14	−32	CGAAAGTATGTTCG	—	Rv0336[a]	—	Hypothetical protein Rv0336
5.14	−32	CGAAAGTATGTTCG	—	Rv0515[a]	—	Hypothetical protein Rv0515
5.14	−105	CGAACACATGTTTG	lexA	Rv2720[a]	COG1974	SOS-response transcriptional repressors
5.11	−122	CGAACAGGTGTTCG	recA	Rv2737c[a]	COG1372	recA
5.08	−87	CGAACAATCGTTCG	—	Rv2595[a]	COG2002	Hypothetical protein Rv2595
5.06	−44	CGAATATGCGTTCG	dnaB	Rv0058[a]	COG0305	Replicative DNA helicase
5.04	−263	GGAACTTGTGTTGG	ubiE	Rv3832c	COG2226	Methylase involved in ubiquinone biosynthesis
5.04	−23	AGAACGGTTGTTCG	splB	Rv2578c[a]	COG1533	DNA repair photolyase
5.02	−6	CGAATATGAGTTCG	—	Rv0071[a]	COG3344	Retron-type reverse transcriptase
5.01	−255	CGAACAAGTGTTGG	—	Rv1414	COG3616	Predicted amino acid aldolase or racemase
4.99	−181	GGAACGCGTGTTTG	—	Rv0750	—	Hypothetical protein Rv0750
4.98	−105	CGAACAACAGTTCG	baeS	Rv0600c	COG0642	Signal transduction histidine kinase
4.98	−186	CGAAGATGCGTTCG	rpsT	Rv2412	COG0268	Ribosomal protein S20
4.95	−242	TGAACGCAAGTTCG	fbpB	Rv1886c	COG0627	fbpB
4.95	−192	CGAACGGGAGTTCG	—	Rv1455	—	Hypothetical protein Rv1455
4.94	−270	AGAACCACCGTTCG	phd	Rv3181c	COG4118	Antitoxin of toxin–antitoxin stability system
4.94	−213	CGAACGACGGTTCG	pe	Rv2099c[a]	—	PE
4.92	−118	CGAACAGGTGTTGG	—	Rv0004	COG5512	Zn-ribbon-containing
4.92	−163	CGAACTTGCGTTCA	—	Rv1887	—	Hypothetical protein Rv1887
4.91	−239	GGAACGCGAGTTCG	fadB2	Rv0468	COG1250	3-hydroxyacyl-CoA dehydrogenase
4.91	−7	TGAACGAATGTTCC	—	Rv0039c	—	Hypothetical protein Rv0039c
4.9	−237	CGAAGCCTTGTTCG	dltE	Rv3174	COG0300	Short-chain dehydrogenase
4.89	−225	GGAAGGTGCGTTCG	frnE	Rv2466c	COG2761	Predicted dithiol-disulfide isomerase
4.88	−8	GGAAGCCATGTTCG	—	Rv0769	COG1028	Hypothetical protein Rv0769
4.88	−186	CGAAGAGGTGTTCG	coxS	Rv0374c	COG2080	Aerobic-type carbon monoxide dehydrogenase
4.88	−186	CGAACCGCAGTTCG	leuA	Rv3534c	COG0119	Isopropyl malate/citramalate synthases
4.85	−195	CGAACGGCTGTTGG	—	Rv2061c	COG3576	Hypothetical protein Rv2061c
4.85	−85	AGAACGGTTGTTGG	accA1	Rv2501c	COG4770	COG4770
4.84	−151	CGAAATTGTGTTCC	nuoB	Rv3146	COG0377	NADH:ubiquinone oxidoreductase
4.84	−217	CAAACATGTGTTCG	—	Rv2719c[a]	—	Hypothetical protein Rv2719c
4.84	−5	CGAACATGTATTCG	—	Rv1702c[a]	—	Hypothetical protein Rv1702c
4.84	−199	CGAAATCTTGTTTG	—	Rv1375	COG1944	Hypothetical protein Rv1375

Score: score of the binding sites, Position: position of the binding site relative to the translation start site, Site: binding site of a regulatory protein, Gene: gene downstream to the binding site, Synonym: synonym of the gene, COG: Cluster of Orthologous Gene code, Product: Gene product. [a] represents the ORFs known to be regulated by the regulator. 'a' symbols are not part of the orginal output of the web server. Source of Genome: NCBI ftp site (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Mycobacterium_tuberculosis_H37Rv/), Accession no. NC_000962.

EXAMPLE: PREDICTION OF LEXA REGULON IN *MYCOBACTERIUM TUBERCULOSIS*

To demonstrate a typical usage of PredictRegulon, we predicted the LexA binding sites and LexA regulon of *M.tuberculosis* using the LexA binding sites of *Bacillus subtilis*. LexA regulators from *B.subtilis* and *M.tuberculosis* share a high sequence identity (45%) at protein level (data not shown). Table 1 lists the known LexA binding sites from *B. subtilis* given as input to the program (2) and Table 2 shows the output of predicted LexA binding sites in *M.tuberculosis*. The site column in Table 2 represents the predicted binding sites of LexA in *M.tuberculosis*. In a typical output the perfect match to the known binding sites and the downstream genes are highlighted with a yellow background, and the rest with score greater than cut-off is shown with a blue background (colours not shown in the table). Eighteen of these genes (indicated by 'a') belonging to the LexA regulon were also observed in data obtained by experimental means by others (9–12). The rest of the matches are potential novel regulatory sites which could be confirmed experimentaly.

The web output of PredictRegulon also contains the hyperlinked gene-synonym and COG number. A click on the former shows the predicted operon context of the regulatory motif while a click on the latter opens a new page showing a description of this gene in the NCBI Conserved Domain Database, which is in turn linked to Pubmed for published information on this gene. These additional links provides users a simple way to browse and understand the functional/physiological implication of the genes that are part of predicted regulon.

ACKNOWLEDGEMENTS

This work is partially supported by the Council of Scientific and Industrial Research (CSIR) NMITLI Grant to A.R. Y.S. and S.K. were recipients of Senior Research Fellowships from CSIR, Govt. of India.

REFERENCES

1. Salgado,H., Santos-Zavaleta,A., Gama-Castro,S., Millan-Zarate,D., Diaz-Peredo,E., Sanchez-Solano,F., Perez-Rueda,E., Bonavides-Martinez,C. and Collado-Vides,J. (2001) RegulonDB (Version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12. *Nucleic Acids Res.*, **29**, 72–74.
2. Munch,R., Hiller,K., Barg,H., Heldt,D., Linz,S., Wingender,E. and Jahn,D. (2003) PRODORIC: prokaryotic database of gene regulation. *Nucleic Acids Res.*, **31**, 266–269.
3. Ishii,T., Yoshida,K., Terai,G., Fujita,Y. and Nakai,K. (2001) DBTBS: a database of *Bacillus subtilis* promoters and transcription factors. *Nucleic Acids Res.*, **29**, 278–280.
4. Shannon,C.E. (1948) A mathematical theory of communication. *Bell Sys. Tech. J.*, 379–423 and 623–656.
5. Benos,P.V., Bulyk,M.L. and Stormo,G.D. (2002) Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res.*, **30**, 4442–4451.
6. Ermolaeva,M.D., White,O. and Salzberg,S.L. (2001) Prediction of operons in microbial genomes. *Nucleic Acids Res.*, **295**, 1216–1221.
7. Salgado,H., Moreno-Hagelsieb,G., Smith,T.F. and Collado-Vides,J. (2000) Operons in *Escherichia coli*: genomic analyses and predictions *Proc. Natl Acad. Sci., USA*, **97**, 6652–6657.
8. Strong,M., Mallick P., Pellegrini,M., Thompson,M.J. and Eisenberg,D. (2003) Inference of protein function and protein linkages in *Mycobacterium tuberculosis* based on prokaryotic genome organization: a combined computational approach. *Genome Biol.*, **4**, R59.
9. Durbach,S.I., Andersen,S.J. and Mizrahi,V. (1997) SOS induction in mycobacteria: analysis of the DNA-binding activity of a LexA-like repressor and its role in DNA damage induction of the recA gene from *Mycobacterium smegmatis*. *Mol. Microbiol.*, **26**, 643–653.
10. Brooks,P.C., Movahedzadeh,F. and Davis,E.O. (2001) Identification of some DNA damage-inducible genes of *Mycobacterium tuberculosis*: apparent lack of correlation with LexA binding. *J. Bacteriol.*, **183**, 4459–4467.
11. Dullaghan,E.M., Brooks,P.C. and Davis,E.O. (2002) The role of multiple SOS boxes upstream of the *Mycobacterium tuberculosis* lexA gene—identification of a novel DNA-damage-inducible gene. *Microbiology*, **148**, 3609–3615.
12. Boshoff,H.I., Reed,M.B., Barry,C.E. and Mizrahi,V. (2003) DNAE2 polymerase contributes to *in vivo* survival and the emergence of drug resistance in *Mycobacterium tuberculosis*. *Cell*, **113**, 183–193.

Computational prediction of DtxR regulon-A Dissection of physiological process controlled by DtxR in *Corynebacterium* species

Sailu Yellaboina, Prachee Chakhaiyar, Seyed Ehetsham Hasnain and Akash Ranjan
EMB India Node, Centre for DNA Fingerprinting and Diagnostics, Hyderabad
akash@cdfd.org.in

Abstract

We developed a user friendly software tool to identify the potential binding sites of any regulatory protein based on Shannon relative entropy method. Known DtxR binding sites of Corynebacterium diphtheriae (C. diphtheriae) were used to generate a position specific reference profile for DtxR which was used to identify the potential DNA binding sites within the upstream sequences of Corynebacterium glutamicum (C. glutamicum) genes. In addition, DtxR regulated operons were also identified taking into account the predicted DtxR regulatory sites and Rho- independent transcription termination sites. The analysis predicted the binding sites upstream to a number of genes/operons which code for proteins involved in hemolysis and haemin transport. The analysis also predicts the binding sites upstream to genes that are involved in iron storage and oxidative stress defense including ferritin, starvation inducible DNA binding protein (Dps) and a homologue of endonuclease VIII (Nei). Both Dps and Nei homologue could be involved in controlling ferrous iron induced DNA damage.

1. Introduction

The diphtheria toxin repressor, DtxR, of *C. diphtheriae* has been shown to be a global transcription regulator that controls the expression of various genes including diphtheria toxin gene in response to iron levels in the host environment [1]. This study aimed to increase understanding of DtxR regulated genes and their role in cellular physiology of *C. glutamicum* and related species.

2. Methods

The complete genome sequence of *C. glutamicum* was downloaded from NCBI ftp site (ftp://ftp.ncbi.nlm.gov/). Experimentally identified DtxR binding sites were collected from literature [2].

DtxR binding site recognition profile was calculated by positional relative entropy method [3]. The relative entropy Q_i at a position *i* in a binding site is defined as

$$Q_i = \sum_{b=A,T,G,C} f_{b,i} \log_{10} \frac{f_{b,i}}{q_b}$$

Where *b* refers to each of the possible base (A, T, G, C), $f_{b,i}$ is observed frequency of each base at the position *i* and q_i is frequency of base *b* in the genome sequence. The contribution of each base to the positional Shannon relative entropy was calculated by multiplying each base frequency with positional Shannon relative entropy.

$$W_{b,i} = f_{b,i}.Q_i$$

Where $W_{b,i}$ refers to the weighted Shannon relative entropy of the base b (A, T, G, C) at position *i.* The DtxR binding site recognition profile was used to scan the upstream sequences of *C. glutamicum genes.* The score of the candidate site is calculated as the sum of the respective positional nucleotide weights. Least score among the experimentally known binding sites was considered as cut-off score. This software tool to screen microbial genomes is freely available.

The gene containing predicted DtxR binding sites in upstream sequence was considered as start gene of the operon. Factor independent transcription terminator was predicted using the GesteR software [4]. The co-directionally transcribed and functionally related genes with intergenic distance less than 100 base pairs were also considered as an operon.

3. Results and discussion

A recognition profile generated from eight known DtxR binding sites from *C. diphtheriae* was used to identify the potential DtxR binding sites in the upstream region of *C. glutamicum* genes. Table 1 lists 23 of these predicted DtxR binding sites in *C. glutamicum* genome. Orthologues of genes labeled with asterisk contain iron-sulfur cluster or known to be regulated by iron in other bacterial species. Whereas the orthologues of genes labeled with double asterisk were known to be regulated by DtxR in *C. diphtheriae*.

Genes in prokaryotes could be organized as an operon allowing more than one gene to be under the control of a

Proceedings of the Computational Systems Bioinformatics (CSB'03)
0-7695-2000-6/03 $17.00 © 2003 **IEEE**

common transcriptional regulatory system. Identification of genes based on *DtxR* binding site prediction could specify only the first gene of the operon. In order to know whether there are any *DtxR* regulated genes located further downstream, we carried out Rho-independent transcription termination sites search and identified a set of genes as a part of the *DtxR* regulated operons (data not shown). Some of the important genes/operons controlled by DtxR are described here.

Table 1. Predicted DtxR binding sites in *C.glutamicum*

Score	*Pos	Binding site	Gene
4.387	-59	GTCGGGCAGCCTAACCTAA	Cgl0649**
4.241	-116	TATGGCTTGCCTAACCTAA	Cgl1415
4.187	-110	TTAGTAAAGGCTCACCTAA	Cgl0493**
4.128	-269	TTAGGTGAGCCTTTACTAA	Cgl0494
4.099	-178	CACGGTGAACCTAACCTAA	Cgl2756*
4.098	-54	TGAGGTTAGCGTAACCTAC	Cgl0958**
4.087	-86	TTTAGGTAACCTAACCTCA	Cgl0787**
4.087	-25	AATGGTTAGGCTAACCTTA	Cgl0125
4.081	-30	TTAGGCTTGCCATACCTAT	Cgl0440*
4.059	-139	GTAGGTGTGGGTAACCTAA	Cgl2178**
4.057	-47	ATAGGATAGGTTAACCTGA	Cgl0627*
4.056	-174	AAAAGGTAGCCTTGCCTAA	Cgl1987
4.054	-133	TAAAGTAAGGCTATCCTAA	Cgl0366*
4.034	-163	TTAAGTTAGCATAGCCTTA	Cgl0384*
3.998	-132	ATAACGCACCCTAACCTTA	Cgl2948
3.998	-212	TTAACTTTGCCCTACCTAA	Cgl2804
3.987	-91	GCACGATGGCCAAACCTAA	Cgl0916
3.962	-54	TTAGGTTAAGCTAATCTAG	Cgl0388*
3.962	-65	CTACTGTGCCCTAACCTAA	Cgl1978
3.957	-80	TCAGGATAGGACAACCTAA	Cgl2943*
3.942	-400	TTAGGATAGCCTTACTTTA	Cgl0365*
3.937	-50	TAAGGATAACCTTGCCTTA	Cgl0335**
3.935	-93	TTAGGTTGTCCTATCCTGA	Cgl2944*
3.928	-196	TTAGGTAAAGCTTGCCTAT	Cgl1672
3.919	-460	TAAGGTTAGCCTAACCATT	Cgl0127*
3.888	-104	TTAAGTCAGTGTTACCTAA	Cgl0928*
3.872	-27	GCTCAATAACCTAACCTAA	Cgl2767
3.855	-186	TTGCATTAGGCTATCCTAA	Cgl3015
3.851	-59	TTATGCTGCGCTAACCTAT	Cgl2474*

The genes Cgl1414 and Cgl1413 are downstream to the gene Cgl1415 that code for Hemolysins containing Cystathionine Beta Synthase (CBS) domains. These genes were similar to the *tlyC* gene of other bacteria [5].

The gene Cgl0384 and Cgl0388 including the downstream gene Cgl0389 are similar to the Haemin transport associated proteins in *C. diphtheriae* and *Corynebacterium ulcerons (C. ulcerons)* [6]. The genes Cgl0385, Cgl0386, Cgl0387 are co-directionally transcribed with the gene Cgl0384 and similar to the *hmu*T, *hmu*U and *hmu*V genes respectively, of the hemin transport system in *C. diphtheriae* and *C. ulcerons*.

Our data show that DtxR could regulate the genes Cgl2474 and Cgl2944c whose products are orthologous to starvation inducible DNA binding protein (Dps) and Nei respectively. Dps in *Escherichia coli (E. coli)* oxidizes ferrous iron to ferric iron using hydrogen peroxide which in turn prevents hydroxyl radical formation by Fenton's reaction [7]. The protein Nei in *E. coli* is a DNA-glycosylase, which removes oxidative products of thymine (5-formyl uracil) and 5-methyl cytosine (5-hydroxymethyluracil) from DNA [8]. The product of the gene Cgl2474 is homologue of ferritin which is involved in iron storage in various bacteria.

In summery, we have observed binding sites of DtxR in upstream regions of the genes involved in iron uptake, iron storage, oxidative stress defense and DNA repair. Our findings highlight an important physiological role of DtxR in regulating the intracellular iron levels as well as in controlling the DNA damage due to Fenton's reactions.

5. References

[1] D..M. Oram, A. Avdalovic, and R.K. Holmes, "Construction and characterization of transposon insertion mutations in *Corynebacterium diphtheriae* that affect expression of the diphtheria toxin repressor (DtxR)". *Journal of Bacteriology*, American Society for Microbiology USA, 2002, pp. 5723-5732.

[2] Y. Qian, J.H. Lee, and R.K., Holmes, "Identification of a DtxR-regulated operon that is essential for siderophore-dependent iron uptake in *Corynebacterium diphtheriae", Journal of Bacteriology,* American Society for Microbiology USA, 2002, pp. 4846-4856.

[3] C. E. Shannon, "A mathematical theory of communication", *Bell System Technical Journal,* University of Illinois Press USA, 1948, pp. 379-423 and 623-656.

[4] S. Unniraman, R. Prakash, and V. Nagaraja, "Alternate paradigm for intrinsic transcription termination in eubacteria.", *Journal of Biological Chemistry,* American Society for Biochemistry and Molecular Biology USA, 2001, pp. 41850-41855.

[5] A.A. Ter Huurne, S. Muir, M. van Houten, M.B. Koopman, J.G. Kusters, B.A. van der Zeijst, and W. Gaastra, "The role of hemolysin(s) in the pathogenesis of Serpulina hyodysenteriae.", *Zentralblatt fur Bakteriologie,* Germany, 1993, pp. 316-325.

[6] M.P. Schmitt, and E.S. Drazek, "Construction and consequences of directed mutations affecting the hemin receptor in pathogenic *Corynebacterium* species.", *Journal of Bacteriology,* American Society for Microbiology USA, 2001, pp. 1476-1481.

[7]. G. Zhao, P. Ceci, A. Ilari, L. Giangiacomo, T.M. Laue, E. Chiancone, and N.D. Chasteen, "Iron and hydrogen peroxide detoxification properties of DNA-binding protein from starved cells. A ferritin-like DNA-binding protein of *Escherichia coli.*", *Journal of Biological Chemistry,* American Society for Biochemistry and Molecular Biology USA, 2002, pp. 27689-27696.

[8]. M. Hori, S. Yonei, H. Sugiyama, K. Kino, K. Yamamoto, and Q.M. Zhang, "Identification of high excision capacity for 5-hydroxymethyluracil mispaired with guanine in DNA of *Escherichia coli* MutM, Nei and Nth DNA glycosylases", *Nucleic Acids Research,* Oxford University Press U.K, 2003, pp. 1191-1196.

Appendix II

Curriculum Vitae

Sailu Yellaboina
Computational and Functional Genomics Group
Centre for DNA Fingerprinting and Diagnostics
HYDERABAD, INDIA
Email: sailu@cdfd.org.in

EDUCATION

Ph.D. (Thesis submitted to Department of Biochemistry, University of Hyderabad), **2005**
Centre for DNA Fingerprinting and Diagnostics, Hyderabad, INDIA
Title of thesis: ***In-silico* prediction of regulons in bacterial genomes**
Thesis supervisor: **Prof Seyed E. Hasnain**, Director, CDFD, Hyderabad

Master of Science (M.Sc), 1998
Department of Biochemistry,
University of Hyderabad, Hyderabad, INDIA

Bachelor of Science (B.Sc), 1996
Kakatiya University, Warangal, INDIA
Subjects: Botany, Zoology and Chemistry

Higher Secondary school (10+2), 1993
Andhrapradesh Secondary School of Education (APSSE), India
Subjects: Physics, Chemistry, Botany, Zoology and English

AWARDS/ HONOURS

- Qualified CSIR (Council of Scientific and Industrial Research) exam, Life Sciences/ Recipient of a 5-year fellowship from CSIR for pursuing a career in research. (1999-2004)
- Qualified ICAR (Indian Council for Scientific and Agricultural Research) - National Eligibility Test for Lectureship in Animal Biochemistry, December 1998.
- Qualified Graduate Aptitude Test in Engineering, 1998.

RESEARCH EXPERIENCE:

Project 1: Prediction of operons and *cis*-regulatory elements in bacterial genomes
[Part of PhD dissertation]

Project 2: Prediction of DtxR regulon in *Corynebacterium diphtheriae*
[Part of PhD dissertation7]

Project 3: Comparative analysis of IdeR regulon in *Mycobacteria*
[Part of PhD dissertation]

Project 4: Prediction of operon links in bacterial genomes

Project 5: Prediction of protein-proteins interactions in *Plasmodium falciparum* genome

Project 6: Distribution of amino acids along proteins sequences
[Project assistant]

Project 7: Activity of topoisomerase-II in Zn^{+2}:deficient rats
[M.Sc final year project, in partial fulfillment of the degree of Master of Science]

Project 8: Study of HIV-GP120 interaction with sulfatide and CD4 using Enzyme Linked imunosorbent assay
[M.Sc Summer project]

Project 9: Immobilization of enzymes on Hydroxy matrice
[M.Sc Summer project]

COMPUTER SKILLS

- **Operating system**: Windows, Macintosh 8.9, 9 as well as OS X (10.2.6 Jaguar), Linux flavors including Mandrake 10.0, RedHat 9.1, Slackware 10.0 and Unix flavors including Irix 6.5 and Sun Solaris.
- **Scripting** - Perl, Python, BASH Shell scripting
- **Programming Languages** - C, C++
- **Markup** - HTML, LaTeX2e, XML
- **Database** - MySQL, mSQL, PostgreSQL
- **Web Programming** - CGI Programming with Perl

MOLECULAR BIOLOGY AND BIOCHEMICAL TECHNIQUES

- Protein purification, estimation, sequencing.
- Protein activity assays and kinetic studies. Immobilization of enzymes, clinical and biochemical assays.
- Study of DNA protein Interactions using Gel mobility shift assays
- Production of antibodies, Antibody-antigen interaction studies and Enzyme Linked Immuno Sorbent Assay.
- Isolation of Plasmid DNA by Benson and Yang method and Adsorption chromatography. Analysis of DNA by electrophoresis, Renaturation kinectics and Colony hybridization.
- Attended 3 day workshop on Microrrays organized by Centre for DNA Fingerprinting and Diagnostics

PUBLICATIONS

1. 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*** [In Press]

2. **Yellaboina S**., Ranjan S., Chakhaiyar Prachee, 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 Microbiology, 4:38***

3. **Yellaboina S**., Seshadri J., Senthil Kumar M and Ranjan A. (2004) Predictregulon: A webserver for the prediction of the regulatory protein binding sites and operons in prokaryote genomes. **Nucleic Acids Research 32, W318-320**

CONFERENCE PROCEEDINGS AND POSTERS

1. **Yellaboina S**., Chakhaiyar P., Hasnain SE., and Ranjan A. (2003) Computational prediction of DtxR regulon-A Dissection of physiological process controlled by DtxR in **Corynebacterium species. Proceedings of IEEE Computational System Bioinformatics Conference 442-443**

2. Chakhaiyar P, **Yellaboina S**, A. Ranjan, Seyed E. Hasnain. 2002. Identification and partial characterization of novel genes of *Mycobacterium tuberculosis* regulated by iron. Presented at the **All India Cell Biology Conference**, December 12-14, 2002. Advanced Centre for Treatment, Research & Education in Cancer, Mumbai, India.

3. **Yellaboina S**, C.K.Mitra and Anusharka Sen. (1999) Distribution of Amino acids along protein sequences. International Conference on Life Sciences in Next Millenium, December 11-14,1999. University of Hyderabad, Hyderabad. India.

TEACHING EXPERIENCE

Taught PhD students, Aug., 2004, Centre for DNA Fingerprinting & Diagnostics (CDFD).

Topics covered were:

- Genome organization
- Gene prediction
- Function prediction by homology
- Computational prediction of protein-DNA and protein-protein interactions.
- Analysis of DNA-protein and protein-protein interaction networks
- Network evolution