Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

Computational identification of composite regulatory sites in 16s-rRNA gene promoters of *Mycobacterium species*

Neeraja Dwivedi[1], Surya Prakash Dwivedi[1], Ajay kumar[1], Vishwa Mohan Katoch[2] and Sanjay Mishra[1,*]

[1]Department of Biotechnology, College of Engineering and Technology, IFTM Campus, Lodhipur-Rajput, Delhi Road, Moradabad 244 001, U.P., India; [2]National Institute for Leprosy (ICMR) and Other Mycobacterial Diseases, Tajganj, Agra, UP, India; Sanjay Mishra* - E-mail: sanjaymishra66@gmail.com; Phone: 91 591 6453829; * Corresponding author

received March 25, 2008; revised May 13, 2008; accepted May 15, 2008; published June 16, 2008

Abstract:
The availability of completely sequenced genomes allow the use of computational techniques to investigate *cis*-acting sequences controlling transcription regulation associated with groups of functionally related genes. Theoretical analysis was performed to assign functions to regulatory systems. The identification of such sites is relevant for locating a promoter at the 5' boundary of a gene. They also allow the prediction of specific gene-expression pattern and response to disturbances in a known signaling pathway. Here, we describe the identification of composite transcription factor (TF) binding sites over promoter regions in16s-rRNA gene for mycobacterium species strains ICC47, ICC67, ICC43 and CMVL700. It is established that the ribosomal gene comprises of sequences that are conserved during evolution and interspersed with divergent regions. Computational identification of known TF-binding sites was performed using TFSITESCAN tool and ooTFD database. The ICC67, ICC47, ICC43 and CMYL700 strains showed 12, 13, 9 and 15 known TF binding sites, respectively. Comparison between strains suggests 9 known TF predicted binding sites to be conserved among them. These data provide basis for the understanding of promoter regulation in 16s-rRNA.

Keywords: composite TF binding sites; promoters; 16s rRNA gene; *Mycobacteria;* regulatory site

Background:
The availability of a number of fully sequenced genomes allow the use of computational techniques to investigate cis-acting elements controlling transcription regulation associated with groups of functionally related genes. The gene expression pattern is integral to the structure of the transcription regulatory regions by specific combinations of TF binding sites. Several computational approaches have been reported to identify regulatory elements during the last decade. Specific TF binding site combinations were identified for muscle-specific promoters in liver-enriched and yeast genes **[1-4]**. Recently, it has been shown that search for specific combinations of two TF site specific composite elements is an effective tool in predicting gene expression patterns for immune-cell specific genes.

Lehman and Neumann established the generic name mycobacterium **[5]**. The first member of this genus to be identified was the *Leprae bacillus* by Hansen **[6]**. Mycobacteria are gram-positive and are usually of the type acid-fast bacilli (AFB). They are non-motile and do not form capsule endospore or conidia. Some species do not grow *in vitro*. They are classified as slow or fast growing bacteria based on growth rate. This genus includes obligate parasites, opportunistic pathogens and saprophytes. These are invariably aerobic with a slightly curved or straight rod measuring about 1-10 micro-meter in size and are occasionally seen as branched filaments. *Mycobacteria* have arabinose and glucose as their principle cell wall sugars.

It is established that the ribosomal gene rich region of both the prokaryotic and eukaryotic genomes comprise of sequences that are conserved during evolution interspersed among divergent regions. 16S rRNA based phylogenetic analyses have contributed to the systematic identification and classification of mycobacteria. 16S rRNA gene, either by direct sequencing **[7]** or by using probes **[8],** has now widely been used for rapid identification of mycobacteria. The clinical importance of several mycobacterial species has increased, especially since the human immunodeficiency virus (HIV) pandemic **[9]**.

The antibiotic susceptibility of some of these species, namely, *M. kansasii*, *M. marinum*, *M. simeae and M. asiaticum* are low **[10]**. The main disease caused by *M. kansasii* is benign pulmonary disease in elderly white males and *M. marinum* is a causative agent for swimming pool granuloma (skin ulcers) among swimmers. Therefore, it is important to understand the regulatory mechanisms in mycobacteria. Here, we describe the prevalence of TF sites in 16s rRNA using predictive tools.

Bioinformation by Biomedical Informatics Publishing Group open access
www.bioinformation.net Hypothesis

Methodology:

Strains

The mycobacterium species strains used in the study were ICC47, ICC67, ICC43 and CMVL700 (Table 1a in supplementary material).

Culturing and biochemical characterization

Strains included in the study were obtained from the Department of microbiology and molecular biology of Central JALMA Institute for leprosy, which is a national respiratory centre for mycobacteria. Strains ware taken and sub-cultured on Lowenstein Jensen media (LJ media). Biochemical and molecular characterization were performed and documented.

DNA isolation

DNA was isolated using standard 'Foaming' protocol **[5]**.

PCR

Amplification of isolated DNA was performed by PCR using the procedure described by Dobner and colleagues (1996) with minor modifications. The PCR amplification of promoter region of 16s rRNA gene was performed using the following set of primers: Primer-1 (Upstream primer) MYC-12 5'CGGAATTCCACATCGATCGCGG - 3' and Primer-2 (Downstream primer) MYC-1 5'-CGGGATCCTGAGCCAGGATCAA - 3'

Agarose gel electrophoresis

The presence and yield of specific PCR product ware analyzed at 3% agarose gel electrophoresis for 2 hrs in 75 volts.

DNA extraction from agarose gel

Elusion of DNA from agarose gel was done by 'Wizard' method.

Performing cyclo sequencing PCR

Extracted DNA was used in cyclic sequencing PCR for making single strand DNA. The parameters for cyclo sequencing PCR were done for 25 cycles with denaturation at 94 ^{0}C for 30 seconds, annealing at 55 ^{0}C for 10 seconds and extension at 60 ^{0}C for 4 minutes.

Purification

The cyclo sequencing PCR product was purified by adding 0.1 Volume of 3M Sodium acetate (pH-4.5) and 2.5 Volume of absolute alcohol. The solutions were mixed and tubes were left at room temperature for 15 minutes. The tubes were then centrifuged at 10000 g for 10 minutes. The supernatant was discarded and the pellet was washed with 70% ethanol (100µl). Subsequently, tubes were air-dried after wash.

Sequencing region of 16s rRNA promoter

The sample was transferred into fresh tube and closed with septa for loading onto sample tray. The sample runs through performance optimized polymer (POP) and electrophorized at 12.1 kV for 3 hours in 1x genetic analyzer buffer. The sequencing of amplicon was carried out using the ABI prism automated DNA sequencer (ABI prism genetic analyzer 310 USA). The sequences generated by the program were compared to their respective wild type sequences using the DNA Star software.

Computational analysis

Recognition of TF composite units

TF composite unit consists of a binding site for a known TF arranged with various flanking motifs and potential targets for additional transcription factors. Such TF composite units could serve as targets for complexes of different transcription factors synergistically regulating gene transcription. The method is designed by discriminating single set of promoter sequences. TF database consist of known TF binding sites and weight matrices were developed for target sites. These parameters were used as training set in prediction and recognition.

Training dataset of 16s rRNA promoters

The training dataset of sequenced 16s rRNA promoters from mycobacterial species were arranged in FASTA format for further analysis.

TF-binding site prediction by TFSITESCAN

TFSITESCAN tool **[11]** is maintained by the Institute for Transcriptional Informatics **[12]**. TFSITESCAN identifies potential transcription factor binding sites in a promoter sequence. The putative binding sites were derived from an object oriented transcription factors database named ooTFD described elsewhere **[13]**.

Discussion:

In the non coding upstream sequences of *Mycobacterium fortuitum* (strain ICC-67), the 16s rRNA promoter contains a total of 12 known binding sites analogue with known transcription factors consisting of two unknown sites. It is seen that 8 binding sites with known TF is common in 16s rRNA genes of *M. fortuitum.* The results were summarized in Table 1b (see supplementary material) and the common sites were listed in Table 1c (see supplementary material). The 16s rRNA promoter in ICC-47 of *Mycobacterium* contains a total of 13 TF binding sites with12 known and 1 unknown TF. The promoter in ICC-43 contains a total of 9 binding sites along with known consecutive TF. One binding site was predicted with no information for transcription factor. The CMYL700 strain 16s rRNA gene of *M. tuberculosis* contains a total of 15 known binding sites with 13 sites of known transcription factors. A comparative study of TF among them suggests conversion between them. It should be noted these predicted data should be clearly confirmed by designing appropriate experiments.

Multiple TF elements have been shown to interact with the upstream region of 16s-rRNA promoter in *Mycobacterium species*. We found 8 TF known binding sites common in the dataset of mycobacterial 16s-rRNA promoter sequences used in this analysis. One known binding site was found

common without transcription factor data. TF binding sites FOX family_CS, NF-Y-consensus and ETS2_H2 had minimum occurrences over promoter sequences and sites GCF_CS, GC_box, AP-2_CS6, KKLF_CS and NF-E1_CSI were found with high number of occurrence.

These factors recognize each of the three CCAAT motifs present in the EIIL promoter at positions -72, -135 and -229. They also identify CCAAT elements in rat albumin and herpes virus thymidine kinase promoters. A mutation is known to reduce thymidine kinase promoter activity *in vivo* and *in vitro*. This abolishes binding of the factor termed CCAAT recognition factor (CRF) and it is distinct from previously identified CCAAT factors. In addition, the upstream factor II (USFII) shares binding sites at position -110 with EIIL promoter and c-fos enhancer adjacent to the serum regulating element. The recognition site for USFII is also found in c-fos promoter, adenovirus early region EIV and EIIa early promoters. A Sp1 recognition site has been identified at position -41, and the binding sites for Sp1, USFII and CRF are required for efficient EIIa-late promoter function. Finally, an additional factor recognizing the consensus element GGGGGGNT has been detected (see Table 1b and Table 1c in supplementary material).

Conclusion:

We described the regulatory elements in four promoters of 16s rRNA gene with different known TF in *Mycobacterium* species. Binding sites with known and unknown transcription factors reveal composite gene regulation over 16s rRNA gene in different species of *Mycobacteria* due to the presence of multiple binding sites and transcription factor data. A total of 9 known TF binding sites were predicted common in 4 promoters studied in 16s RNA gene. The details of each TF binding sites were summarized in Table 1b (supplementary material) with consensus patterns of occurrences. Higher number of occurrence strongly supports the presence of binding sites which might have a role in gene regulation. Details of each TF binding sites are summarized in Table 2 (see supplementary material). It should be stated that these data provide a skeleton to understand the basis of transcription regulation in Mycobacteria. Nonetheless these data require confirmation by appropriate experimental data.

Acknowledgement:

The present work was supported by a joint venture of the laboratory facility at National Institute for Leprosy (ICMR) and other Mycobacterial Diseases, Agra, U.P., India and CET, IFTM, Moradabad, U.P., India. An institutional research promotion grant to the Department of Biotechnology, College of Engineering & Technology, Moradabad, U.P., India is also acknowledged. The authors are grateful to Prof. R. M. Dubey (Managing Director, CET, IFTM, Moradabad, U.P., India) for providing the necessary facilities and encouragement. The authors are also thankful to all faculty members of the Department of Biotechnology, College of Engineering & Technology, Moradabad, U.P., India, for their generous help and suggestions during the course of experimental work and manuscript preparation.

References:

[01] W. W. Wasserman, *et al*, *J. Mol. Biol.*, 278: 167 (1998) [PMID: 9571041]

[02] K. Frech, *et al., In Silico Biology*, 1: 0005 (1998) [PMID: 11471240]

[03] F. Tronche, *et al., J. Mol. Biol.*, 266: 231 (1997) [PMID: 9047360]

[04] A. Brazma, *et al., Proc. of the German Conference on Bioinformatics GCB'97, Germany,* (H. W. Mewes and D. Frishman eds.), 57 (1997)

[05] P. Dobner, *et al., J Clin Microbiol.*, 34: 866 (1996) [PMID: 8815098]

[06] K. E. Kembsell, *et al*, *J Clin Microbiol*, 138: 1717 (1992) [PMID: 1382114]

[07] D. Ghosh, *Nucleic Acids Research*, 27: 315 (1999) [PMID: 9847215]

[08] H. A Cox and V. M Katoch, *FEBS letters*, 195: 194 (1986) [PMID: 3002853]

[09] J. E. Clark-Curtiss, *Mcfadden J Surrey University Academic Press,* 77 (1990)

[10] E. A. Aghanzani, *et al., J Clin Microbiol.*, 134: 98 (1986) [PMID: 8748282]

[11] http://www.ifti.org/cgi-bin/ifti/Tfsitescan.pl

[12] http://www.ifti.org/

[13] http://www.ifti.org/

Edited by P. Kangueane

Citation: Dwivedi *et al.*, Bioinformation 2(8): 363-369 (2008)

Supplementary material

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

S. No.	Strain number	Organism name
1	ICC 47	*Mycobacterium fortuitum*
2	ICC 67	*Mycobacterium fortuitum*
3	ICC 43	*Mycobacterium fortuitum*
4	CMVL 700	*Mycobacterium tuberculosis*

Table 1a: Strain number and organism names used in this study is given.

S. No	Promoter sequence	Site name	Length	Binding sites /TF	Position	Score	Occurrence	EXP value
1	*M. fortuitum* ICC-67	hFASAT-1-ARE	8	ARET-11	15	7	2	5.46E-02
		FOX Family CS	7	FOX proteins	15	6	1	5.94E-01
		ETS2 H2 site	11	H2a/H2b	69	8	6	7.46E-03
		NF-Y-Consensus	13	NF-Y	95	10	2	3.41E+00
		GCF CS	7	GCF	117	7	2	5.94E-01
		GC box	7	ETF	119	7	1	3.62E-01
		NF-E1 CSI	8	GATA-1	125	8	1	3.62E-01
		C/EBP-CPSI	11	C/EBP	249	8	2	7.40E-03
		AP-2 CS6	8	AP-2	396	6	2	4.46E-01
		TFIID-ADA	8	TFIID	298	6	3	1.40E-02
		KKLF CS	9	KKLF	367	7	4	5.45E-02
		GAr C1	9	Unknown	394	7	1	3.49E-03
		W-element CS	8	Unknown	251	7	4	5.93E-01
2	*M. fortiutum* ICC-47	KKLF CS	9	KKLF	450	7	1	5.49E-02
		FOX Family CS	7	FOX proteins	44	6	1	6.41E-01
		ETS2 H2 site	11	H2 a/H2b	67	8	2	8.41E-03
		SW15 Consensus	6	SW15	70	6	2	6.42E-01
		SCB consensus	7	SCB	75	6	2	2.26E-01
		NF-Y-Consensus	13	NF-Y	92	12	2	9.64E-02
		GC box	10	ETF	110	8	1	6.16E-02
		NF-E1 CSI	8	GATA-1	124	8	2	4.00E-02
		GCF CS	7	GCF	138	6	2	6.41E-01
		Sp1-tk1	10	Sp1	287	10	1	8.81E-04
		AP-2 CS6	8	AP-2	288	7	2	5.96E-01
		TFIID-ADA	8	TFIID	300	8	1	1.41E-02
		W-element CS	8	Unknown	39	7	2	6.40E-01
3	*M. fortuitum* ICC-43	FOX Family CS	7	FOX proteins	16	6	3	1.42E-01
		GCF CS	7	GCF	58	6	3	5.97E-01
		ETS2 H2 site	11	H2 a/H2b	68	8	1	9.46E-03
		NF-Y-Consensus	13	NF-Y	94	10	2	3.44E+00
		NF-E1 CSI	8	GATA-1	124	8	1	3.64E-01
		TFIID-MBP	8	TFIID	131	6	1	2.03E-01
		GC box	7	ETF	148	7	4	3.65E-01
		AP-2 CS6	8	AP-2	188	8	2	6.41E-02
		KKLF CS	9	KKLF	450	7	1	5.49E-02
		W-element CS	8	Unknown	39	7	2	6.40E-01
		BRE-G2	11	Unknown	176	8	1	7.46E-03
4	*M.*	FOX family CS	7	FOX protein	16	7	1	5.73E-01

	tuberculosis CMVL-700	GCF CS	7	GCF	63	7	3	3.46E-01
		ETG2 H2 site	11	H2a/H2b	73	8	1	5.73E-01
		BRE	7	BRE	57	6	1	6.97E-03
		NF-E1 CSI	8	GATA-1	129	8	1	3.54E-01
		FN-GCE-A	9	Egr-1	153	9	1	3.29E-03
		GC box	7	ETF	436	6	4	2.64E-02
		C/EBP-CPSI	11	C/EBP	252	8	1	6.97E-03
		GaEII-late	8	GaEII	357	6	3	1.32E-02
		NF-Y-consensus	13	NF-Y	374	10	1	3.22E+00
		AP-2 CS6	8	AP-2	426	7	5	5.72E-01
		KKLF CS	9	KKLF	426	7	5	5.15E-02
		JCV repeat seq.	8	JCV	427	7	6	1.91E-01
		BRE-G2	11	unknown	181	8	2	6.97E-03
		W-element CS	8	unknown	254	7	3	5.72E-01

Table 1b: Predicted transcription factors with their binding sites data in 16s-RNA promoter sequences of mycobacteria.

S. No	Common binding site with length	Reference	Consensus sequence	Number of TF binding site occurrences in *Mycobacterium* species strain			
				ICC 67	ICC 47	ICC 43	CMVL 700
1	FOX_family_CS (7)	B. Wang et al., *J Biol Chem.* **278:** 24259 (2003)	TRTTKRY	1	1	2	1
2	ETS2_H2_ site (11)	G. J. Mavrothalassitis & T. S. Papas *Cell Growth Differ.* **2:** 215 (1991)	GAGACTGACGA	2	1	1	1
3	GCF_CS (7)	S. Faisst & S. Meyer, *Nucleic Acids Res.* **20:** 3 (1992)	SCGSSSC	2	2	3	2
4	GC_box (7)	R. Kageyama et al., *J Biol Chem.* **264:** 15508 (1989)	CCCSCSS	1	2	1	4
5	AP-2_CS6 (8)	S. Faisst & S. Meyer, *Nucleic Acids Res.* **20:** 3 (1992)	CCCMNSSS	3	2	2	5
6	KKLF_CS (9)	S. Uchida et al., *Mol Cell Biol.* **20:** 7319 (2000)	GGGGNGGNG	4	1	1	5
7	NF-Y-consensus (13)	R. Mantovani *Nucleic Acids Res.* 1998 **26:** 1135.	BVDCCAATWWD	1	1	2	2
8	NF-E_ CSI (8)	L. Wall et al., *Genes* Dev. 1988 **2:** 1089.	MYWATCWY	1	2	2	10
9	W-element CS*	Z. Yang et al., *Proc Natl Acad Sci U S A.* 1990 **87:**9226.	WGNAMCYK	1	2	1	3

Table 1c: Binding site and consensus sequences. Note: *** =** known binding site without TF data at ooTFD database.

S.No.	Common TF binding sites	consensus sequence	Top of form Function Bottom of form
1	FOX family CS	TRTTKRY	Multiple domains define the expression and regulatory properties of Foxp1 forkhead transcriptional repressors. Foxp1 proteins have diverse functional roles in different cell and tissue types
2	ETS2 H2 site	GAGACTGACGA	Positive and negative factors regulate the transcription of the ETS2 gene via an oncogene-responsive-like unit within the ETS2 promoter region.
3	NF-E1 CSI	MYWATCWY	The human beta-globin gene 3' enhancer contains multiple binding sites for an erythroid-specific protein.
4	GCF CS	SCGSSSC	Vertebrate-encoded transcription factors.
5	GC box	CCCSCSS	Transcription factor ETF stimulates the expression of the epidermal growth factor receptor (EGFR) gene which does not have a TATA box in the promoter region.
6	AP-2 CS6	CCCMNSSS	Vertebrate-encoded transcription factors.
7	KKLF CS	GGGGNGGNG	KKLF was found to be abundantly expressed in the liver, kidneys, heart, and skeletal muscle, and immuno-histochemistry revealed the nuclear localization of KKLF protein in interstitial cells in heart and skeletal muscle, stellate cells, and fibroblasts in the liver. Transcriptional regulation of the CLC-K1 promoter by myc-associated zinc finger protein and kidney-enriched like factor, a novel zinc finger repressor.
8	NF-Y-consensus	BVDCCAATWWD	The CCAAT box is one of the most common elements in eukaryotic promoters, found in the forward or reverse orientation. Among the various DNA binding proteins that interact with this sequence, only NF-Y (CBF, HAP2/3/4/5) has been shown to absolutely require all 5 nt.
9	W-element CS	WGNAMCYK	The interferon gamma (IFN-gamma) response region of the human class II major histo-compatibility complex gene, DPA, has been localized to a 52-base-pair (bp) DNA fragment in the proximal promoter at -107 to -55 bp after transfection into HeLa cells of a series of 5', 3', and gap deletion mutants linked to a reporter gene, human growth hormone, as well as of synthetic oligonucleotides fused to the heterologous promoter thymidine kinase.

Table 2: Details of putative TF binding sites with consensus sequence and function found common in promoter training data set sequences of Mycobacterial species 16s-RNA.