BMC Genomics

BioMed Central

Research article

Open Access

GntR family of regulators in *Mycobacterium smegmatis*: a sequence and structure based characterization

Vaibhav Vindal, Katta Suma and Akash Ranjan*

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

Email: Vaibhav Vindal - vaibhav@cdfd.org.in; Katta Suma - suma@cdfd.org.in; Akash Ranjan* - akash@cdfd.org.in

* Corresponding author

Published: 23 August 2007

BMC Genomics 2007, **8**:289 doi:10.1186/1471-2164-8-289

Received: 21 November 2006

Accepted: 23 August 2007

This article is available from: http://www.biomedcentral.com/1471-2164/8/289

Abstract

Background: *Mycobacterium smegmatis* is fast growing non-pathogenic mycobacteria. This organism has been widely used as a model organism to study the biology of other virulent and extremely slow growing species like *Mycobacterium tuberculosis*. Based on the homology of the N-terminal DNA binding domain, the recently sequenced genome of *M. smegmatis* has been shown to possess several putative GntR regulators. A striking characteristic feature of this family of regulators is that they possess a conserved N-terminal DNA binding domain and a diverse C-terminal domain involved in the effector binding and/or oligomerization. Since the physiological role of these regulators is critically dependent upon effector binding and operator sites, we have analysed and classified these regulators into their specific subfamilies and identified their potential binding sites.

Results: The sequence analysis of *M. smegmatis* putative GntRs has revealed that FadR, HutC, MocR and the YtrA-like regulators are encoded by 45, 8, 8 and 1 genes respectively. Further out of 45 FadR-like regulators, 19 were classified into the FadR group and 26 into the VanR group. All these proteins showed similar secondary structural elements specific to their respective subfamilies except MSMEG_3959, which showed additional secondary structural elements. Using the reciprocal BLAST searches, we further identified the orthologs of these regulators in *Bacillus subtilis* and other mycobacteria. Since the expression of many regulators is auto-regulatory, we have identified potential operator sites for a number of these GntR regulators by analyzing the upstream sequences.

Conclusion: This study helps in extending the annotation of *M. smegmatis* GntR proteins. It identifies the GntR regulators of *M. smegmatis* that could serve as a model for studying orthologous regulators from virulent as well as other saprophytic mycobacteria. This study also sheds some light on the nucleotide preferences in the target-motifs of GntRs thus providing important leads for initiating the experimental characterization of these proteins, construction of the gene regulatory network for these regulators and an understanding of the influence of these proteins on the physiology of the mycobacteria.

Background

Being a fast growing, non-pathogenic mycobacteria, *Mycobacterium smegmatis* has been widely used as a model organism to study the biology of other virulent and extremely slow growing species like *M. tuberculosis*. The genome of *M. smegmatis*, as listed at the TIGR site, contains a large number of putative GntR-like regulators. These regulators play an important role in the cellular physiology. Many such regulators are involved in regulation of gene expression in response to various oxidized substrates related to either amino acid metabolism or at the branch points of various other metabolic pathways.

The GntR family of bacterial regulators is named after the *Bacillus subtilis* transcription regulator- GntR- a repressor of the gluconate operon [1]. Regulators of this family possess a conserved N-terminal domain that is involved in the DNA binding. Based on this conservation, these proteins can easily be recognized by a Conserved Domain Database (CDD) search [2]. However, the C-terminal domain, which is involved in the effector binding and/or oligomerization (E-b/O), is quite diverse and heterogeneous. As a consequence of this heterogeneity, the GntR regulators have been further classified into six subfamilies (FadR, HutC, MocR, YtrA, AraR and PlmA) [3,4]. The members of subfamilies possess conserved secondary structural features specific to their subfamily and interact with a limited number of molecules [5]. Considering these conserved secondary structural features in sequence analysis, GntR regulators are defined as a part of specific subfamily [6]. Earlier, we have characterized GntR regulators from *M. tuberculosis* [7]. In present study putative GntR regulators from *M. smegmatis* are classified into their specific subfamilies. Further, suitable orthologs of the *M. smegmatis* GntRs were also identified using reciprocal BLAST searches in *M. tuberculosis, M. avium, M. bovis, M. ulcerans, M. sp.* KMS, *M. sp.* MCS, *M. vanbaalenii* PYR-1 and *B. subtilis*. To identify the DNA targets of these regulators, we utilized the information about the nucleotide preferences for regulators of a given subfamily. All the upstream DNA sequences of the GntR coding genes were scanned to locate potential palindromes that matched the nucleotide preference criteria [5].

Results and discussion

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

Unrooted tree of the *M. smegmatis* GntRs was constructed with the classified representatives of all subfamilies (Table 1) [5]. Among all putative *M. smegmatis* GntRs two proteins (MSMEG_1043 and MSMEG_2323) were found to be identical in sequence, hence only one of them MSMEG_1043 was taken for the classification. Each branch of the constructed tree represents a subfamily. Bootstrapping, involving 1000 replicates, shows all subfamily branches clustered with high bootstrap values. FadR subfamily is divided into two groups, FadR and VanR (Figure 1).

FadR-like proteins of M. smegmatis

Of all the putative GntRs, 45 proteins were classified as the FadR-like regulators. These subfamily members are further classified into two groups FadR and VanR where the C-terminal effector binding and/or oligomerization domain length is about 170 and 150 amino acid residues respectively comprising all α-helices [5]. Among all FadR-like regulators, 19 regulators were clustered as members of the FadR group while 26 for the VanR group (Table 2). To study secondary structural features both the group members were dealt with separately. C-terminal domain of all the members of FadR group were predicted with seven α-helices except MSMEG_2599. All the regulators showed distinguishable predicted secondary structural features specific to this subfamily (Figure 2 and Figure 3) [5]. Secondary structural patterns of the regulator MSMEG_3959 revealed an extra secondary structural element, which could be significant in studying protein family evolution. FadR-like regulators are known to be involved in the regulation of gene expression in response to oxidized substrates related to either amino acid metabolism or at the branch point in various metabolic pathways such as glycolate [8], pyruvate [9], lactate [10], malonate [11] or gluconate [12]. One of FadR-like classified transcriptional regulator MSMEG_6700 is known to be involved in the regulation of piperidine and pyrrolidine metabolism [13]. These results provide a starting point for a detailed biochemical and genetic characterization of *M. smegmatis* FadR-like regulators.

HutC-like proteins of M. smegmatis

Contrary to the FadR-like regulators, the regulators of this subfamily consist of both α-helices and β-sheet structures in the C-terminal domain. We identified eight GntRs as members of this subfamily (Table 2). All these members showed distinguishable predicted secondary structural features specific to this subfamily (Figure 4) [5]. These regulators are known to acquire the same protein fold as *Escherichia coli* UbiC; hence it is also named as UbiC transcription regulator-associated (UTRA) domain [14]. This effector-binding domain responds to various ligands like histidine (HutC) [15], long chain fatty acids [16], trehalose 6-phosphate [17] or alkylphosphonate [18]. A range of known ligands, specific to many HutC-like regulators, will help in characterizing the classified *M. smegmatis* regulators.

MocR-like protein of M. smegmatis

Among all the putative GntR regulators, eight were classified as members of the MocR subfamily (Table 2). All the eight regulators showed distinguishable predicted second-

Table 1: Details of GntR regulators used as representative from all subfamilies

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

Table 2: List of Classified *M. smegmatis* GntR regulators

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

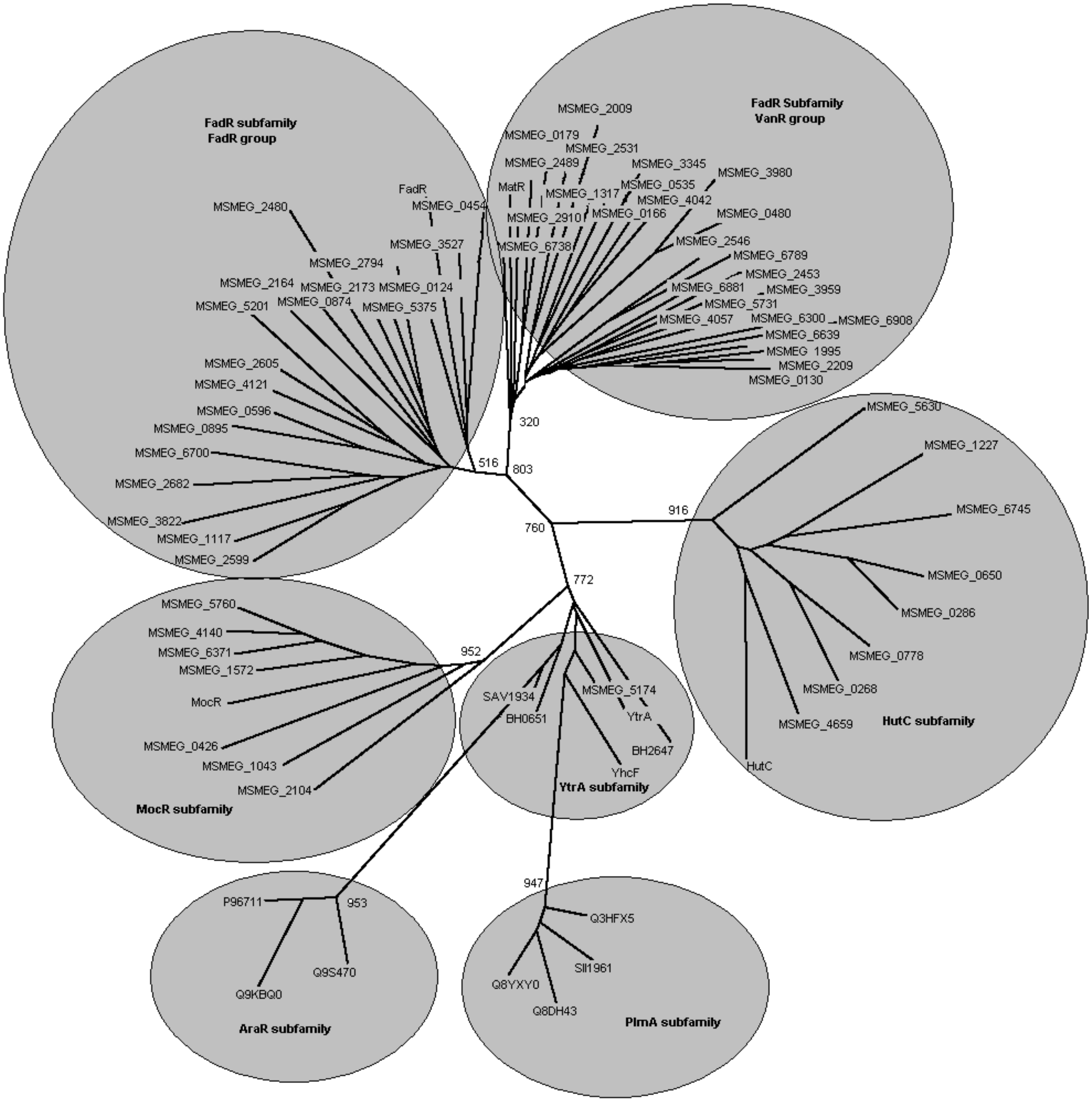
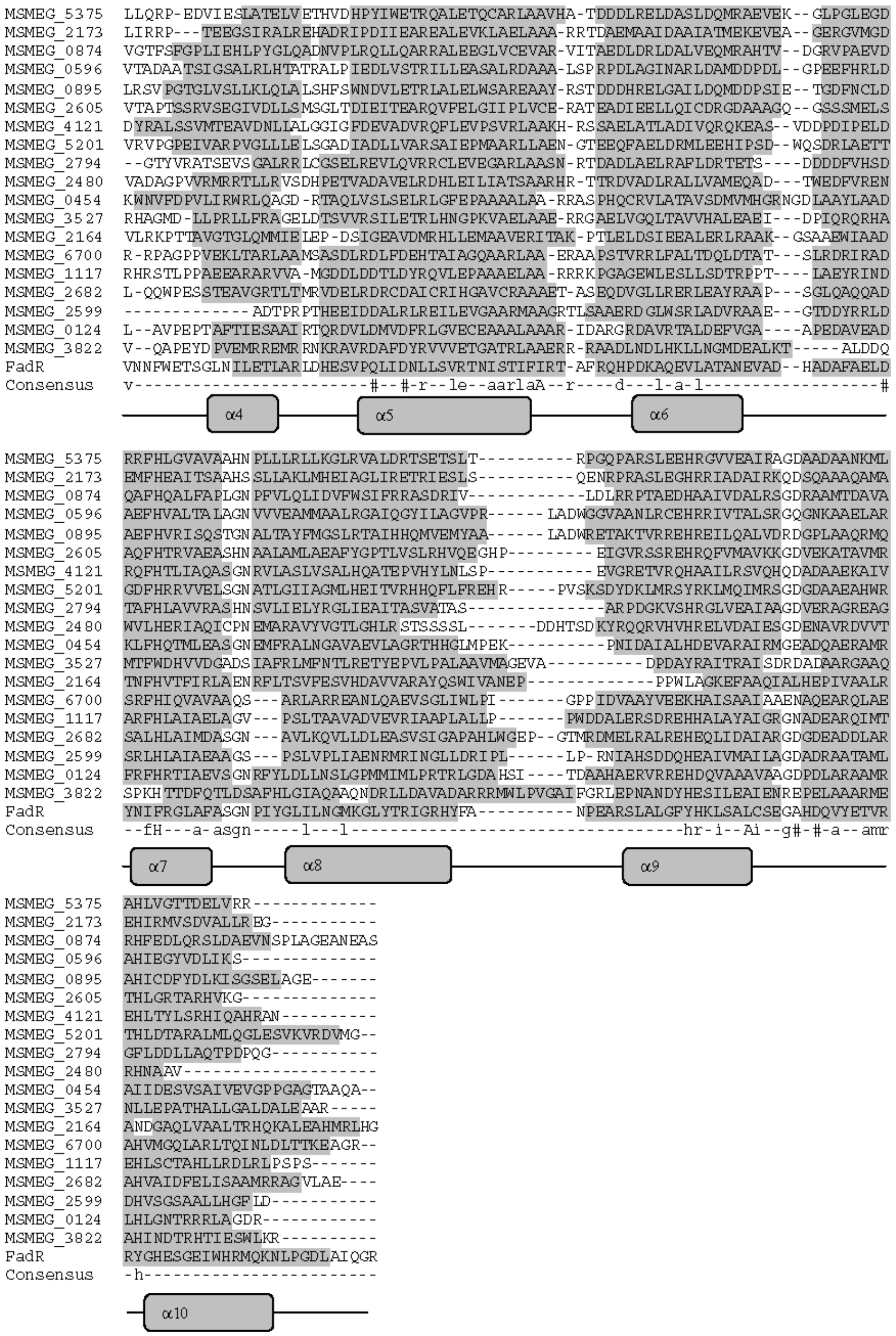


Figure 1

Unrooted tree of the proteins of GntR family regulators of *M. smegmatis* including representatives of all subfamily regulators from different Bacterial Genomes with 1000 bootstrap replicates. All the GntR regulators are clustered into six subfamilies. FadR subfamily is branched again into two groups (FadR and VanR). (Abbreviations are as indicated in Table 1 and Table 2).



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

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

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

Figure 2

Structure based sequence analysis of *M. smegmatis* GntR-like regulators by the multiple sequence alignment of the C-terminal domains of GntR regulators belonging to FadR Subfamily (FadR group). Abbreviations are as indicated in Table 1. Consensus sequence from the multiple sequence alignment has been drawn. High and low consensus levels were fixed arbitrarily at 80% and 40% of identity and are represented respectively by the capital and lowercase letters. Consensus symbol ! used for anyone of IV; $ is anyone of LM; % is anyone of FY; # is anyone of NDQEBZ. In graphical representation α-helix region and β-sheet regions are highlighted with light and dark gray background.

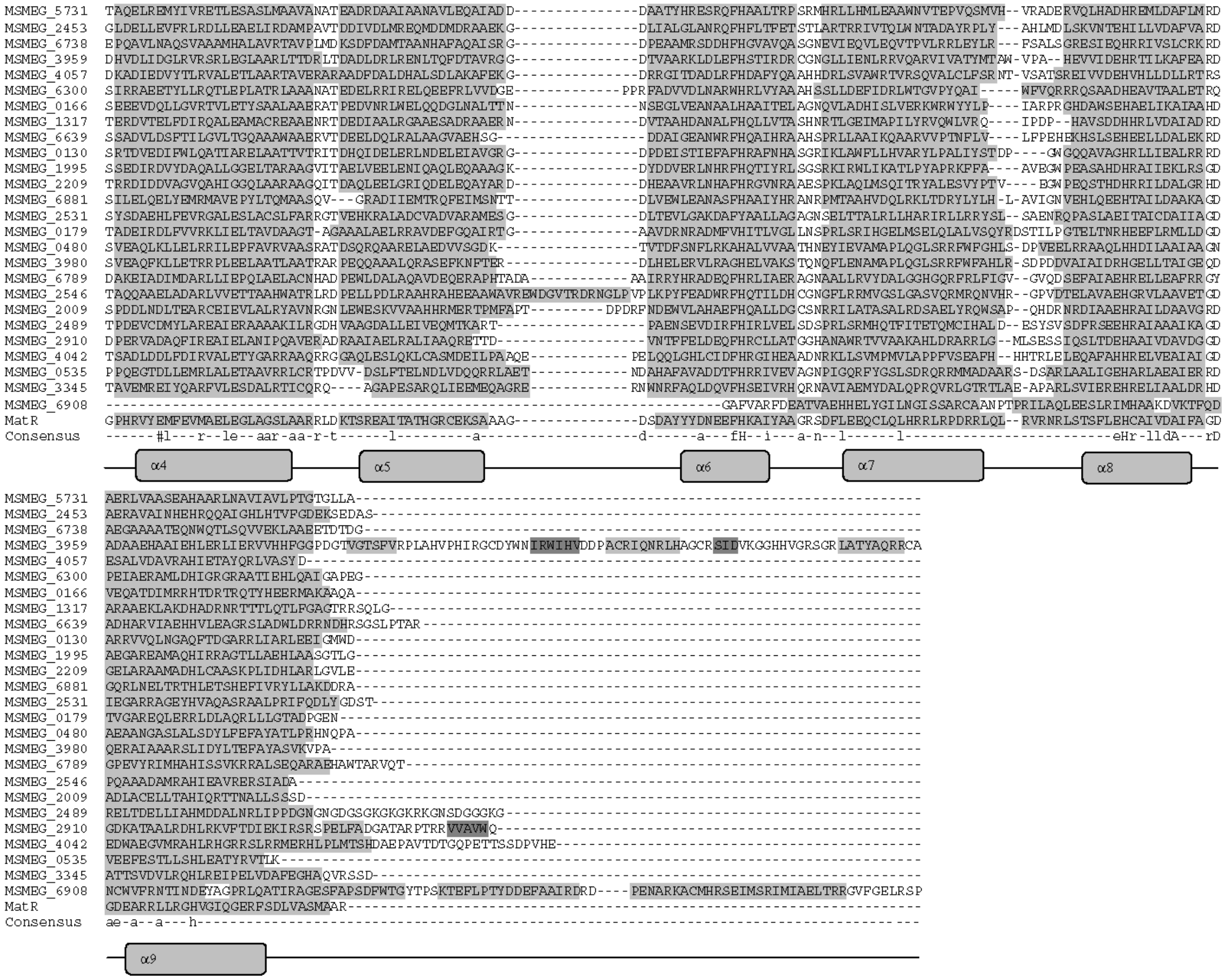
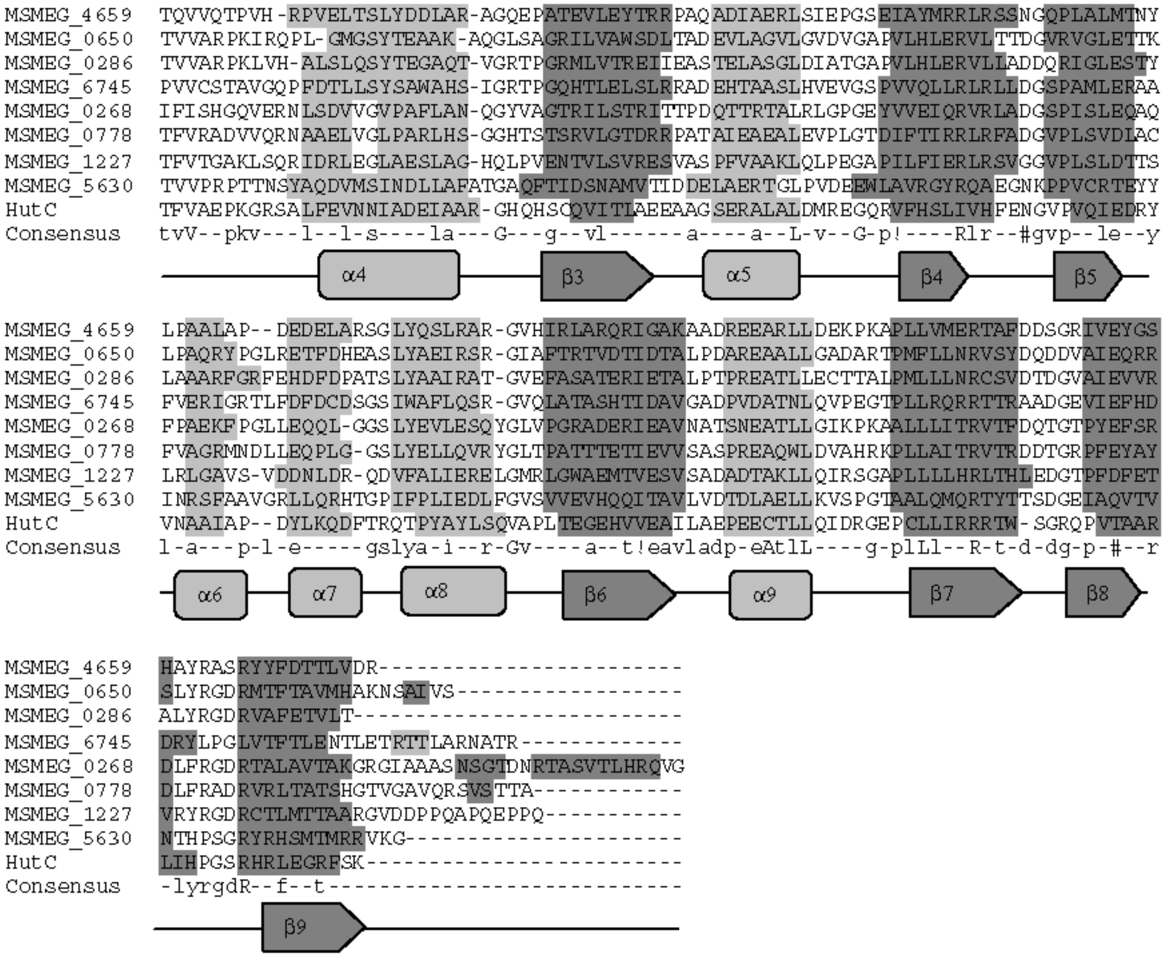


Figure 3
Structure based sequence analysis of *M. smegmatis* GntR-like regulators by the multiple sequence alignment of C-terminal domains of GntR regulators belonging to FadR Subfamily (VanR group). Abbreviations are as indicated in Table 1. Consensus sequence from the multiple sequence alignment has been drawn. High and low consensus levels were fixed arbitrarily at 80% and 40% of identity and are represented respectively by the capital and lowercase letters. Consensus symbol ! used for anyone of IV; $ is anyone of LM; % is anyone of FY; # is anyone of NDQEBZ. In graphical representation α-helix region and β-sheet regions are highlighted with light and dark gray background.

ary structural features specific to this subfamily (Figure 5) [5]. MocR-like regulators show homology to the class I aminotransferase proteins [19], which requires pyridoxal 5'-phosphate (PLP) as a co-factor. All MocR-like regulators exhibit a PLP attachment site with a conserved lysine residue, which is also evident in the classified MocR-like regulators (Figure 5). It would thus be interesting to study the role of pyridoxal phosphate regulation in the classified regulators. [20].

YtrA-like protein of M. smegmatis

The YtrA subfamily is the least represented GntR-like regulator in the bacterial genomes. Among all *M. smegmatis* GntR regulators, only one regulator MSMEG_5174, showed the signatures of the YtrA subfamily member (Table 2, Figure 6). YtrA possesses a reduced C-terminal domain with only two α-helices. The average length of the putative effector binding and/or oligomerization domain is about 50 amino acids [5]. YtrA from *B. subtilis* is an experimentally explored regulator, which is part of a large self-regulated operon. This operon consists of genes encoding the ATP binding cassette (ABC) transport sys-



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

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

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

Figure 4

Structure based sequence analysis of *M. smegmatis* GntR-like regulators by the multiple sequence alignment of C-terminal domains of GntR regulators belonging to the HutC Subfamily. Abbreviations are as indicated in Table 1. Consensus sequence from the multiple sequence alignment has been drawn. High and low consensus levels were fixed arbitrarily at 80% and 40% of identity and are represented respectively by the capital and lowercase letters. Consensus symbol ! used for anyone of IV; $ is anyone of LM; % is anyone of FY; # is anyone of NDQEBZ. In graphical representation α-helix region and β-sheet regions are highlighted with light and dark gray background.

tems in addition to the YtrA [21]. It would be interesting to study further, whether MSMEG_5174 has any role in modulating such an operon.

Operator/binding site analysis

We have tabulated a list of potential operator sites near the perfect palindrome sequence with conserved residues, which are found to be specific for most of the subfamily members (Table 3) [5]. We did not find an operator sequence in the upstream sequences of all the remaining regulators. All the predicted sites were found to be in the upstream region from the translation start site except MSMEG_2599. Identification of these sites is an important step to understand the GntR associated regulon or the gene regulatory network in the genome [22-25].

Ortholog prediction

We have found a number of *M. smegmatis* GntR regulators that are orthologs of proteins from the other species of mycobacteria and *B. subtilis* (Table 4). As orthologs typically share the same function, these regulators could serve as a model to study homologues from the other species of mycobacteria. These characterized orthologs may provide clues for initiating detailed biochemical characterization of *M. smegmatis* proteins. Many putative orthologs were experimentally known like Rv0165c that is involved in regulation of *mce1* operon [6]; GntR, a transcriptional

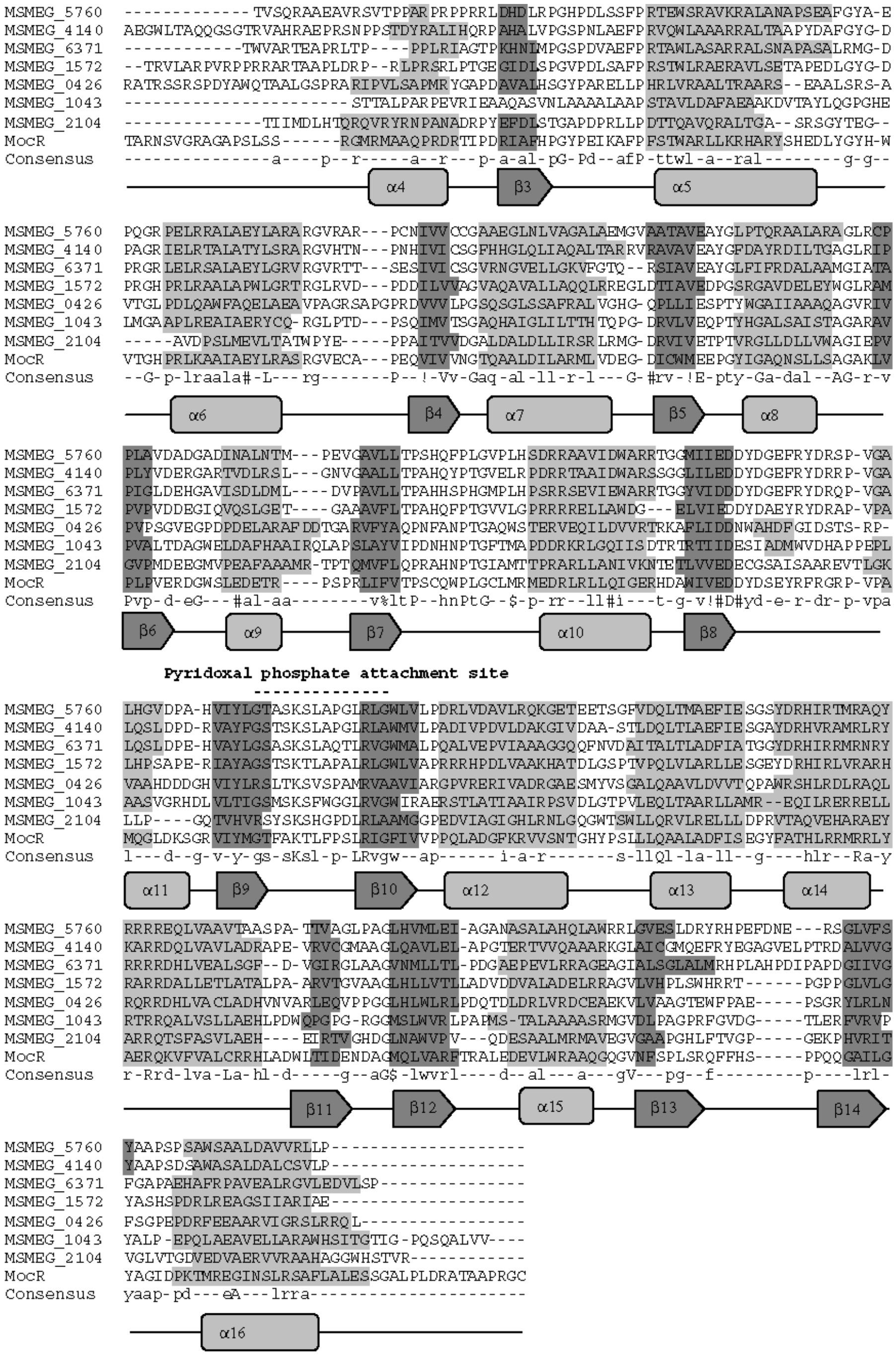
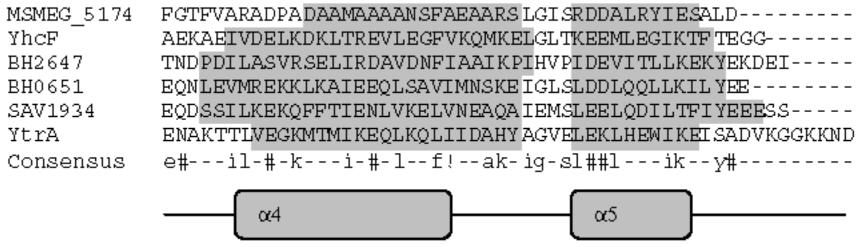

MSMEG_5760 -------------TVSQRAAEAVRSVTPPARPRPPRRLDHDLRPGHPDLSSFPRTEWSRAVKRALANAPSEAFGYA-E
MSMEG_4140 AEGWLTAQQGSGTRVAHRAEPRSNPPSTDYRALIHQRPAHALVPGSPNLAEFPRVQWLAAARRALTAAPYDAFGYG-D
MSMEG_6371 ------------TWVARTEAPRLTP----PPLRIAGTPKHNIMPGSPDVAEFPRTAWLASARRALSNAPASALRMG-D
MSMEG_1572 --TRVLARPVRPPRRARTAAPLDRP--RLPRSRLPTGEGIDLSPGVPDLSAFPRSTWLRAERAVLSETAPEDLGYG-D
MSMEG_0426 RATRSSRSPDYAWQTAALGSPRARIPVLSAPMRYGAPDAVALHSGYPARELLPHRLVRAALTRAARS--EAALSRS-A
MSMEG_1043 -----------------------STTALPARPEVRIEAAQASVNLAAAALAAPSTAVLDAFAEAAKDVTAYLQGPGHE
MSMEG_2104 --------------TIIMDLHTQRQVRYRNPANADRPYEFDLSTGAPDPRLLPDTTQAVQRALTGA--SRSGYTEG--
MocR TARNSVGRAGAPSLSS------RGMRMAAQPRDRTIPDRIAFHPGYPEIKAFPFSTWARLLKRHARYSHEDLYGYH-W
Consensus ---------------a----p--r-----a--r---p-a-al-pG-Pd--afP-ttwl-a--ral--------g-g--
α4
β3
α5
MSMEG_5760 PQGRPELRRALAEYLARARGVRAR---PCNIVVCCGAAEGLNLVAGALAEMGVAATAVEAYGLPTQRAALARAGLRCP
MSMEG_4140 PAGRIELRTALATYLSRARGVHTN---PNHIVICSGFHHGLQLIAQALTARRVRAVAVEAYGFDAYRDILTGAGLRIP
MSMEG_6371 PRGRLELRSALAEYLGRVRGVRTT---SESIVICSGVRNGVELLGKVFGTQ--RSIAVEAYGLFIFRDALAAMGIATA
MSMEG_1572 PRGHPRLRAALAPWLGRTRGLRVD---PDDILVVAGVAQAVALLAQQLRREGLDTIAVEDPGSRGAVDELEYWGLRAM
MSMEG_0426 VTGLPDLQAWFAQELAEAVPAGRSAPGPRDVVVLPGSQSGLSSAFRALVGHG-QPLLIESPTYWGAIIAAAQAGVRIV
MSMEG_1043 LMGAAPLREAIAERYCQ-RGLPTD---PSQIMVTSGAQHAIGLILTTHTQPG-DRVLVEQPTYHGALSAISTAGARAV
MSMEG_2104 -----AVDPSLMEVLTATWPYE-----PPAITVVDGALDALDLLIRSRLRMG-DRVIVETPTVRGLLDLLVWAGIEPV
MocR VTGHPRLKAAIAEYLRASRGVECA---PEQVIVVNGTQAALDILARMLVDEG-DICWMEEPGYIGAQNSLLSAGAKLV
Consensus --G-p-lraala#-L---rg-------P--!-Vv-Gaq-al-ll-r-l---G-#rv-!E-pty-Ga-dal--AG-r-v
α6
β4
α7
β5
α8
MSMEG_5760 PLAVDADGADINALNTM---PEVGAVLLTPSHQFPLGVPLHSDRRAAVIDWARRTGGMIIEDDYDGEFRYDRSP-VGA
MSMEG_4140 PLYVDERGARTVDLRSL---GNVGAALLTPAHQYPTGVELRPDRRTAAIDWARSSGGLILEDDYDGEFRYDRRP-VGA
MSMEG_6371 PIGLDEHGAVISDLDML----DVPAVLLTPAHHSPHGMPLHPSRRSEVIEWARRTGGYVIDDDYDGEFRYDRQP-VGA
MSMEG_1572 PVPVDDEGIQVQSLGET----GAAAVFLTPAHQFPTGVVLGPRRRRELLAWDG---ELVIEDDYDAEYRYDRAP-VPA
MSMEG_0426 PVPSGVEGPDPDELARAFDDTGARVFYAQPNFANPTGAQWSTERVEQILDVVRTRKAFLIDDNWAHDFGIDSTS-RP-
MSMEG_1043 PVALTDAGWELDAFHAAIRQLAPSLAYVIPDNHNPTGFTMAPDDRKRLGQIISDTRTRTIIDESIADMWVDHAPPEPL
MSMEG_2104 GVPMDEEGMVPEAFAAAMR-TPTQMVFLQPRAHNPTGIAMTTPRARLLANIVKNTETLVVEDECGSAISAAREVTLGK
MocR PLPVERDGWSLEDETR----PSPRLIFVTPSCQWPLGCLMRMEDRLRLLQIGERHDAWIVEDDYDSEYRFRGRP-VPA
Consensus Pvp-d-eG---#al-aa-------v%ltP--hnPtG--$-p-rr--ll#i---t-g-v!#D#yd-e-r-dr-p-vpa
β6
α9
β7
α10
β8
Pyridoxal phosphate attachment site
MSMEG_5760 LHGVDPA-HVIYLGTASKSLAPGLRLGWLVLPDRLVDAVLRQKGETEETSGFVDQLTMAEFIESGSYDRHIRTMRAQY
MSMEG_4140 LQSLDPD-RVAYFGSTSKSLAPGLRLAWMVLPADIVPDVLDAKGIVDAA-STLDQLTLAEFIESGAYDRHVRAMRLRY
MSMEG_6371 LQSLDPE-HVAYLGSASKSLAQTLRVGWMALPQALVEPVIAAAGGQQFNVDAITALTLADFIATGGYDRHIRRMRNRY
MSMEG_1572 LHPSAPE-RIAYAGSTSKTLAPALRLGWLVAPRRRHPDLVAAKHATDLGSPTVPQLVLARLLESGEYDRHIRLVRARH
MSMEG_0426 VAAHDDDGHVIYLRSLTKSVSPAMRVAAVIARGPVRERIVADRGAESMYVSGALQAAVLDVVTQPAWRSHLRDLRAQL
MSMEG_1043 AASVGRHDLVLTIGSMSKSFWGGLRVGWIRAERSTLATIAAIRPSVDLGTPVLEQLTAARLLAMR--EQILRERRELL
MSMEG_2104 LLP----GQTVHVRSYSKSHGPDLRLAAMGGPEDVIAGIGHLRNLGQGWTSWLLQRVLRELLLDPRVTAQVEHARAEY
MocR MQGLDKSGRVIYMGTFAKTLFPSLRIGFIVVPPQLADGFKRVVSNTGHYPSLLLQAALADFISEGYFATHLRRMRRLY
Consensus l---d--g-v-y-gs-sKsl-p-LRvgw--ap-----i-a-r-------s-llQl-la-ll--g----hlr--Ra-y
α11
β9
β10
α12
α13
α14
MSMEG_5760 RRRREQLVAAVTAASPA-TTVAGLPAGLHVMLEI-AGANASALAHQLAWRRLGVESLDRYRHPEFDNE---RSGLVFS
MSMEG_4140 KARRDQLVAVLADRAPE-VRVCGMAAGLQAVLEL-APGTERTVVQAAARKGLAICGMQEFRYEGAGVELPTRDALVVG
MSMEG_6371 RRRRDHLVEALSGF--D-VGIRGLAAGVNMLLTL-PDGAEPEVLRRAGEAGIALSGLALMRHPLAHPDIPAPDGIIVG
MSMEG_1572 RARRDALLETLATALPA-ARVTGVAAGLHLLVTLLADVDDVALADELRRAGVLVHPLSWHRRT------PGPPGLVLG
MSMEG_0426 RQRRDHLVACLADHVNVARLEQVPPGGLHLWLRLPDQTDLDRLVRDCEAEKVLVAAGTEWFPAE-----PSGRYLRLN
MSMEG_1043 RTRRQALVSLLAEHLPDWQPGPG-RGGMSLWVRLPAPMS-TALAAAASRMGVDLPAGPRFGVDG-----TLERFVRVP
MSMEG_2104 ARRQTSFASVLAEH----EIRTVGHDGLNAWVPV---QDESAALMRMAVEGVGAAPGHLFTVGP-----GEKPHVRIT
MocR AERQKVFVALCRRHLADWLTIDENDAGMQLVARFTRALEDEVLWRAAQGQGVNFSPLSRQFFHS-----PPQQGAILG
Consensus r-Rrd-lva-La-hl-d-----g--aG$-lwvrl----d--al---a---gV---pg--f---------p----lrl-
β11
β12
α15
β13
β14
MSMEG_5760 YAAPSPSAWSAALDAVVRLLP--------------------
MSMEG_4140 YAAPSDSAWASALDALCSVLP--------------------
MSMEG_6371 FGAPAEHAFRPAVEALRGVLEDVLSP---------------
MSMEG_1572 YASHSPDRLREAGSIIARIAE--------------------
MSMEG_0426 FSGPEPDRFEEAARVIGRSLRRQL-----------------
MSMEG_1043 YALP-EPQLAEAVELLARAWHSITGTIG-PQSQALVV----
MSMEG_2104 VGLVTGDVEDVAERVVRAAHAGGWHSTVR------------
MocR YAGIDPKTMREGINSLRSAFLALESSGALPLDRATAAPRGC
Consensus yaap-pd---eA---lrra---------------------
α16

Figure 5

Structure based sequence analysis of *M. smegmatis* GntR-like regulators by the multiple sequence alignment of C-terminal domains of GntR regulators belonging to the MocR Subfamily. Abbreviations are as indicated in Table 1. Consensus sequence from the multiple sequence alignment has been drawn. High and low consensus levels were fixed arbitrarily at 80% and 40% of identity and are represented respectively by the capital and lowercase letters. Consensus symbol ! used for anyone of IV; $ is anyone of LM; % is anyone of FY; # is anyone of NDQEBZ. In graphical representation α-helix region and β-sheet regions are highlighted with light and dark gray background.



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

Figure 6

Structure based sequence analysis of *M. smegmatis* GntR-like regulators by the multiple sequence alignment of the C-terminal domains of GntR regulators belonging to YtrA Subfamily. Abbreviations are as indicated in Table 1. Consensus sequence from the multiple sequence alignment has been drawn. High and low consensus levels were fixed arbitrarily at 80% and 40% of identity and are represented respectively by the capital and lowercase letters. Consensus symbol ! used for anyone of IV; $ is anyone of LM; % is anyone of FY; # is anyone of NDQEBZ. In graphical representation α-helix region and β-sheet regions are highlighted with light and dark gray background.

repressor of gluconate operon [12]; YcbG, involved in utilization of D-glucarate and D-galactarate [26]; YcnF, involved in utilization of gamma-aminobutyrate [27]. However, we did not find the orthologs for all *M. smegmatis* GntRs in other pathogenic species.

Our results help in extending the annotation of GntRs encoded in the *M. smegmatis* genome. We have classified putative *M. smegmatis* GntRs into four subfamilies. Though in the present study, we have made an attempt to explore *M. smegmatis* GntR regulators, this approach could also be effectively employed to extend the GntR family classification in other bacterial species as well.

Conclusion

This analysis has shown that *M. smegmatis* is equipped with large number of GntR-like regulators, belonging to four subfamilies. It further suggests that the GntR regulatory repertoires of *M. smegmatis* are far more complex than in *M. tuberculosis*. Indeed, additional GntR regulators possibly control a subset of genes required for adapting to a range of environmental conditions. One of the FadR-like regulators shows additional secondary structural ele-

Table 3: List of predicted potential operator sites

Subfamily	Regulator	Potential Operator sequence
FadR	MSMEG_0124	--CCACT**GT**TCA**AC**G**A**GCG---
	MSMEG_0179	-AAGA**T**C**GT**CCG**AC**A**A**TT----
	MSMEG_0454	--CAA**T**C**GT**CAT**AC**G**A**TTG---
	MSMEG_0596	--GTG**T**G**GT**CAG**AC**C**A**CAC---
	MSMEG_0895	-----**T**C**GT**GGG**AC**G**A**------
	MSMEG_2164	-----CC**GT**TGA**AC**GG------
	MSMEG_2480	---ACCG**GT**GGC**AC**C**A**GGGT--
	MSMEG_2599	----ACC**GT**GGG**AC**GGT-----
	MSMEG_2682	-----TG**G**CAAG**AC**C**A**------
	MSMEG_2910	CCTTGAT**GT**CCC**AC**A**A**CG----
	MSMEG_3527	-----**T**G**GT**AAG**AC**C**A**------
	MSMEG_3822	-----**T**T**GT**TACT**C**A**A**------
	MSMEG_3959	--TTGCC**G**CGCG**AC**A**A**------
	MSMEG_3980	-----**T**G**GT**GAT**AC**ACCA----
	MSMEG_4057	----T**T**C**GT**GTC**AC**A**A**GTCGAA
	MSMEG_6789	----T**T**T**GT**GTC**AC**A**A**A-----
HutC	MSMEG_0268	-----ACC**GT**C**TA**C**A**TCGT---
	MSMEG_0650	------TG**GT**C**TA**T**AC**CA----
YtrA	MSMEG_5174	---**G**CCA**T**CATG**TA**GTG**C**----

Preferred nucleotides in potential operator sites are printed in bold

Table 4: Orthologs of *M. smegmatis* GntR-like regulators in other bacterial species

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

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

ments, suggesting a possible origin of a new group within the FadR subfamily. Identified orthologs from *M. smegmatis* could serve as a model to decipher molecular regulation events taking place in the pathogenic mycobacteria. Potential operator sites were also identified based on the nucleotide recognition preferences of GntR-like regulators.

Methods

Selection of GntR-like Members

The sequences of *M. smegmatis MC2* were downloaded from the Institute for Genomic Research Comprehensive Microbial Resource [28]. Apart from classified GntR regulators or proteins annotated as GntR-like regulator, other putative GntRs from *M. smegmatis* proteome were selected using GntR Pfam profile [29]. Among all predicted GntRs one protein (MSMEG_3400) was discarded for this study because of its unusual size (741 amino acid) and its annotation as glutamyl-tRNA(Gln) amidotransferase subunit A. Rest of the GntR regulators were retrieved from the SWISS-PROT/TrEMBL sequence database as per their Swiss-Prot ID (Table 1). Additionally published and annotated genome sequences of *M. tuberculosis*, *M. avium subsp. paratuberculosis*, *M. bovis*, *M. ulcerans*, *M. sp* KMS, *M. sp. MCS*, *M. vanbaalenii* PYR-1 and *Bacillus subtilis* were downloaded from the NCBI ftp site [30].

Secondary structure prediction

The secondary structural features of all bacterial GntR regulators including the *M. smegmatis* GntRs were analyzed (Table 1 and Table 2). Secondary structure predictions were made using Jpred [31], SsPro [32] and 3DPSSM [33]. A consensus of all the secondary structure predictions was considered for a better validity.

Multiple sequence alignments and Phylogenetic tree construction

Multiple sequence alignment was generated with MULTIALIN [34]. Distances between aligned proteins were computed with the PROTDIST program using the Dayhoff PAM matrix [35]. The FITCH program estimated phylogenies from distances in the matrix data using the Fitch-Margoliash algorithm [36]. The phylogenetic tree was drawn using the TREEVIEW program with incorporation of bootstrap values that were obtained involving 1000 replicates [37]. PROTDIST and FITCH programs are included in the PHYLIP package developed by Felsenstein [38].

Operator site analysis

To study the upstream region of GntR-like regulators, we considered sequences from 400 bases upstream to 50 bases downstream from the translation start site. As many GntR regulators are reported to recognize palindromes and also exhibit nucleotide recognition preferences among the same subfamily [5], we utilised these clues to scan the upstream sequences.

Reciprocal BLAST

Reciprocal BLAST hits are frequently utilized to identify the orthologs in two species [39,40]. In this method we searched for the best reciprocal BLAST hit for *M. smegmatis* GntR proteins with *M. tuberculosis*, *M. avium*, *M. bovis*, *Mycobacterium ulcerans*, *Mycobacterium sp* KMS, *Mycobacterium sp.* MCS, *Mycobacterium vanbaalenii* PYR-1 and *B. subtilis*.

Abbreviations

M. tuberculosis – *Mycobacterium tuberculosis*

M. bovis – *Mycobacterium bovis*

M. avium para. – *Mycobacterium avium subsp. paratuberculosis*

M. smegmatis – *Mycobacterium smegmatis*

M. ulcerans – *Mycobacterium ulcerans*

M. sp KMS – *Mycobacterium sp.* KMS

M. sp. MCS – *Mycobacterium sp.* MCS

M. vanbaalenii PYR-1 – *Mycobacterium vanbaalenii* PYR-1.

Authors' contributions

VV carried out the operator site prediction, subfamily data analysis, ortholog search and drafted the manuscript. KS participated in the multiple sequence alignment and structure based manual adjustment. AR participated in the study design and coordination. All authors read and approved the final manuscript.

Acknowledgements

We thank TIGR for making *M. smegmatis* mc2-155 sequences available through their web portal. Research in AR's laboratory is supported by the grants from Department of Biotechnology, Department of Science and Technology, Indian Council of Medical Research (ICMR), Council of Scientific & Industrial Research (CSIR) NMITLI. VV is supported by the UGC fellowship. KS is supported by the CSIR NMITLI Grant. Authors acknowledge the assistance of Ms. P. Umadevi, SRF and Mr. Rohan Misra, JRF for proof reading the manuscript and making useful suggestions.

References

1. Haydon DJ, Guest JR: **A new family of bacterial regulatory proteins.** *FEMS* 1991, **63:**291-295.
2. Marchler-Bauer A, Anderson JB, Cherukuri PF, DeWeese-Scott C, Geer LY, Gwadz M, He S, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Liebert CA, Liu C, Lu F, Marchler GH, Mullokandov M, Shoemaker BA, Simonyan V, Song JS, Thiessen PA, Yamashita RA, Yin JJ, Zhang D, Bryant SH: **CDD: a Conserved Domain Database for protein classification.** *Nucleic Acids Res* 2005, **33:**D192-196.
3. Rigali S, Schlicht M, Hoskisson P, Nothaft H, Merzbacher M, Joris B, Titgemeyer F: **Extending the classification of bacterial tran-**

scription factors beyond the helix-turn-helix motif as an alternative approach to discover new cis/trans relationships. *Nucleic Acids Res* 2004, **32**:3418-3426.

4. Lee MH, Scherer M, Rigali S, Golden JW: **PlmA, a new member of the GntR family, has plasmid maintenance functions in *Anabaena sp*. strain PCC 7120.** *J Bacteriol* 2003, **185**:4315-4325.
5. Rigali S, Derouaux A, Giannotta F, Dusart J: **Subdivision of the helix-turn-helix GntR family of bacterial regulators in the FadR, HutC, MocR, and YtrA subfamilies.** *J Biol Chem* 2002, **277**:12507-12515.
6. Casali N, White AM, Riley LW: **Regulation of the *Mycobacterium tuberculosis mce1* operon.** *J Bacteriol* 2006, **188**:441-449.
7. Vindal V, Ranjan S, Ranajan A: **In silico analysis and characterization of GntR family of regulators from *Mycobacterium tuberculosis*.** *Tuberculosis* 2007, **87**:242-247.
8. Pellicer MT, Fernandez C, Badia J, Aguilar J, Lin EC, Baldom L: **Cross-induction of *glc* and *ace* operons of *Escherichia coli* attributable to pathway intersection. Characterization of the glc promoter.** *J Biol Chem* 1999, **274**:1745-1752.
9. Quail MA, Guest JR: **Purification, characterization and mode of action of PdhR, the transcriptional repressor of the *pdhR-aceEF-lpd* operon of *Escherichia coli*.** *Mol Microbiol* 1995, **15**:519-529.
10. Nunez MF, Pellicer MT, Badia J, Aguilar J, Baldoma L: **The gene *yghK* linked to the *glc* operon of *Escherichia coli* encodes a permease for glycolate that is structurally and functionally similar to L-lactate permease.** *Microbiology* 2001, **147**:1069-1077.
11. Lee HY, An JH, Kim YS: **Identification and characterization of a novel transcriptional regulator, MatR, for malonate metabolism in *Rhizobium leguminosarum bv. trifolii*.** *Eur J Biochem* 2000, **267**:7224-7230.
12. Reizer A, Deutscher J, Saier MH Jr, Reizer J: **Analysis of the gluconate (*gnt*) operon of *Bacillus subtilis*.** *Mol Microbiol* 1991, **5**:1081-1089.
13. Poupin P, Ducrocq V, Hallier-Soulier S, Truffaut N: **Cloning and characterization of the genes encoding a cytochrome P450 (PipA) involved in piperidine and pyrrolidine utilization and its regulatory protein (PipR) in *Mycobacterium smegmatis* mc2155.** *J Bacteriol* 1999, **181**:3419-3426.
14. Aravind L, Anantharaman V: **HutC/FarR-like bacterial transcription factors of the GntR family contain a small molecule-binding domain of the chorismate lyase fold.** *FEMS Microbiol Lett* 2003, **222**:17-23.
15. Allison SL, Phillips AT: **Nucleotide sequence of the gene encoding the repressor for the histidine utilization genes of *Pseudomonas putida*.** *J Bacteriol* 1990, **172**:5470-5476.
16. Quail MA, Dempsey CE, Guest JR: **Identification of a fatty acyl responsive regulator (FarR) in *Escherichia coli*.** *FEBS Lett* 1994, **356**:183-187.
17. Matthijs S, Koedam N, Cornelis P, De Greve H: **The trehalose operon of *Pseudomonas fluorescens* ATCC 17400.** *Res Microbiol* 2000, **151**:845-851.
18. Chen CM, Ye QZ, Zhu ZM, Wanner BL, Walsh CT: **Molecular biology of carbon-phosphorus bond cleavage. Cloning and sequencing of the *phn* (*psiD*) genes involved in alkylphosphonate uptake and C-P lyase activity in *Escherichia coli* B.** *J Biol Chem* 1990, **265**:4461-4471.
19. Sung MH, Tanizawa K, Tanaka H, Kuramitsu S, Kagamiyama H, Hirotsu K, Okamoto A, Higuchi T, Soda K: **Thermostable aspartate aminotransferase from a thermophilic *Bacillus* species. Gene cloning, sequence determination, and preliminary x-ray characterization.** *J Biol Chem* 1991, **266**:2567-2572.
20. Magarvey N, He J, Aidoo KA, Vining LC: **The *pdx* genetic marker adjacent to the chloramphenicol biosynthesis gene cluster in *Streptomyces venezuelae* ISP5230: functional characterization.** *Microbiology* 2001, **147**:2103-2112.
21. Yoshida KI, Fujita Y, Ehrlich SD: **An operon for a putative ATP-binding cassette transport system involved in acetoin utilization of *Bacillus subtilis*.** *J Bacteriol* 2000, **182**:5454-5461.
22. Yellaboina S, Seshadri J, Kumar MS, Ranjan A: **PredictRegulon: a web server for the prediction of the regulatory protein binding sites and operons in prokaryote genomes.** *Nucleic Acids Res* 2004, **32**:W318-320.
23. Yellaboina S, Ranjan S, Vindal V, Ranjan A: **Comparative analysis of iron regulated genes in mycobacteria.** *FEBS Lett* 2006, **580**:2567-2576.
24. Ranjan S, Seshadri J, Vindal V, Yellaboina S, Ranjan A: **iCR: a web tool to identify conserved targets of a regulatory protein across the multiple related prokaryotic species.** *Nucleic Acids Res* 2006, **34**:W584-587.
25. Ranjan S, Gundu RK, Ranjan A: **MycoperonDB: a database of computationally identified operons and transcriptional units in Mycobacteria.** *BMC Bioinformatics* 2006, **7(Suppl 5)**:S9.
26. Hosoya S, Yamane K, Takeuchi M, Sato T: **Identification and characterization of the *Bacillus subtilis* D-glucarate/galactarate utilization operon *ycbCDEFGHJ*.** *FEMS Microbiol Lett* 2002, **210**:193-199.
27. Belitsky BR, Sonenshein AL: **GabR, a member of a novel protein family, regulates the utilization of gamma-aminobutyrate in *Bacillus subtilis*.** *Mol Microbiol* 2002, **45**:569-83.
28. **The Institute for Genomic Research Comprehensive Microbial Resource** [http://cmr.tigr.org/tigr-scripts/CMR/CmrHomePage.cgi]
29. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14**:755-763.
30. **National Center for Biotechnology Information(NCBI) FTP site** [ftp://ftp.ncbi.nih.gov/genomes/Bacteria/]
31. Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ: **JPred: a consensus secondary structure prediction server.** *Bioinformatics* 1998, **14**:892-893.
32. Cheng J, Randall AZ, Sweredoski MJ, Baldi P: **SCRATCH: a protein structure and structural feature prediction server.** *Nucleic Acids Res* 2005, **33**:W72-76.
33. Kelley LA, MacCallum RM, Sternberg MJ: **Enhanced genome annotation using structural profiles in the program 3D-PSSM.** *J Mol Biol* 2000, **299**:499-520.
34. Corpet F: **Multiple sequence alignment with hierarchical clustering.** *Nucleic Acids Res* 1988, **16**:10881-10890.
35. Young CL, Barker WC, Tomaselli CM, Dayhoff MO: *From Atlas of Protein Sequence and Structure Volume 5*. Issue Suppl 3 Edited by: Dayhoff MO. National Biochemical Foundation, Silver Spring, MD; 1979:73-93.
36. Fitch WM, Margoliash E: **Construction of phylogenetic trees.** *Science* 1967, **155**:279-284.
37. Page RD: **TreeView: an application to display phylogenetic trees on personal computers.** *Comput Appl Biosci* 1996, **12**:357-358.
38. Felsenstein J: **PHYLIP-Phylogeny Interface Package.** *Cladistics* 1989, **5**:164-166.
39. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403-410.
40. Fulton DL, Li YY, Laird MR, Horsman BG, Roche FM, Brinkman FS: **Improving the specificity of high-throughput ortholog prediction.** *BMC Bioinformatics* 2006, **28;7**:270.