Published online 21 May 2009 *Nucleic Acids Research, 2009, Vol. 37, Web Server issue* ***W193–W201***
doi:10.1093/nar/gkp388

RegAnalyst: a web interface for the analysis of regulatory motifs, networks and pathways

Deepak Sharma*, Debasisa Mohanty and Avadhesha Surolia

National Institute of Immunology, Aruna Asaf Ali Marg, New Delhi 110067, India

Received February 6, 2009; Revised April 18, 2009; Accepted April 29, 2009

ABSTRACT

RegAnalyst is a user-friendly web interface that integrates MoPP (Motif Prediction Program), MyPatternFinder (pattern detection tool) and MycoRegDB (mycobacterial promoter and regulatory elements database). Since motif discovery is a challenging task, numerous tools have been developed over the past few years. Strikingly, the existing programs were not successful in detecting the known consensus in all mycobacterial (epitomizing degenerate) datasets even in the absence of noise and their performance was further reduced in the presence of noise. Consequently, MoPP, a *de novo* and greedy (for degeneracy) 'inexact' word-based tool that is tailored to enumerate significantly conserved degenerate oligonucleotide motifs was developed. Benchmarking on datasets from MycoRegDB and SCPD (http://rulai.cshl.edu/SCPD/) indicate that MoPP (i) consistently outperforms other motif discovery tools on highly degenerate as well as less degenerate datasets and (ii) successfully detects completely degenerate motifs (with no two instances of a pattern being exactly the same) even in the presence of noise. We have also developed another accessory program, MyPatternFinder, that scans a given sequence or genome to find exact or approximate matches to a query motif of any length identified by MoPP or any other user-defined motif. RegAnalyst will be a valuable tool for *in silico* analysis of regulatory networks and can be accessed at http://www.nii.ac.in/~deepak/RegAnalyst.

INTRODUCTION

Although transcriptional regulation is one of the most fundamental processes for all forms of life, it still remains an intriguing and challenging subject for biomedical research. Experimental endeavors towards understanding the regulation of genes are laborious, time-consuming and expensive but can be substantially accelerated with the use of *in silico* methods. Computational identification of transcription factor binding sites has proved to be extremely valuable for deciphering complex regulatory networks in functional genomic studies (1,2). Therefore, a variety of computational algorithms for identifying regulatory motifs from DNA sequences, with or without additional information, have been developed over the past few years (1–6). A motif can be represented as a word of length l that occurs in q sequences with k mismatches (7). Motif detection is acknowledged to be challenging, with various problems potentially requiring different algorithms or ensembles of different methods (8). Additionally, often a transcription factor recognizes a highly diversified (i.e. degenerate) set of elements that vary from each other at many positions (high k values). Such high degeneracy (as observed in mycobacteria) poses another obstacle in detecting motifs. A database of promoter and regulatory elements from various mycobacterial species, MycoRegDB, was created with the primary aim of addressing high levels of degeneracy. Surprisingly, the existing programs were not able to detect the obscured mycobacterial motifs very satisfactorily. Therefore, MoPP (Motif Prediction Program), an exhaustive motif discovery tool based on 'inexact' word detection was developed with a focus to detect highly degenerate regulatory elements. Analysis of various mycobacterial datasets from MycoRegDB unambiguously proves the ability of MoPP to identify degenerate motifs in the absence or presence of noise (i.e. background genomic sequences). Furthermore, limited tests suggest that MoPP may be useful in eukaryotes. We used MoPP to identify candidate binding sites in several well studied regulons of *Saccharomyces cerevisiae*. Our results indicate that MoPP outperforms other motif discovery programs on less degenerate datasets (such as those from yeast) as well.

Along with the growth of available genomic information (6,9), our knowledge of organism specific motifs such as promoters, Shine Dalgarno and regulatory sequences has increased (10–17). The ability to search genomic sequences to locate particular patterns in DNA is of

*To whom correspondence should be addressed. Tel: +91 11 26703833; Fax: +91 11 26162125; Email: deepak@nii.res.in deepak.aiims@gmail.com

considerable importance and also helps in designing primers with engineered restriction sites for use in molecular biology experiments. The program MyPatternFinder, which we describe here, is useful for detection of user-tailored motifs in DNA sequences. It uses an exact search method along with an alignment technique to find both exact and approximate copies (with/without indels). Its ability to detect copies with insertions and/or deletions (to any desired level) is unique.

We demonstrate the utility of MyPatternFinder, by successfully identifying and validating distinct motifs (such as promoters or hypoxia consensus sequences) in *Mycobacterium tuberculosis* which differ significantly from those present in other bacterial species, and detection of which proved to be difficult using existing tools. Bacterial persistence is a hallmark of tuberculosis and is thought to result from bacterial adaptation to the prevailing environment within tuberculous lesions and granulomas that are believed to be deficient in oxygen and/or nutrient supply (18). A whole genome microarray analysis revealed widespread changes in gene expression when *M. tuberculosis* was briefly subjected to *in vitro* hypoxic conditions (19). Among the genes that were induced was the two-component regulatory system *devR-devS* suggesting its possible role in mycobacterial latency. Recently, DevR (Rv3133c/DosR) was also reported to be a transcriptional regulator of the hypoxic response in *M. tuberculosis* (13). A hypoxia consensus motif (5′-TTSG GGACTWWAGTCCCSAA-3′) or a variant thereof was detected upstream of nearly all *M. tuberculosis* genes rapidly induced by hypoxia (12,13).

METHODS

MycoRegDB

Transcription start points (TSPs) and regulatory elements experimentally identified in various mycobacterial species [*M. tuberculosis* (strains H37Rv and CDC1551), *M. bovis*, *M. leprae*, *M. smegmatis* and *M. avium* subsp. *paratuberculosis*] were compiled.

MoPP

MoPP is an exhaustive motif discovery tool that is tailored to enumerate significantly conserved degenerate oligonucleotide patterns. Figure 1 shows the schematic representation of MoPP's algorithm. In the first step, MoPP identifies patterns that are overrepresented in the input dataset (FASTA format). By default, the program initially searches for motifs that are ≥80% identical and present in ≥70% of the sequences (high stringency). Subsequently, the stringency is reduced to detect motifs that are ≥70% identical and present in ≥60% of the sequences after masking out the motifs already found (medium stringency). Finally, the stringency is reduced to detect motifs that are ≥60% identical and present in ≥50% of the sequences after masking out the motifs already found (low stringency). Using advanced options, a user also has the freedom to specify the cut-offs for percent identity and percent sequences that should contain the motif. Consensus sequence (at each position a nucleotide or set of nucleotides present in ≥60% of the sequences is selected) and enrichment (ratio of copy number in input dataset to that in the non-coding regions of the genome) are then computed for each of the patterns.

In the second step, exact/degenerate matches to these consensus sequences are also searched for in the input dataset. Furthermore, for every group of similar patterns/consensus sequences identified in the first step, a consensus sequence is computed and searched for in the input dataset (as in the first step). All the patterns identified in the first and second steps are ranked on the basis of copy number (R_{cp}) and enrichment (R_{en}). The final score of each pattern is given by $(1/R_{avg}) \times 100$, where $R_{avg} = (R_{cp} + R_{en})/2$.

MoPP's performance was compared with other programs by motif level success rate score mS_r, which is defined as the number of target motif groups N_p that have at least one correctly predicted binding site divided by the total number of target motifs M [$mS_r = N_p/M$] (8). However, the programs YMF, PRISM and Oligo only report the detected motif, its statistical score(s) and/or count, but do not explicitly provide the binding sites or their locations. Therefore, while calculating mS_r we have considered the detection of 'motif' instead of 'binding site' (on the presumption that if motif has been correctly detected then at least one binding site would definitely have been predicted correctly). Furthermore, to avoid any bias due to different number of motifs predicted by various programs, we have considered only the top five motifs for each program as suggested by Tompa *et al.* (20). The scalability issue, as to how the algorithm performance changes with the motif width and the sequence length, is also addressed (8). Therefore, yeast datasets for various motif lengths (6–10 bp) each with different margin sizes (extending on both sides of target motifs) of 50, 100, 200, 300, 400, 500 and 800 bp were generated and analyzed with MoPP by mS_r as well as performance coefficient at binding site level [*sPC*] (8). The *sPC* score indicates whether predicted binding sites overlap with true binding sites (those that have ≥75% matches with the consensus) and is defined as, $sPC = sTP/(sTP + sFP + sFN)$, where *sTP* is the number of predicted binding sites which overlaps with the true binding sites by at least 1 nucleotide, *sFP* is the number of predicted binding sites which have no overlaps with the true binding sites and *sFN* is the number of true binding sites that have no overlaps with any predicted binding sites.

In principle, MoPP has the capability to detect motifs of any length. However, by default, the program searches only for motif widths of 6–8 bp. In case 8-mer motifs are detected, user can repeat the search with longer motif width(s) of his interest. The algorithm also gives user the freedom to allow single/multiple hits of the motif in each input sequence.

MyPatternFinder

The algorithm that is used in MyPatternFinder is mentioned briefly as follows. In the first step (Option A), the input pattern of length N is aligned with the first N bases of the DNA sequence and percentage score is computed

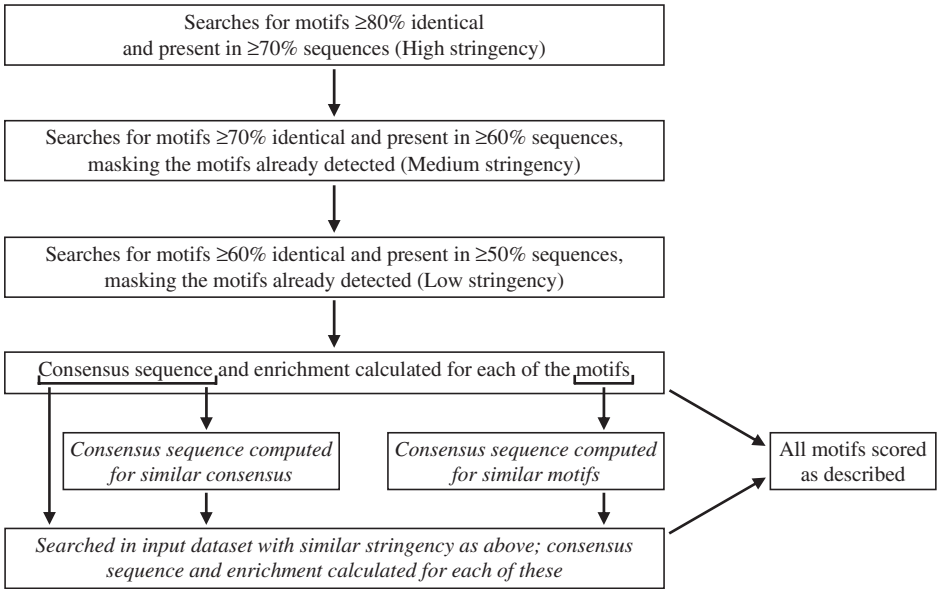

Figure 1. Schematic representation of MoPP's algorithm. The unparalleled ability of MoPP in detecting degenerate motifs is due to the steps indicated in italics.

(1 for every match, 0 for a mismatch) in a sliding window with 1 base shift, along the entire sequence [the method was also incorporated in our previously developed program Spectral Repeat Finder (21)]. If indels are permitted (Option B), the input pattern of length N is aligned with the first M bases ($M = N +$ number of mutations allowed by the user) of the DNA sequence using ClustalW (22) (with gap opening and gap extension penalties of 2.0 each); this allows for indels, and as before, a score can be computed in a sliding window. In the second step, windows where the percentage score exceeds a desired threshold are identified; if there are overlapping patterns, the one with the highest score is considered.

Flexibility has been incorporated into the MyPatternFinder algorithm, so that target patterns can be specified precisely, or with standard abbreviations B, D, H, V, K, M, W, R, S, Y and N if desired. However, in Option B no ambiguous bases can be specified in the query sequence since it is based on ClustalW. Query motifs can also be chosen from a list of available consensus sequences (these will constantly be updated). The first version of MyPatternFinder offers the choice of 34 distinct annotated DNA motifs [15 prokaryotic promoter elements (including 7 from mycobacteria), 4 eukaryotic promoter elements, 9 transcription factors and 6 response elements]. Searches can be carried out in various completely sequenced genomes (choice of >600 organisms is available at present and it will be kept up-to-date in future) and a detailed visualization of the patterns detected along with their positions is provided.

RESULTS AND DISCUSSION

MycoRegDB

MycoRegDB is currently the only available database of promoter/regulatory elements across various mycobacterial species. The first release of MycoRegDB (Supplementary Figure S1) contains 290 annotated DNA motifs (174 promoters and 116 transcription factor binding sites) described in 81 research papers. For each database entry, MycoRegDB gives a variety of information such as gene annotation, CDS positions, promoter/regulatory sequence (with TSP/binding site explicitly marked), TSP-CDS/Motif-CDS distance and hyperlinks to relevant reference(s). Wherever applicable, it also provides hyperlinks to gene information from TubercuList, BCGList and Leproma (http://genolist.pasteur.fr/). These resources are helpful for (i) retrieving DNA/protein sequences, (ii) knowing family classification of genes, and (iii) providing cross-references to UniProt, PDB, PFAM and COG databases. The MycoRegDB will be kept up-to-date in future releases.

Mycobacterial promoters are quite divergent

Among the 174 promoters in MycoRegDB, 118 are those for which the TSPs have been experimentally defined. Of these, for a large subset of 95 promoters the sigma factor(s) recognizing them is/are not known. A majority of these promoters are possibly regulated by the housekeeping sigma factor SigA (23). Alignment of the –10 and –35 regions revealed that there is only ~60% conservation with the known SigA consensus in both these regions (Supplementary Figure S2). Only one of the –10 regions and six of the –35 regions showed perfect match to the –10 and –35 consensus, respectively. This indicates that there exists considerable degeneracy in mycobacterial promoters. Furthermore, our analysis does not seem to suggest that –35 regions are conserved to lesser extent in comparison to –10 regions (17). This discrepancy could possibly be due to accumulation of additional data over recent years. The remaining 23 promoters were divided into subsets on the basis of the involvement of a given sigma factor (SigC: 1, SigD: 6, SigF: 1, SigH: 10 and

Table 1. Performance comparison of MoPP with five popular motif finders on mycobacterial datasets

Regulon	Consensus	Size	MoPP[a]	YMF	PRISM	SCOPE	Oligo	MEME
MycoSigA-10	TATAMT	95	**TAYAVT** (1)[b]	TATtrW (5) TATtAW (6)	t**TAcAAT** (3)	**TANDVT**gk (2)	**TAgACT** (1) **TAcAAT** (2)	**TAgACT** (1)
MycoSigA-35	TTGACW	95	c**TKGAC** (1) c**TBGAC** (3)	TTGACW (6)	gnh**WTGACW** (1)	wy**TTGMMW** (1)	**TTGACT** (2)	**TTGACT** (1)
MycoSigA50bp	TATAMT	95	**TATACT** (2) **TAKACT** (3)	TAgWCW (14)	tTAcAAT (14)	ataTHDMAY (2)[c]	TAgACT (6)	TATtAT (11)
	TTGACW		**TRACT**a (1) **TaKACT** (3)	TaGWCW (14) TWGACW (22)			**TTGACT** (4)	**TTGACT** (2)
MycoSigD-10	WNATGT[d]	6	g**TTATG** (1) g**TTABG** (4)	ACATaT (15)				
MycoSigD-35	GTAACG	6	g**GWAWC** (3)	g**GTAAC** (2)			**GTAACG** (1)	**GTAACG** (1)
MycoSigH-10	SGTTS	10	t**CGTT** (1) g**CGKT** (2)	SGTTar (21)	c**GGTT** (3)	c**GGTT** (3)	g**CGTT** (1)	c**CGTT** (2)
MycoSigH-35	SGGAAC	10	**GGGAAt** (1) **GGGAAY** (2)	**GGGAAC** (1) **GGGAAY** (2)	**CGGAA** (2)	**CGGAA** (2)	**GGGAAC** (1)	**GGGAAC** (1)
MycoSigL-10	CGTGTC	5	**CGTGTC** (1)	**GTGTC**a (5)			**GTGTC**a (1) **CGTGTC** (2)	**GTGTC**a (1)
MycoSigL-35	TGAACC	5	t**TGAAC** (1) b**TGAAC** (2)	**TGWACY** (3) **TGAACY** (5)	**TGAAC** (1)	b**TGAAC** (1)	**TGAACC** (1)	c**TGAAC** (1)
mS_r			**1.0**	**0.4**	**0.5**	**0.5**	**0.8**	**0.8**

[a]Weeder could not be compared since the background file was not available.
[b]Pattern is highlighted in bold if it matches the consensus with not more than one mismatch and ranks among top five. Number in parenthesis indicates rank of the pattern.
[c]Not considered a match since ≥80% of the matching residues are degenerate nucleotides or matching with degenerate nucleotides.
[d]According to MtbRegList (www.usherbrooke.ca/vers/MtbRegList).

SigL: 5). Here again, the promoter elements were quite degenerate although less than SigA dataset (Supplementary Figure S3). However, it would be important to point out that these datasets are small in size and the level of degeneracy is expected to increase as more data get accumulated.

Evaluating MoPP and other motif prediction programs on mycobacterial datasets in absence of noise

The –10 and –35 regions from the SigA, SigD, SigH and SigL class (Supplementary Figures S2 and S3) were then used to evaluate MoPP, YMF (24), Oligo (25), MEME (26), PRISM (27) and SCOPE (28). MoPP was successful in detecting the known consensus in all eight datasets (Table 1). However, even in the absence of noise, the existing programs were not totally successful; MEME and Oligo, closely followed MoPP, with being successful in seven datasets (Table 1). The ensemble program SCOPE was able to detect the consensus in only five datasets. This enhanced ability of MoPP to detect highly degenerate motifs is because the algorithm (i) deduces the consensus sequences in three different ways, and (ii) allows imperfections not only in the initial step but also each time it searches for matches to the consensus in the input dataset (Figure 1).

Evaluating MoPP and other motif prediction programs on mycobacterial datasets in presence of noise

Input sequences for motif finding programs typically consist of motifs buried in noise. Therefore, to simulate real scenario, we made a dataset (MycoSigA50bp) by extracting 50 bp sequences upstream of TSPs (encompassing both –10 and –35 regions) for the SigA regulon. This formed an ideal dataset since it contained 95 genes with highly degenerate motifs. Here also, MoPP was successful in detecting both –10 and –35 consensus sequences (Table 1 and Figure 2). None of the other programs was able to detect the –10 consensus for which only a single perfect match occurred in the whole dataset. However, both MEME and Oligo were able to find a pattern TTGACT that matched with the –35 consensus since there existed six exact occurrences of this pattern in the dataset. This illustrates the ability of MoPP in detecting a completely degenerate motif (with not even two instances of exact match to the consensus) in the presence of noise.

MoPP is not restricted to mycobacteria

To demonstrate that MoPP algorithm is not organism specific, we compared MoPP against other programs on 20 well-characterized *S. cerevisiae* regulons (http://rulai.cshl.edu/SCPD/). MoPP was able to detect the known consensus in 12 of 20 regulons (Supplementary Table S1). Interestingly, MoPP ($mS_r = 0.60$) outperformed all other programs including SCOPE ($mS_r = 0.55$), which combines the output of three different programs. The overall comparison of MoPP with other tools across a total of 30 datasets derived from mycobacteria (with or without noise) and yeast also revealed that MoPP ($mS_r = 0.74$) outperformed SCOPE ($mS_r = 0.54$) (Figure 3a). MoPP was followed by MEME and Oligo which had an mS_r of 0.64 and 0.57, respectively. However, it would be important to point out that the superior performance of MoPP was primarily because of its ability in detecting highly degenerate motifs present in mycobacterial datasets wherein it outperformed other programs by 20–60%.

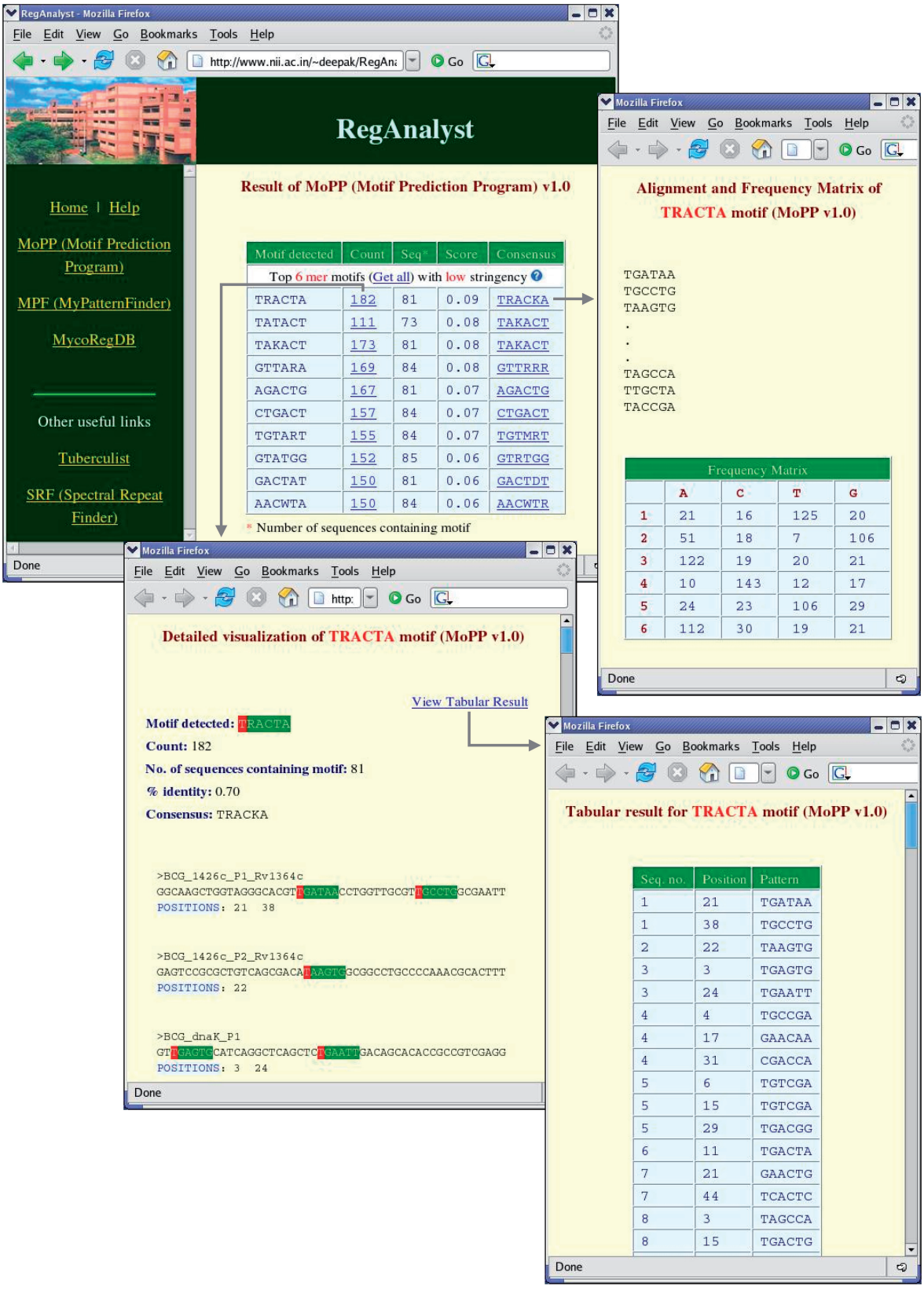

Figure 2. A typical output of MoPP on analysis of a large (95 genes) and highly degenerate dataset, MycoSig50bp. MoPP successfully identified both –10 (motifs ranked 2 and 3) and –35 consensus (motifs ranked 1 and 3) sequences. For each of the detected motif, user can view (i) a colored display of patterns along with their positions (by clicking on the count link), (ii) a tabular output of patterns and their positions and (iii) alignment and frequency matrix of patterns (by clicking on the consensus sequence).

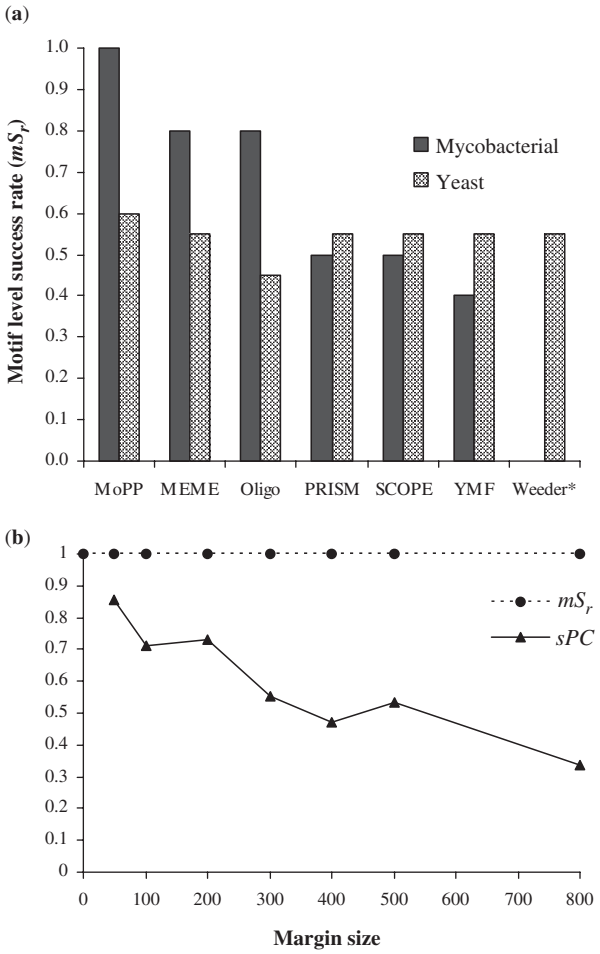

Figure 3. (**a**) Performance comparison of MoPP with other motif discovery tools on 30 datasets derived from mycobacteria and yeast. *Weeder could not be assessed on mycobacterial datasets since the background file was not available. (**b**) Scalability of MoPP in terms of motif level success rate (mS_r) and performance coefficient at binding site level (*sPC*) with respect to the sequence length (margin size).

Furthermore, MoPP's motif level success rate (mS_r) was not affected by sequence length and/or motif width since it is an exhaustive enumeration program (Figure 3b). These results are consistent with similar observations for MEME (8). It would also be important to mention that for each dataset (irrespective of the margin size) the motif (with $\geq$80% matches with the consensus) was correctly identified. The prediction accuracy at the binding site level (*sPC*) on yeast datasets (Figure 3b) was also higher (for all margin sizes) in comparison to those observed for other programs on *Escherichia coli* datasets (8).

Detecting known consensus sequences by MyPatternFinder

The consensus motif sequences of 7 of the 13 *M. tuberculosis* sigma factors (SigA, SigC, SigD, SigE, SigF, SigH and SigL) have been recently published (10,15). As representative examples, MyPatternFinder was used to search the exact consensus motifs of three sigma factors SigA, SigF and SigH (Table 2; complete details are available at http://www.nii.ac.in/~deepak/MyPattern/supl/sigma). No exact copy of the motif for the primary housekeeping sigma factor, SigA, was found and only four copies of the SigF motif could be located (15). This corroborates our observation that there exists considerable flexibility in promoter recognition and a search for promoter sequences must necessarily accommodate mismatches in sequence or spacing of the bipartite elements. We were indeed able to detect 20 copies of the SigA motif by allowing one mismatch with the consensus sequence, several of which were present upstream of various genes (Table 2). Some of these could possibly also be active in *E. coli* since they are almost identical to *E. coli* σ^{70} consensus promoter sequence; such comparisons with promoters of another organism(s) such as *E. coli* can help in predicting whether the organism(s) is a good candidate for studying these mycobacterial promoters (29). Another interesting finding was that out of the 150 exact copies of SigH motif identified, more than 80% were not present in the upstream region of genes but rather within the protein-coding regions.

Using MyPatternFinder, we also searched for the hypoxia consensus motif (13) in the *M. tuberculosis* H37Rv genome. Complete details of the best 100 motifs identified are available at http://www.nii.ac.in/~deepak/MyPattern/supl/hypmotif. We were not only able to detect all the motifs reported by Park *et al.* (13), but also identified a number of additional motifs among which several were positioned upstream of coding regions (Table 3). Although most of these genes were not hypoxia responsive by microarray analysis (13), one of the genes, *Rv3318* (*sdhA*), was repressed in hypoxia in *M. tuberculosis* H37Rv:Δ*dosR* (13) while another, *Rv1039c* (*PPE15*), was significantly induced within artificial granulomas in mice (30) substantiating our results. Further analysis revealed that a number of motifs (with significantly high scores) were present within protein-coding regions of genes, a majority of which were also not regulated by hypoxia. The possible significance of this observation is unclear at present.

The utility of this server is also not limited to mycobacterial sequences: we screened for thyroid hormone response elements (TREs) which are regulatory sequences known to exist upstream of metallothionein genes (31). The metallothionein protein protects the cell against excess concentrations of heavy metals, by binding the metal and removing it from the cell. The gene is expressed at a basal level, but is induced to greater levels of expression by heavy metal ions (such as cadmium) or by glucocorticoids. The TRE has a binding site for transcription factor AP1 and this interaction is part of the mechanism for constitutive expression. Furthermore, this binding reaction is one of the mechanisms (not necessarily the only modality) by which phorbol esters such as TPA (an agent that promotes tumors) trigger a series of transcriptional changes. The TRE motif (TGACTCA) was identified, in 1–6 copies, upstream of various human metallothionein genes (*MT1E*, *MT1K*, *MT2*, *MT3* and *MT4*) when the pattern was allowed to contain indels (Supplementary Figure S4; details are available at http://www.nii.ac.in/~deepak/MyPattern/supl/TRE). It is noteworthy to mention that motif discovery in datasets derived

Table 2. Detection of exact sigma consensus sequences in the complete *M. tuberculosis* H37Rv genome by MyPatternFinder

Sigma factor	Consensus sequence (Ref.)	Total number of hits	Gene[a,b]	Distance from start codon[c]
SigA	TTGACW-N_{17}-TATAMT[d] (15)	0	–	–
	TTGACW-N_{16-21}-TATAMT (15,17)	0	–	–
	TTGACW-N_{16-21}-TATAMT (15,17)	20[e]	Rv0068[f]	−84
			Rv0305c (*PPE*)	−163
			Rvnr01 (*16S rRNA*)	−225
			Rv1403c	−84
			Rv2011c[f]	−50
			Rv2487c (*PE_PGRS*)	−288
			Rv2578c[f]	−35
			Rv3082c (*virS*)[f]	−44
			Rv3760	−485
SigF	GTTT-N_{17}-GGGTAT (15)	4	Rv1248c (*sucA*)	−358
			Rv3287c (*rsbW*/*usfX*)[g]	−35
			Rv3349c	−264
SigH	SGGAAC-N_{17-22}-SGTTS (15)	150	Rv0384c (*clpB*)[g]	−72
			Rv0474	−150
			Rv0563 (*htpX*)	−78
			Rv0569	−475
			Rv1072	−79
			Rv1535	−93
			Rv1786	−448
			Rv1792	−112
			Rv1883c	−217
			Rv2018	−182
			Rv2184c	−178
			Rv2308	−34
			Rv2334 (*cysK*)	−364
			Rv2373c (*dnaJ2*)	−138
			Rv2466c[g]	−77
			Rv2525c	−345
			Rv2694c	−96
			Rv2745c	−66
			Rv2804c	−384
			Rv2839c (*infB*)	−313
			Rv3179	−321
			Rv3482c	−248
			Rv3597c (*lsr2*)	−203
			Rv3832c	−481
			Rv3913 (*trxB2*)[g]	−66

[a]Gene is reported only if the distance of consensus sequence is ≤500 bp upstream of the start codon and it has a non-coding upstream region of ≥25 bp.
[b]According to Cole *et al.* (36).
[c]Location is relative to the translation start site as determined at http://genolist.pasteur.fr/TubercuList, except for Rv3287c (*rsbW*/*usfX*), where location is relative to transcription start site according to Beaucher *et al.* (37).
[d]W = A/T; M = A/C; S = G/C.
[e]By allowing one mismatch in the consensus sequence.
[f]Also predicted to be an *E. coli* σ^{70} promoter with one mismatch.
[g]Involvement of the particular sigma factor has been experimentally verified (37,38).

Table 3. Hypoxia responsive motifs present upstream of genes[a]

Sequence[b]	Score[c]	Gene[d]
ccGGGGAtgAAcGTCCCCgc	11.8486	Rv1039c (*PPE15*)[e]
TgCGGGACTAcAaTCCCGgg	11.7186	Rv1811 (*mgtC*)
ggCGGGACTATgGTCgCGAc	11.414	Rv1552 (*frdA*)
gTCGGGgCggTgGTCCCCgg	11.2576	Rv0345
TTGGGGcCaTccGgCCCGgA	11.195	Rv0877
aTaGtGACaTTcGaCCCGAA	10.8046	Rv3318 (*sdhA*)[f]
aTCGGGcCgAAcGTCaCGAt	10.761	Rv1824
cTCGGGACaTTAcTtCCGtt	10.7435	Rv1881c (*lppE*)
caCGGGACgAgcaTCCCCAg	10.7301	Rv2194 (*qcrC*)
cTCGGGtgTgAgGTCCCatA	10.6815	Rv2221c (*glnE*)
gcCaGGACgTcgGgCCCGAg	10.5356	Rv1256c (*cyp130*)

[a]In addition to those detected by Park *et al.* (13).
[b]Lower case characters show disagreement to motif consensus.
[c]Calculated as mentioned in Park *et al.* (13).
[d]According to Cole *et al.* (36).
[e]Induced in artificial granulomas (30).
[f]Repressed in hypoxia in *M. tuberculosis* H37Rv:Δ*dosR* (13).

from large complex genomes pose certain additional challenges, and the speed and performance of the two algorithms (MoPP and MyPatternFinder) were not assessed on such datasets (e.g. genome-wide ChIP-chip or ChIP-seq datasets).

Validation of motifs detected

MyPatternFinder was used to detect matches to the various sigma consensus elements upstream of the experimentally determined TSPs in the *Rv3134c-devR-devS* operon (32). The $P2_{Rv3134c}$ promoter showed similarity to both *M. tuberculosis* SigA consensus as well as *E. coli* σ^{70} consensus. As predicted, the $P2_{Rv3134c}$ promoter was indeed found to be functional in both *M. smegmatis* [model for studying *M. tuberculosis* promoters since the transcriptional machinery is well conserved between the

two organisms (33)] and *E. coli* (32) substantiating the results of MyPatternFinder.

Distant matches to the DevR consensus motif were also identified in the region encompassing the *devR* upstream region, *Rv3134c* coding sequence and *Rv3134c* upstream region (32). Although these low scoring Dev boxes did not show interaction with DevR (34), their comparison with various high scoring Dev boxes revealed the importance of C_8 base in the consensus motif (35).

CONCLUSION

We have unambiguously proved the efficacy of MoPP (i) in prokaryotes and lower eukaryotes, (ii) in detecting motifs of various lengths, (iii) in detecting highly degenerate as well as less degenerate motifs, and (iv) in the presence of high noise (large sequence lengths). Similarly, the utility of MyPatternFinder has been shown (i) in prokaryotes and small eukaryotic sequences, (ii) in short sequences as well as complete less complex genomes, and (iii) for various consensus sequences (sigma factors, hypoxia motifs and TREs). Thus, both MoPP and MyPatternFinder work efficiently for smaller, less complex genomes and may also be useful for higher eukaryotes with larger, more complex genomes. The patterns detected using MyPatternFinder have been experimentally validated. The detection of conserved motifs (by MoPP) and user-defined patterns of interest (by MyPatternFinder) in genomic sequences should facilitate the understanding of gene expression and regulatory pathways in biological systems.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

D.S. is grateful to Dr R. Ramaswamy, Dr J. S. Tyagi, Dr G. P. S. Raghava and Dr Biju Issac for their valuable help during development of the MyPatternFinder program.

FUNDING

Research Fellowships from Department of Biotechnology (DBT), Government of India and Indian National Science Academy (INSA); core and BTIS project grants from DBT (to D.M.); Centre of Excellence by DBT (to A.S.). Funding for open access charge: Department of Biotechnology.

Conflict of interest statement. None declared.

REFERENCES

1. Wyrick,J.J. and Young,R.A. (2002) Deciphering gene expression regulatory networks. *Curr. Opin. Genet. Dev.*, **12**, 130–136.
2. Duret,L. and Bucher,P. (1997) Searching for regulatory elements in human noncoding sequences. *Curr. Opin. Struct. Biol.*, **7**, 399–406.
3. Brazma,A., Jonassen,I., Vilo,J. and Ukkonen,E. (1998) Predicting gene regulatory elements *in silico* on a genomic scale. *Genome Res.*, **8**, 1202–1215.
4. Banerjee,N. and Zhang,M.Q. (2002) Functional genomics as applied to mapping transcription regulatory networks. *Curr. Opin. Microbiol.*, **5**, 313–317.
5. Ohler,U. and Niemann,H. (2001) Identification and analysis of eukaryotic promoters: recent computational approaches. *Trends Genet.*, **17**, 56–60.
6. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
7. D'Haeseleer,P. (2006) What are DNA sequence motifs? *Nat. Biotechnol.*, **24**, 423–425.
8. Hu,J., Li,B. and Kihara,D. (2005) Limitations and potentials of current motif discovery algorithms. *Nucleic Acids Res.*, **33**, 4899–4913.
9. Kauer,G. and Blocker,H. (2003) Applying signal theory to the analysis of biomolecules. *Bioinformatics*, **19**, 2016–2021.
10. Rodrigue,S., Provvedi,R., Jacques,P.E., Gaudreau,L. and Manganelli,R. (2006) The σ factors of *Mycobacterium tuberculosis*. *FEMS Microbiol. Rev.*, **30**, 926–941.
11. Camp,E., Badhwar,P., Mann,G.J. and Lardelli,M. (2003) Expression analysis of a tyrosinase promoter sequence in zebrafish. *Pigment Cell Res.*, **16**, 117–126.
12. Florczyk,M.A., McCue,L.A., Purkayastha,A., Currenti,E., Wolin,M.J. and McDonough,K.A. (2003) A family of *acr*-coregulated *Mycobacterium tuberculosis* genes shares a common DNA motif and requires Rv3133c (*dosR* or *devR*) for expression. *Infect. Immun.*, **71**, 5332–5343.
13. Park,H.D., Guinn,K.M., Harrell,M.I., Liao,R., Voskuil,M.I., Tompa,M., Schoolnik,G.K. and Sherman,D.R. (2003) Rv3133c/*dosR* is a transcription factor that mediates the hypoxic response of *Mycobacterium tuberculosis*. *Mol. Microbiol.*, **48**, 833–843.
14. Puopolo,K.M. and Madoff,L.C. (2003) Upstream short sequence repeats regulate expression of the alpha C protein of group B *Streptococcus*. *Mol. Microbiol.*, **50**, 977–991.
15. Manganelli,R., Provvedi,R., Rodrigue,S., Beaucher,J., Gaudreau,L. and Smith,I. (2004) σ factors and global gene regulation in *Mycobacterium tuberculosis*. *J. Bacteriol.*, **186**, 895–902.
16. Hoh,J., Jin,S., Parrado,T., Edington,J., Levine,A.J. and Ott,J. (2002) The p53MH algorithm and its application in detecting p53-responsive genes. *Proc. Natl Acad. Sci. USA*, **99**, 8467–8472.
17. Gomez,M. and Smith,I. (2000) In Hatfull,G.F. and Jacobs,W.R. Jr (eds), *Molecular genetics of Mycobacteria*. ASM Press, Washington, DC, pp. 111–129.
18. Wayne,L.G. and Sohaskey,C.D. (2001) Nonreplicating persistence of *Mycobacterium tuberculosis*. *Annu. Rev. Microbiol.*, **55**, 139–163.
19. Sherman,D.R., Voskuil,M., Schnappinger,D., Liao,R., Harrell,M.I. and Schoolnik,G.K. (2001) Regulation of the *Mycobacterium tuberculosis* hypoxic response gene encoding α-crystallin. *Proc. Natl Acad. Sci. USA*, **98**, 7534–7539.
20. Tompa,M., Li,N., Bailey,T.L., Church,G.M., De Moor,B., Eskin,E., Favorov,A.V., Frith,M.C., Fu,Y., Kent,W.J. *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, **23**, 137–144.
21. Sharma,D., Issac,B., Raghava,G.P. and Ramaswamy,R. (2004) Spectral Repeat Finder (SRF): identification of repetitive sequences using Fourier transformation. *Bioinformatics*, **20**, 1405–1412.
22. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
23. Hu,Y. and Coates,A.R. (1999) Transcription of two sigma 70 homologue genes, *sigA* and *sigB*, in stationary-phase *Mycobacterium tuberculosis*. *J. Bacteriol.*, **181**, 469–476.
24. Sinha,S. and Tompa,M. (2003) YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res.*, **31**, 3586–3588.
25. van Helden,J., Andre,B. and Collado-Vides,J. (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.*, **281**, 827–842.

26. Bailey,T.L., Williams,N., Misleh,C. and Li,W.W. (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.*, **34**, W369–W373.
27. Carlson,J.M., Chakravarty,A., Khetani,R.S. and Gross,R.H. (2006) Bounded search for de novo identification of degenerate cis-regulatory elements. *BMC Bioinformatics*, **7**, 254.
28. Carlson,J.M., Chakravarty,A., DeZiel,C.E. and Gross,R.H. (2007) SCOPE: a web server for practical *de novo* motif discovery. *Nucleic Acids Res.*, **35**, W259–W264.
29. Verma,A., Kinger,A.K. and Tyagi,J.S. (1994) Functional analysis of transcription of the *Mycobacterium tuberculosis* 16S rDNA-encoding gene. *Gene*, **148**, 113–118.
30. Karakousis,P.C., Yoshimatsu,T., Lamichhane,G., Woolwine,S.C., Nuermberger,E.L., Grosset,J. and Bishai,W.R. (2004) Dormancy phenotype displayed by extracellular *Mycobacterium tuberculosis* within artificial granulomas in mice. *J. Exp. Med.*, **200**, 647–657.
31. Lewin,B. (1997) *Genes VI*. Oxford University Press, New York, pp. 847–883.
32. Bagchi,G., Chauhan,S., Sharma,D. and Tyagi,J.S. (2005) Transcription and autoregulation of the *Rv3134c-devR-devS* operon of *Mycobacterium tuberculosis*. *Microbiology*, **151**, 4045–4053.
33. Bashyam,M.D., Kaushal,D., Dasgupta,S.K. and Tyagi,A.K. (1996) A study of mycobacterial transcriptional apparatus: identification of novel features in promoter elements. *J. Bacteriol.*, **178**, 4847–4853.
34. Chauhan,S. and Tyagi,J.S. (2008) Cooperative binding of phosphorylated DevR to upstream sites is necessary and sufficient for activation of the Rv3134c-*devRS* operon in *Mycobacterium tuberculosis*: implication in the induction of DevR target genes. *J. Bacteriol.*, **190**, 4301–4312.
35. Chauhan,S. and Tyagi,J.S. (2008) Interaction of DevR with multiple binding sites synergistically activates divergent transcription of *narK2-Rv1738* genes in *Mycobacterium tuberculosis*. *J. Bacteriol.*, **190**, 5394–5403.
36. Cole,S.T., Brosch,R., Parkhill,J., Garnier,T., Churcher,C., Harris,D., Gordon,S.V., Eiglmeier,K., Gas,S., Barry,C.E. 3rd. *et al.* (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, **393**, 537–544.
37. Beaucher,J., Rodrigue,S., Jacques,P.E., Smith,I., Brzezinski,R. and Gaudreau,L. (2002) Novel *Mycobacterium tuberculosis* anti-σ factor antagonists control σ^F activity by distinct mechanisms. *Mol. Microbiol.*, **45**, 1527–1540.
38. Manganelli,R., Voskuil,M.I., Schoolnik,G.K., Dubnau,E., Gomez,M. and Smith,I. (2002) Role of the extracytoplasmic-function σ factor σ^H in *Mycobacterium tuberculosis* global gene expression. *Mol. Microbiol.*, **45**, 365–374.