Discovery of DNA operators for TetR and MarR family transcription factors from *Burkholderia xenovorans*


Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

Determining transcription factor (TF) recognition motifs or operator sites is central to understanding gene regulation, yet few operators have been characterized. In this study, we used a protein-binding microarray (PBM) to discover the DNA recognition sites and putative regulons for three TetR and one MarR family TFs derived from *Burkholderia xenovorans*, which are common to the genus *Burkholderia*. We also describe the development and application of a more streamlined version of the PBM technology that significantly reduced the experimental time. Despite the genus containing many pathogenically important species, only a handful of TF operator sites have been experimentally characterized for *Burkholderia* to date. Our study provides a significant addition to this knowledge base and illustrates some general challenges of discovering operators on a large scale for prokaryotes.

INTRODUCTION

Transcription factors (TFs) are central components of gene regulation. TFs enhance or suppress gene transcription by binding to specific DNA sequences (recognition motifs). Knowing the DNA recognition motif of a TF enables identification of its operator sites in a genome, and hence discovery of the genes regulated by the TF. Despite the importance of DNA operator sites in reconstructing and comparing transcriptional regulatory networks, few of them have been characterized. For example, DNA recognition motifs for only about a third of the 333 predicted TFs in the model organism *Escherichia coli* have been experimentally characterized (Gama-Castro et al., 2008). The fraction of TFs characterized in other organisms, even for the clinically significant ones, is generally much lower. For example, the genus *Burkholderia* contains numerous pathogens of humans, animals and plants, including *Burkholderia pseudomallei* (causative agent of melioidosis) and the *Burkholderia cepacia* complex (opportunistic pathogens in cystic fibrosis patients) (Coenye & LiPuma, 2003), yet only a handful of *Burkholderia* TF DNA recognition sequences are known (for examples see Grove, 2010; Weingart et al., 2005).

Abbreviations: cTGP, thermostable green fluorescent protein; EMSA, electrophoretic mobility shift assay; IMG, Integrated Microbial Genomes; PBM, protein-binding microarray; TF, transcription factor.

Nine supplementary tables are available with the online version of this paper.

In this study, we attempted to characterize DNA recognition half-sites, and then to discover putative operator sites for ten TetR and five MarR family TFs from *Burkholderia xenovorans* that are common to the genus *Burkholderia*. A major obstacle to discovering DNA recognition motifs for individual TFs had been the lack of a high-throughput method. The recent development of a universal protein-binding microarray
(PBM) provided a promising solution to this problem (Berger et al., 2006; Mukherjee et al., 2004). In the current version of the PBM technique, the binding affinity of a protein of interest is screened with all theoretically possible decamer DNA sequences (Berger et al., 2006; Mukherjee et al., 2004). We used PBM as the primary technique for our motif discovery effort. Despite the successful application of the PBM technique to hundreds of eukaryotic TFs from yeast (Zhu et al., 2009), nematode (Grove et al., 2009), mouse (Badis et al., 2009; Wei et al., 2010) and human (Wei et al., 2010), the use of PBMs for bacterial TF characterization is extremely limited (Gordon et al., 2011; Pompeani et al., 2008). We report successful characterization of several TFs and demonstration of a more streamlined version of the PBM method.

**METHODS**

**Constructing phylogenetic trees for the TetR and MarR family TFs.** We searched the Integrated Microbial Genomes (IMG) database for protein sequences that are annotated as either TetR or MarR family TFs for *B. xenovorans* and six additional representative species from the genus *Burkholderia*: *B. multivorans*, *B. cenocepacia*, *B. ambifaria*, *B. thailandensis*, *B. vietnamiensis* and *B. pseudomallei*. Our search yielded a total of 247 and 134 protein sequences for TetR and MarR family TFs, respectively. For proteins of each TF family, we performed multiple sequence alignment using the CLUSTAL W algorithm to align and then identify their signature DNA-binding domains (Ramos et al., 2005; Wilkinson & Grove, 2006). Finally, we performed phylogenetic reconstruction using the maximum-likelihood statistical method for each family of TF proteins based on the sequences of their DNA-binding domains.

**Cloning and expression of glutathione-S-transferase (GST)–TF and GFP–TF fusion proteins.** GST–TF fusions were constructed by cloning the target TFs from the *B. xenovorans* into the pGEX-KG expression vector. The sequences of the primers used to amplify the target TF genes are provided in Supplementary Table S1 (available by downloading the online version of this paper). The PCR products were ligated and cloned into an expression vector. The sequences of the primers used to amplify the GST–TF fusions were constructed by cloning cTGP and Bxe_B3018/Bxe_B0886 sequences into a pGEX-KG vector using the BamHI and NcoI restriction sites. Various concentrations of GST–TFs were allowed to bind with the DNA in the presence of 2.5 % non-fat milk solution, 0.2 mg BSA ml⁻¹, and about 50 μg single-stranded DNA ml⁻¹ extracted from salmon testes. Bound TFs were then reacted with Alexa488-labelled α-GST for visualization. Microarray slides were scanned using a Genepix professional 4200A scanner at 5 μm resolution, exciting at 488 nm for the Alexa488 protein signals and at 532 nm for the Cy3 DNA signals, and then analysed by GenPixPro 6.1 software. Multiple laser power settings at a constant PMT level were used to capture the complete range of the protein signals. The protein signal data were then combined using a linear regression method (http://arep.med.harvard.edu/masliner/supplement.htm) (Dudley et al., 2002). The downstream normalization and motif discovery calculations were performed using custom software (http://the_brain.bwh.harvard.edu/PBMAssaySuite/index.html) (Berger & Bulyk, 2009). Sequence logos were obtained from the position weight matrices of the DNA motifs (gapped and ungapped octomers) with the highest E-scores using the ‘enologo’ program (http://www.benoslab.pitt.edu/cgi-bin/enologos/enologos.cgi) (Workman et al., 2005). For the cTGP-fused proteins, PBM experiments were performed in an identical manner except that the reaction with Alexa488-labelled α-GST and the subsequent washing steps were omitted.

**Electrophoretic mobility shift assays (EMSAs) and pull-down experiments.** For EMSAs, dsDNA (1 μM) was incubated with GST-tagged TFs (8 μM) in 1 x PBS buffer in the presence of 10 % (v/v) glycerol at room temperature for 30 min. Free and protein-bound DNAs were separated on a 1 % TBE agarose gel, and then visualized by staining with ethidium bromide solution. For pull-down assays, the binding experiments were performed as described for gel-shift assays, and then each binding reaction (10 μl) was incubated with 2 μl Ni-NTA agarose resin at room temperature. Protein-bound DNAs were separated by precipitating the Ni-NTA agarose resin and then eluted using 250 mM imidazole. Elutions were mixed with denaturing gel-loading buffer and then run on 1 % TAE-agarose gel. DNA was visualized using ethidium bromide staining. Three independent experiments for each DNA sequence were normalized with respect to the native operator sequence and then averaged.

**RESULTS**

**Selection of the conserved TetR and MarR family TFs from *B. xenovorans***

We used phylogenetic analysis to identify the subset of TetR and MarR family TFs in *B. xenovorans* that are
common to the genus *Burkholderia* (see Methods for details). Briefly, our search in the IMG database resulted in a total of 57 and 28 proteins for *B. xenovorans* that are annotated as TetR and MarR family TFs, respectively. To identify the TFs conserved across the genus, we chose six additional representative species from the genus: *B. multivorans*, *B. cenocepacia*, *B. ambifaria*, *B. thailandensis*, *B. vietnamiensis* and *B. pseudomallei*. Among these species, the number of annotated TetR and MarR family TFs varies from 20 to 54 and 9 to 23, respectively. We performed a phylogenetic reconstruction for each TF family (Fig. 1). TF orthologues formed tight clusters in the trees. We found at least 16 TetR family TFs and nine MarR family TFs from *B. xenovorans* that formed tight clusters with TFs from most of the other representative species (Fig. 1). Ten TetR family and five MarR family *B. xenovorans* TFs were chosen for experimental characterization; these TFs had orthologues present in at least five other species (Fig. 1, Table 1).

**PBM with GST-tagged TFs**

We used a PBM technique to discover DNA half-sites for the 15 selected TFs. First, we attempted to clone, overexpress and purify the selected TFs (Table 1). Only 12 TFs were adequately expressed in *E. coli* and successfully purified (Table 1). Next, we performed two or more PBM assays with each purified TF to identify its half-site recognition sequence. Nine of the 12 TFs produced qualitative results resembling successful PBM assays—that is, relatively low non-specific binding and a small set of strong binding events (Mukherjee et al., 2004). However, after using the standard data analysis routine to obtain a list of enriched octomers (with no, one or two gaps) ranked by their enrichment (E) scores (Berger & Bulyk, 2009), five of the nine TFs yielded DNA motifs that were not reproducible, and in most cases had inadequate enrichment scores or E-scores (less than 0.40; typically E-scores ≥ 0.45 indicate good enrichments). These TFs were not studied further. The remaining four TFs yielded consensus DNA motifs with high E-scores (Table 1, Fig. 2 and Supplementary Table S2, available with the online version of this paper). The TFs were Bxe_B0886, Bxe_B3018, Bxe_A0425 – members of the TetR family – and Bxe_A3929 – a member of the MarR family. For all TFs, we also performed an enrichment analysis for 4-gap tetramers (gap length varying from 0 to 12), but the results yielded no new DNA motif with a high E-score.

**From PBM motifs to the discovery of putative regulons**

Online genome-wide scans (http://rsat.ulb.ac.be/rsat/) (van Helden, 2003) with the PBM-derived half-site sequences alone yielded a large number of candidate regulatory regions (data not shown). The length of the spacer between the half-sites is vital to narrow down the list. To determine the spacer length between half-sites, we sought at least one credible operator site upstream of each TF. TetR and MarR family TFs are often auto-regulatory (Ramos et al., 2005; Wilkinson & Grove, 2006). Therefore, the first logical place to seek a candidate operator site is upstream of each TetR or MarR family TF gene. Using this strategy, we found putative operators consistent with the PBM-derived motifs for three of the four TFs (Bxe_B3018, Bxe_A0425 and Bxe_A3929) in the intergenic region upstream of each respective gene as described below.

**TF Bxe_B0886.** We did not find the PBM-derived motif (G/A)GTAAAC(T/C) in the 96 bp intergenic region upstream of the Bxe_B0886 gene or in the region upstream of Bxe_B0887, which is the only local gene that has conserved linkage with Bxe_B0886 across the genus *Burkholderia*. The Bxe_B0886 TF bound its upstream intergenic site in EMSAs, but also exhibited non-specific binding with many other DNA sequences (data not shown). Using an online genome scanning algorithm (http://rsat.ulb.ac.be/rsat/) (van Helden, 2003), we found some candidate sites compatible with the PBM-derived half-site for Bxe_B0886. The most promising candidates were five dimeric binding sites [T(T/A)(N)T(T/A)] that were conserved across the genus *Burkholderia* (see Supplementary Table S3, available with the online version of this paper). Validation of TF binding to these sequences in EMSAs was not successful because we were unable to distinguish between specific and non-specific interactions. Therefore, an operator sequence for Bxe_B0886 is unresolved.

**TF Bxe_B3018.** In contrast with Bxe_B0886, we found a credible operator site for Bxe_B3018 upstream of its own gene (Fig. 3a). An 18 bp candidate operator occurred 10 to 27 bp upstream of Bxe_B3018 in the 108 bp intergenic region. The operator consisted of an inverted repeat of the PBM-derived heptamer half-site C(G/A)TTCTC separated by a 2 bp spacer (Fig. 3a). The operator was validated by demonstrating specific binding of the GST-tagged, Bxe_B3018 TF to a synthetic 36 bp dsDNA representing the operator sequence in an EMSA assay (Fig. 3b). We also tested the stringency of the spacer length between the two half-sites by deleting or inserting nucleotides (Fig. 3b). GST-Bxe_B3018 did not exhibit binding to any non-native operator sequences with altered spacer lengths (Fig. 3b). Identical results were obtained from a pull-down assay, in which a polyhistidine-tagged Bxe_B3018 protein was incubated with the DNAs and then precipitated with Ni-NTA resin (results not shown). The importance of spacer length in controlling TF binding has been demonstrated previously; with the *E. coli* TetR protein, alteration of spacer length by even one nucleotide eliminated protein binding (Wissmann et al., 1988). We also examined the effect of spacer nucleotide composition on protein binding in pull-down assays. All four palindromic combinations of spacer nucleotides (i.e. AT, TA, CG and GC) permitted binding with Bxe_B3018, but the relative affinity differed (Fig. 3c). Collectively, the results suggest that the predicted operator site is authentic and the spacer length, but not the identity of palindromic spacer nucleotides, is critical for stable Bxe_B3018 binding.
TetR family tree

MarR family tree
The predicted Bxe_B3018 operator site was highly conserved upstream of the orthologues of Bxe_B3018 across the genus *Burkholderia* (see Supplementary Table S4, available with the online version of this paper), consistent with the predicted regulatory role for this site. Also, the operator fell within predicted sigma-70 binding sites for both Bxe_B3018 and the divergently oriented Bxe_B3019 gene, which encodes a major facilitator superfamily transporter protein (Chain *et al.*, 2006). Bxe_B3018 and Bxe_B3017 (encoding a protein of unknown function) constitute a predicted operon (Dam *et al.*, 2007; Mao *et al.*, 2009) with only 13 nt between the two genes. The gene block comprising Bxe_B3017, Bxe_B3018, and Bxe_B3019 is conserved across the genus *Burkholderia*, suggesting a functional linkage among the genes. Taken together, we predict that Bxe_B3018 regulates transcription of at least three genes; itself, Bxe_B3017 and Bxe_B3019.

To identify additional Bxe_B3018 regulatory sites, we scanned the intergenic regions of the *B. xenovorans* genome for the substring, AGAA(N)_6TTCT, derived from the four most conserved ‘core’ nucleotides found in the PBM-derived half-site for Bxe_B3018 (Fig. 2b). Matches occurred at 18 intergenic locations, upstream of 23 genes, excluding the previously verified Bxe_B3018/Bxe_B3019 operator site (see Supplementary Table S5, available with the online version of this paper). These 18 sites had multiple mismatches relative to the 18 bp Bxe_B3018/Bxe_B3019 operator site (Supplementary Table S5). We tested GST-Bxe_B3018 binding to four of these 18 sites representing a wide range of mismatches using EMSA (Fig. 4) or in pull-down assays using a polyhistidine-tagged Bxe_B3018 protein (data not shown). None of the sites we tested exhibited binding with Bxe_B3018 (Fig. 4b), suggesting no regulatory relationship between these sites and Bxe_B3018. The negative binding results indicated that the nucleotides flanking the AGAA sequences in each half-site make critical contributions to Bxe_B3018 binding.

**TF Bxe_A0425.** Four half-sites containing no, one or two mismatches with the PBM-derived motif for Bxe_A0425, (G/A)GT(A/C)ACA, occurred between 35 and 76 bp upstream of the Bxe_A0425 gene (Fig. 5a). All of the four sites were found to be conserved to varying degrees upstream of Bxe_A0425 orthologues across the genus *Burkholderia* (see Supplementary Table S6, available with the online version of this paper). The affinity of Bxe_A0425 for this region was tested by EMSA with oligonucleotides containing two, three, or all four of the half-sites (Fig. 5b). Protein binding occurred with the oligonucleotide containing all four half-sites and the oligonucleotide containing only the first two half-sites. An 8 bp spacer separates the first two half-sites. Together, these two half-sites contain only one mismatch relative to the PBM-derived motif and represent a credible operator site. We predict that this 22 bp operator site upstream of Bxe_A0425 controls transcription of Bxe_A0425 as well as

**Table 1.** Summary of expression, purification and PBMs of the TetR and MarR family TFs derived from *B. xenovorans*

<table>
<thead>
<tr>
<th>TF</th>
<th>Family</th>
<th>Expression/purification</th>
<th>DNA binding in PBM</th>
<th>Consensus DNA motif from PBM (max 8 nt)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bxe_B0886</td>
<td>TetR</td>
<td>Yes</td>
<td>Yes</td>
<td>(G/A)TTAAC(T/C)</td>
</tr>
<tr>
<td>Bxe_B3018</td>
<td>TetR</td>
<td>Yes</td>
<td>Yes</td>
<td>C(G/A)TTCTC</td>
</tr>
<tr>
<td>Bxe_A0425</td>
<td>TetR</td>
<td>Yes</td>
<td>Yes</td>
<td>C(G/A)GT(A/C)ACA</td>
</tr>
<tr>
<td>Bxe_A3929</td>
<td>MarR</td>
<td>Yes</td>
<td>Yes</td>
<td>ATATATCA</td>
</tr>
<tr>
<td>Bxe_B1590</td>
<td>TetR</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Bxe_A4039</td>
<td>TetR</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Bxe_A3239</td>
<td>TetR</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Bxe_A2299</td>
<td>MarR</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Bxe_B2842</td>
<td>MarR</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Bxe_A0707</td>
<td>TetR</td>
<td>Yes</td>
<td>No</td>
<td>NA</td>
</tr>
<tr>
<td>Bxe_B0959</td>
<td>TetR</td>
<td>Yes</td>
<td>No</td>
<td>NA</td>
</tr>
<tr>
<td>Bxe_B2027</td>
<td>MarR</td>
<td>Yes</td>
<td>No</td>
<td>NA</td>
</tr>
<tr>
<td>Bxe_A0289</td>
<td>TetR</td>
<td>No</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Bxe_A3309</td>
<td>TetR</td>
<td>No</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Bxe_A0003</td>
<td>MarR</td>
<td>No</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

http://mic.sgmjournals.org
the divergently oriented gene, Bxe_A0426, encoding a major facilitator superfamily transporter system gene. The Bxe_A0425/Bxe_A0426 gene block is conserved across the genus Burkholderia, suggesting functional linkage consistent with our prediction of coordinated expression.

To find additional candidate operator sites for Bxe_A0425, we performed EMSAs with double-stranded DNA oligonucleotides representing each site. All three oligonucleotides exhibited specific binding to GST-Bxe_A0425, demonstrating tolerance of variation in the nucleotide composition and spacer length of the binding site and suggesting that Bxe_A0425 may regulate the genes downstream of these binding sites. The Bxe_A0425 binding site upstream of Bxe_B1789 was also found to be conserved across the genus, further suggesting a regulatory role for this site.

**TF Bxe_A3929.** Bxe_A3929 was the only MarR family protein that yielded a consensus half-site upon PBM data analysis (Table 1, Fig. 2d). The resulting DNA motif, ATATATCA, was found to be partially present (TATATATCA) in the 85 bp intergenic region upstream of the Bxe_A3929 gene (Fig. 7a). An identical motif was also found upstream of many Bxe_A3929 orthologues (see Supplementary Table S7, available with the online version of this paper), suggesting a regulatory role for this motif. We found a nearby DNA sequence that constituted an inverse repeat of the motif with two mismatches with the PBM-derived octomer DNA motif for Bxe_A3929 (Fig. 7a). When tested in EMSA, a 34 bp DNA representing the region containing both the motifs exhibited binding to GST-Bxe_A3929 (Fig. 7b). Furthermore, when the less conserved second DNA motif was partially truncated, no substantial protein binding was observed (Fig. 7b). The region is a credible operator site for Bxe_A3929. The results suggest that tolerance of sequence variation differs greatly for the two half-sites in the Bxe_A3929 operator.

The linkage of Bxe_A3929 with Bxe_A3930, Bxe_A3928 and Bxe_A3927 in the genome is conserved across the genus, consistent with a functional relationship among the genes. The Bxe_A3929 gene encodes a cystathionine beta synthase (CBS) domain-containing a putative voltage-gated chloride channel protein. The separation between the Bxe_A3929 and Bxe_A3930 genes is only 48 bp. A sigma70 binding site is not apparent in the 48 bp intergenic region. These two genes are a predicted operon (Dam et al., 2007; Mao et al., 2009); we predict the operon is regulated by Bxe_A3929. The Bxe_A3928 and Bxe_A3927 genes, which code for a protein containing a CBS domain and a Crp/FNR-family TF, respectively, share a common 351 bp upstream intergenic region (Fig. 6a). Interestingly, we found the exact same Bxe_A3929 monomer-binding motif found upstream of the Bxe_A3929 gene in this 351 bp intergenic region. As in the region upstream of Bxe_A3929, we also found that this sequence is phylogenetically conserved upstream of the Bxe_A3928/Bxe_A3927 orthologues across the genus Burkholderia. In addition, we found an AT-rich sequence partially resembling the PBM-derived motif in the vicinity of this conserved motif. When checked by EMSA, a dsDNA oligonucleotide derived from this region exhibited binding to GST-Bxe_A3929, suggesting that Bxe_B3929 could also regulate Bxe_A3928 and/or Bxe_A3927 genes. Altogether, Bxe_A3929 appears to be involved in regulating an ion transport pathway in B. xenovorans that involves one or more of the Bxe_A3930, Bxe_A3928 and Bxe_A3927 genes.

Despite positive EMSA results with two oligonucleotides (Fig. 7), we could not determine a common theme for the second half-site required for Bxe_A3929 binding. When we searched for the conserved PBM-derived half-site TATATCA...
in the entire \textit{B. xenovorans} genome, we found 69 additional intergenic regions across the genome that contained the search string (data not shown). Five of them contain a second AT-rich region near the PBM motif like the one upstream of the Bxe\_A3929 and Bxe\_A3928 genes (TATATC\_An\_12TAT, see Supplementary Table S8, available with the online version of this paper). But, in none of the five intergenic regions, the PBM motif was found to be phylogenetically conserved across the genus \textit{Burkholderia}, suggesting a less likely regulatory role for these regions.

\textbf{Method improvement}

Although PBMs can be successfully performed using GST-tagged proteins, use of GST as a fusion partner has several potential drawbacks. First, GST itself can dimerize in solution (Riley \textit{et al.}, 1996). When fused with a bacterial TF, which itself can dimerize, unusual multimers may arise and alter the binding of the TF to DNA in the PBM assay. Second, detection of GST requires additional steps, namely reaction with Alexa488-labelled \(\alpha\)-GST and subsequent washing to remove unbound antibody. These steps account for 33\% of the approximately 3 h assay procedure (Berger \& Bulyk, 2009). Based on these considerations, the use of the GST tag may be suboptimal for high-throughput motif discovery.

To develop a more streamlined PBM technique, we cloned, expressed, and purified Bxe\_B3018 and Bxe\_B0886 fused with a thermostable green fluorescent protein (cTGP), and tested the fusion proteins in single-step PBM experiments, i.e. protein binding to the DNA without the additional reaction with Alexa488-labelled antibody step. GST and cTGP have very similar molecular masses (26 and 25 kDa, respectively), and the excitation profiles of Alexa488 and cTGP are almost identical, enabling seamless integration of cTGP with the signal detection and processing steps in the standard PBM protocol.

We found that the results of the PBM experiments with cTGP-tagged and GST-tagged proteins were consistent (compare Figs 2 and 8; see Supplementary Table S9, available with the online version of this paper, for E-scores). However, we did note some differences. One difference was the conservation of degenerate positions in consensus sequences derived from the PBM experiments. For example, the core nucleotides, TTAA or TTCT, for the native (36-nt) 5'-CTTGACGAGG
\textit{GAGAATG}
AGTT
CGTTCTA
CCCTTCGTTC
-3'
Bxe_B0886 and Bxe_B3018 recognition motifs appeared unaffected by the choice of the TF fusion partner, but the nucleotides flanking the cores appeared to differ slightly in the degree of conservation between the two tags for both TFs (compare Figs 2 and 8). With the limited data, we could not determine if these differences were simply the result of experimental variation. Additional systematic experiments are required to draw more definitive conclusions.

**DISCUSSION**

Our principal goal was to discover regulatory modules for 15 TetR and MarR family TFs selected from *B. xenovorans*. To our knowledge this is the first attempt to use the PBM technique to characterize operators for multiple TFs from bacteria. Our success rate in discovering operators for proteins that were effectively purified was 25%. Only four of the 12 successfully purified TFs yielded specific DNA

---

**Fig. 4.** Evaluation of four additional putative Bxe_B3018 operator sites. (a) Sequence comparisons of the four candidate operator sites and the region upstream of the Bxe_B3018 gene containing a common theme, AGAA(N)₆TTCT. Dashes indicate a match with the Bxe_B3018 upstream sequence. (b) EMSAs with Bxe_B3018 and the four candidate operator sites.

**Fig. 5.** Evaluation of Bxe_A0425 binding to its own upstream region. (a) Four PBM-derived 7 bp half-sites (underlined and numbered), (G/A)GT(A/C)ACA, containing no, one or two mismatches, were found upstream of the Bxe_A0425 gene. (b) Three dsDNA oligonucleotides, A0425_4, A0425_3, A0425_2, derived from the region upstream of the Bxe_A0425 gene were tested by using EMSAs for their ability to bind GST-Bxe_A0425 TF.
motifs upon routine PBM data analysis and only three of these led to discovery of credible operators. Multiple factors may have contributed to this low success rate.

First, the stringency of DNA binding specificity may vary among TFs (Lozada-Chávez et al., 2008). The low DNA binding specificity of a TF can be compensated in vivo in various ways, including a high expression level (Lozada-Chávez et al., 2008). It is possible that some TFs we attempted to work with have less stringent DNA binding preferences and therefore failed to yield consensus DNA motifs upon routine PBM data analysis. Second, although many TetR and MarR family TFs do not require any additional cofactors to bind to their cognate DNAs, exceptions do exist (Wilkinson & Grove, 2006). Therefore, some of the TFs we attempted might require additional cofactors to specifically interact with their cognate DNA motifs. Third, the GST tag on the TFs may have interfered with dimerization of some of the TFs or with binding to the PBM. We did not try to address this possibility by cloning and expressing only the DNA-binding domain of each TF or by placing the tag on the C terminus of the TFs. Fourth, the current PBM has been developed to discover TF monomer-binding DNA sites. The present PBM design may be suboptimal for dimeric or multimeric TFs, resulting in a success rate that is inevitably lower for bacterial TFs than for eukaryotic TFs.

PBM assays versus biochemical assays

Although the results of PBM assays, EMSAs and pull-down experiments were generally consistent, some differences were observed. For example, in PBM assays, the presence of

---

**Fig. 6.** Evaluation of three additional putative Bxe_A0425 operator sites. (a) Sequence comparisons of the three candidate operator sites and the region upstream of the Bxe_A0425 gene containing a common theme TGT(A(T) N10–15) TAACA. Dashes indicate a match with Bxe_A0425 upstream sequence. The black lines indicate spacers of variable length and nucleotide compositions. (b) EMSAs to test Bxe_A0425 binding to the three candidate operator sites. A non-specific dsDNA was used as the negative control.

---

**Fig. 7.** Evaluation of putative operator sites for Bxe_A3929. (a) The region upstream of both Bxe_A3929 and Bxe_A3928/Bxe_A3927 genes contains a partial PBM motif (solid underline, the first half-site) and a stretch of nucleotides that weakly constitutes an inverse repeat of the PBM motif (dashed underline, the second half-site). (b) EMSAs to test Bxe_A3929 binding to dsDNAs derived from the sequences upstream of Bxe_A3929 and Bxe_A3928/Bxe_A3927 genes. A3929_1 and A3929_2 were derived from upstream of the Bxe_A3929 gene. In A3929_2, the second half-site was partially truncated. A3928_2 was derived from the intergenic region between the Bxe_A3928 and Bxe_A3927 genes.
a single half-site on a DNA oligonucleotide was adequate for Bxe_B3018 binding (Fig. 3b). Also, some sequence variations were well tolerated. In stark contrast, EMSAs and pull-down assays were more stringent; a dimeric binding site was required, the correct spacer length between half-sites was critical (Fig. 3b) and less sequence variation in the half-sites was tolerated (Fig. 4b). We predict that these differences between the assays arise from differences in the concentration and mobility of the DNA probes. In a PBM, dsDNAs are immobilized at each spot on the microarray slide by their 3′ ends, potentially creating a higher local concentration of identical probes. Consequently, if one monomer of a putative Bxe_B3018 dimer finds its preferred site, then an identical site may be readily available from another DNA molecule for the other monomer to bind to form a stable ternary complex, probably eliminating the requirement for the presence of two correctly spaced half-sites in a single DNA molecule for binding by a dimeric TF.

From PBM results to regulon discovery

After discovering DNA recognition motifs with PBM experiments, there are at least two obstacles that can impede discovery of candidate regulons in a genome of interest. First, a consensus PBM-derived motif represents the strongest binding DNA sequences on the PBM. However, a TF may exploit much weaker binding sites in vivo that do not strongly resemble the primary PBM motif (Berger et al., 2008). Therefore, the failure to find a credible operator for a TF of interest may occur if the operator site in vivo deviates substantially from the PBM-derived half-site, and cannot be easily discovered using the PBM-derived half-site as a search pattern.

Second, the PBM consensus motif only represents a monomer-binding site; whereas, we expect that most, if not all, the TetR and MarR family TFs used in this study function as dimers (or other higher order multimers) and their operator sites in vivo consist of a direct or inverted repeat of their respective monomer-binding sequences. The spacer length between half-sites is a crucial piece of information that drastically reduces the number of false positives arising from genome-wide scans for binding sites. If the spacer length cannot be determined easily, discovery of genuine operators and reconstruction of regulons is undermined. There are algorithms that can derive spacer lengths from PBM data (Pompeani et al., 2008), but our attempt to use these algorithms was unsuccessful.

Given these two obstacles, a critical step in regulon discovery is to find at least one credible operator site in a genome such that the orientation of half-sites and the spacer length are defined. For TFs with known targets, obtaining this information is relatively straightforward, but we did not have any prior information about the regulatory roles for the four TFs, Bxe_B0886, Bxe_B3018, Bxe_A0425, and Bxe_A3929, that resulted in consensus DNA motifs in PBM experiments. Since many TetR and MarR family TFs are known to auto-regulate (Ramos et al., 2005; Wilkinson & Grove, 2006), searching for an operator in the intergenic region upstream of each TF gene was a sensible alternate strategy. This strategy worked for Bxe_B3018, Bxe_A0425 and Bxe_A3929, but will fail if TFs are not auto-regulatory.

Phylogenetic conservation of TF regulatory modules

Our focus on TFs conserved across the Burkholderia genus increases the impact of the findings by enabling extension of results to other Burkholderia species. The tight clustering of orthologous protein sequences suggests that protein function is conserved for the TFs we examined (Fig. 1). Operator sites for the Bxe_B3018, Bxe_A0425 and Bxe_A3929 TFs were also conserved across the genus (Supplementary Tables S3, S5 and S6). This illustrates that our findings with B. xenovorans TFs are applicable to orthologues in other Burkholderia species. It is possible that there are additional operators in individual species and that the complete regulon for each TF varies slightly among the species. The ability to compare the location and composition of operators among genomes is a central component of regulatory genomics (Vingron et al., 2009) and illustrates the value of characterizing operator sequences for as many TFs in organisms of interest as possible.

Conclusions

In B. xenovorans, approximately 6 % of the total 8724 genes are annotated as TFs. None of the TFs previously had operator sequences that had been experimentally characterized. Our work provides operators and putative regulons for 20 % of the TetR and MarR family TFs that are common to the genus Burkholderia and about 2–3 % of the total TFs that are common to the genus. Our findings
provide a first step towards a comprehensive catalogue of common cis-regulatory elements in Burkholderia that can be used to understand similarities and differences in regulons and regulatory behavior within this ecologically and clinically important group of species.

ACKNOWLEDGEMENTS

We sincerely thank Dr Martha Bulyk (Harvard Medical School) for several discussions about the PBM. This work was supported by a grant from the Los Alamos National Laboratory Directed Research program (grant no. DR20090117).

REFERENCES


Edited by: D. W. Ussery