Quantitative whole-genome analysis of DNA-protein interactions by in vivo methylase protection in *E. coli*

Saeed Tavazoie[1] & George M. Church*

*Department of Genetics, Harvard Medical School, Boston, MA, 02115. [1]Graduate Program in Biophysics, Harvard University, Boston, MA 02115. *Corresponding author (e-mail:church@rascal.med.harvard.edu).*

Received 24 February 1998; accepted 29 April 1998

A global methylation-based technique was used to identify, display, and quantitate the in vivo occupancy of numerous protein-binding sites within the *Escherichia coli* genome. The protein occupancy profiles of these sites showed variation across different growth conditions and genetic backgrounds. Of the 25 sites identified in this study, 24 occurred within 5′ noncoding regions. Protein occupancy at 13 of these sites was supported by independent biochemical and genetic evidence. Most of the remaining 12 sites fell upstream of genes with no previously known function. A multivariate statistical analysis was utilized to group such uncharacterized genes with well-characterized ones, providing insights into their function based on a common pattern of transcriptional regulation.

Keywords: genomics, gene expression, cluster analysis

Of central importance to cellular physiology are interactions between transcriptional regulatory proteins and their DNA binding sites. It is becoming increasingly evident that genetic regulatory networks, linked through DNA-protein interactions, establish and maintain patterns of gene expression crucial to processes such as adaptation and cellular differentiation. Novel experimental approaches that allow the simultaneous in vivo quantitation of many DNA-protein interactions are crucial for understanding the global architecture and dynamics of gene regulatory networks.

Interactions between DNA and proteins have been extensively studied and many systems have been developed for their detailed analysis. Although these methods provide detailed information, they require a priori knowledge of the DNA sequence and/or DNA-binding protein(s) involved and are generally limited to analysis of interactions at single loci. Furthermore, many of these techniques such as traditional in vivo footprinting[1] perturb the physiologic state of DNA-protein interactions, as the requisite agents such as dimethyl sufate (DMS) and ultraviolet irradiation are lethal to cells.

We have developed a methylation-based method to display and quantitate the in vivo occupancy of DNA sites. Our approach relies on the efficient methylation of the *Escherichia coli* chromosome by DNA methyltransferases in vivo. Wild-type *E. coli* strains express Dam, a DNA methyltransferase that methylates the N6 position of adenine in GATC sequences. This methylase does not have a cognate endonuclease in *E. coli*, but serves important functions in mismatch repair and chromosomal replication[2]. We and others have observed that 0.1–0.2% of the roughly 20,000 *E. coli* adenine methyltransferase (Dam methylase) targets are found to be undermethylated[3-5]. These undermethylated GATCs are only found at specific chromosomal sites where DNA-bound proteins protect the DNA from enzymatic methylation by sterically inhibiting the access of DNA methyltransferase. These protected sites are therefore footprints for in vivo bound DNA-binding proteins and can be identified through the use of methylation-sensitive endonucleases that cleave such sequences. We originally cloned seven of these sites and showed that they fall within the 5′ noncoding regions of *E. coli* genes[4].

The whole-genome screening method allows us to display, identify, and quantitate a large number of these sites simultaneously. In addition to corroborating independent biochemical and genetic evidence for protein binding at 13 of these sites, we identified 10 new protein-binding sequences upstream of previously uncharacterized genes. The quantitative nature of this approach indicates the growth conditions in which such genes are switched on or off, thereby providing insights into their potential role in cellular physiology.

Results

In a typical experiment (Fig. 1A), *E. coli* strains expressing native and/or foreign DNA methyltransferases were grown under a given condition. Genomic DNA was then isolated, purified, and digested with a methylation-sensitive endonuclease, yielding fragments whose ends were protected from methylation in vivo. A radioactively labeled oligonucleotide linker was then ligated to the cohesive ends of the genomic fragments. To resolve these fragments on a gel, the linker-ligated fragments were then cut with a second enzyme, chosen to cut frequently and give fragments in the range of 100 to 1000 bp. The resulting fragments were electrophoresed through a 6% nondenaturing polyacrylamide gel and the pattern of methylase protection was then visualized on x-ray film or Phosphorimager scan. The intensity of a given band reflected the strength of in vivo protection and correlated with the extent of protein occupancy at that site. In this fashion, it was possible to obtain a protection profile revealing the extent of protein occupancy at numerous sites simultaneously.

Once the protection pattern was visualized, a specific protected site could be identified by excision of the corresponding band from the polyacrylamide gel, ligation of a second linker, followed by PCR and sequencing (Fig. 1B). An excision template was generated using the autoradiogram. Fragments of interest were cut out of the gel and crushed, and the DNA was eluted into a high-salt buffer. A Y-shaped, partially double-stranded linker was then ligated to the fragments. The resulting fragments constituted two major populations: (1) the methylase-protected subset that has the primary linker at the 5′ end and the secondary linker at the 3′ end, and (2) a subset, resulting from digestion with the second endonuclease, that has the secondary linker at both ends. The Y-shaped linker allowed for specific amplification of the methylase-protected fragments[6]. This linker had a 3′ overhang used to ligate to fragments generated by the second endonuclease, an 11 bp middle region of complementarity, and a 5′ end that contains a 16 bp region of noncomplementarity, giving rise to its schematized Y-shape. These fragments were then used as templates for PCR. The four potential templates were labeled 1–4 in figure 1B. In the first round, the 5′ primer (P1) annealed to the 3′ end of template 2 and extended DNA synthesis through the fragment, into the secondary linker, synthesizing complementary sequence, to which the 3′ primer (P2) can anneal to. Template 2 became exponentially amplified in subsequent rounds. Because templates 1, 3, and 4 have no sites for primer annealing, they were not amplified. The PCR product was then gel-purified and cycle-sequenced.

Dam methylase protection. We performed a methylase-protection experiment using the native *E. coli* methylase Dam (Fig. 2). Two strains of *E. coli* were used: CA 8000 and CA 8445-1 (ref. 7). CA 8445-1 contains a null *crp* allele (*crp*-) and is otherwise isogenic with CA 8000 (*crp*+). The *crp* gene encodes the cAMP receptor protein (Crp, also known as catabolite activator protein). This global regulator binds cAMP and regulates the expression of many genes by either activating or repressing transcription[8]. Genomic DNA was harvested from *crp*- and *crp*+ strains grown in batch culture to either log phase (optical density $[\mathrm{OD}]_{600}$ of 0.57 for *crp*- and 0.60 for *crp*+) or stationary phase (OD_{600} of 1.4 for *crp*- and 1.5 for *crp*+). The methylation-sensitive endonuclease DpnII (cuts at non-methylated GATCs) was used to identify the in vivo protected sites. Ethidium bromide–stained agarose gels of DpnII digested genomic DNA showed indistinguishable overall methylation levels for the four growth phase/strain combinations. The specific pattern of methylase protection differed both with respect to growth phase and strain (Fig. 2). A number of these protected sites was identified and each was labeled with the downstream gene/open reading frame (ORF) that it presumably regulates. All labeled sites were sequenced in all four conditions. The most striking differences

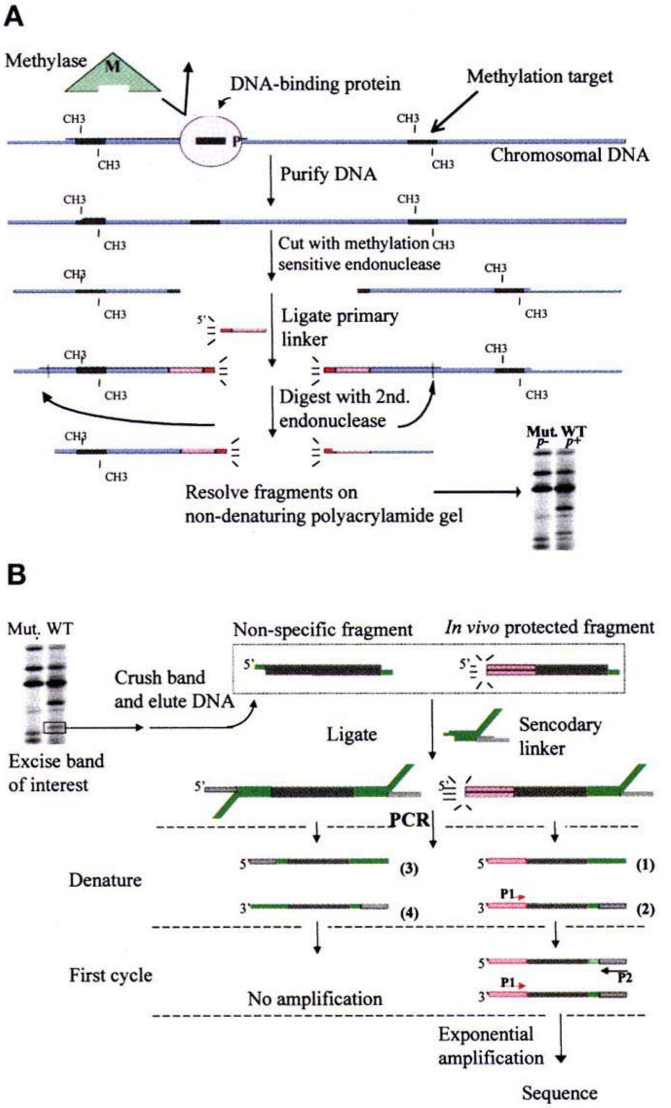

Figure 1. Protocol for displaying and identifying in vivo protected sites. (A) After the genomic DNA is purified and digested with a methylation-sensitive endonuclease, a radioactive linker is ligated to the ends of genomic fragments. The fragments are digested with a second nonmethylation-sensitive enzyme and separated through a 6% polyacrylamide gel. The pattern of methylase protection is then visualized on x-ray film or Phosphorimager scan. (B) Fragments of interest are excised from the gel, and the DNA is eluted into a high-salt buffer. A Y-shaped, partially double-stranded linker is then ligated to the ends of fragments. PCR amplification of the ligation product using primers P1 and P2 only amplifies the methylase-protected subset of fragments (the ones that contain the primary linker at one end and the Y-shaped linker at the other). The PCR product is then cycle sequenced.

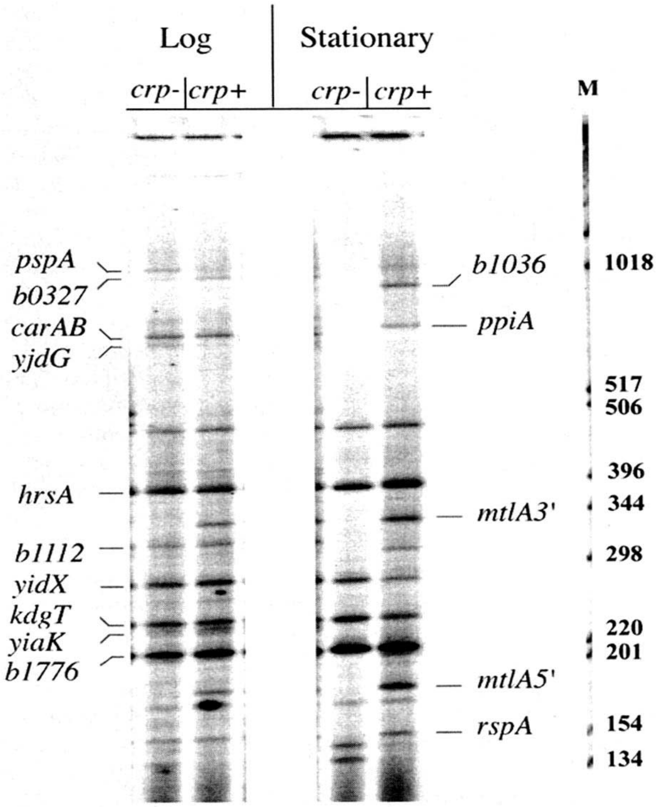

Figure 2. Methylase-protection display of isogenic *crp*- and *crp*+ strains of *E. coli* grown in batch culture to mid log, and stationary phase. A subset of these protected fragments was identified and labeled according to downstream gene/ORF name.

were seen among Crp-dependent methylase protections. For example, protections upstream of *mtlA* were seen only in the *crp+* strains in both log and stationary phase. Nonetheless, variation of growth phase also altered protection profiles. For instance, protection upstream of *carAB* and *b1112* was only transiently present during log phase. The pattern of protection across the majority of sites seems to be a more complicated function of genetic background and growth phase. For example, the protection upstream of *rspA* was present equally in *crp-* and *crp+* cells during log phase, but was observed exclusively in the *crp+* strain during stationary phase.

All 23 of the identified Dam-protected sites fell within the 5′ noncoding region of *E. coli* genes and putative ORFs (Table 1). Their downstream genes/ORFs, the corresponding map position, and the 15 bases flanking the protected sequences have been determined. Several independent lines of evidence support protein binding at these protected sites including regulation by the DNA-binding protein on physiologic basis, protein binding on the basis of DNA-binding consensus sequence, and DNase I footprinting. The DNA-binding proteins belong to diverse families of transcriptional regulatory proteins with distinct sequence specificities and modes of regulation (Table 1). In the case of the Crp, the overrepresentation is a result of the partial overlap between the methylation target sequence (GATC) and Crp half-site binding consensus (TCACA).

Table 1. Characterization of protected sites.

Gene	Map	Sequence context	Bound protein	Evidence	Reference
carAB	1	5′CAGAACAGGTTAGAT**GATC**TTTTTGTCGCTTAAT	CarP	P/D	20
gcd	3	TCGAATTGTGATGAC**GATC**ACACATGTTAAAACC	Crp	P/S	24
b0327	7	CGGTGTATAAAAAAT**GATC**TCATGCAGATGTTTT			
fep	13	AACATATCCAAATAA**GATC**GATAACGATAATTAA	Fur	P/S	23
hrsA	16	CAATCAAGTGAAATT**GATC**ACATAATGCTATTGT	Fnr	S	This study
b1036	23	CTGAAGAATAAGTGT**GATC**TACCTCACTCATAAC	Crp	S	This study
b1112	25	TATCATTTAGTTATC**GATC**GTTAAGTAATTGCTT			
pspA	29	TTGATTCTTCAATCA**GATC**TTTATAAATCAAAAA	IHF	S/D	29
rspA	36	ACTCCGGCTTTTTTC**GATC**TTTATACTTGTATGG			
b1776	40	TAAAACACAGATAAT**GATC**TGCGTTTTACAACTC			
flhD	42	ATAATGCGTGATGCA**GATC**ACACAAAACACTCAA	Crp	P/S	25
cdd	48	AATTAATGAGATTCA**GATC**ACATATAAAGCCACA	Crp	P/S/D	22
yffE	55	TACCTCACTTCTCGT**GATC**AAGATCACATTCTCG	Crp	S	This study
gut	61	TATCTTTCATTTTGC**GATC**AAAATAACACTTTTA	Crp	P/S	27
yjdG	66	TTTATAGATTAATCT**GATC**TACCCATTTGTGGGT	IHF	S	This study
ppiA	75	TTAAGAGGTGATTTT**GATC**ACGGAATAAAAAAGT	Crp	P/S/D	16
yhiP	78	TTCCATCATTAGTGT**GATC**ATCTGGTTATTTTCT			
mtlA	81	ATATCTTGTGATTCA**GATC**ACAAATATTCAACAA	Crp	P/S/D	17
yiaK	81	GCCTTGTTAAAAAGT**GATC**GATATATTTGAAATC			
yidX	83	TTGTACTACAATTTA**GATC**ACAAAAAGAACAATG			
kdgT	88	AATTGATGTGGTTTT**GATC**ACTTTTATTGATTAA	Fnr	S	This study
*yihU/V**	88	TGCTTGTCTGTTTTT**GATC**GTATTTGTAATTTAT			
proP	93	TCCATGTGTGAAGTT**GATC**ACAAATTTAAACACT	Crp	P/S	28

Evidence for protein binding was based on: physiological data (P), DNA-binding consensus sequence (S), and Dnase I footrpinting (D). Established consensus binding motifs are underlined. *In the case of *yihU/V*, the protected site fell within the intergenic region between the divergent coding regions *yihU* and *yihV* and therefore could not be assigned to either gene.

SssI methylase protection. We also carried out an initial study to see whether sites protected from methylation by the SssI DNA methyltransferase (M.SssI) fall within known or suspected protein-binding sequences. M.SssI, isolated from *Spiroplasma* sp. strain MQ1 (ref. 9), methylates the cytosine in 5′CG, increasing the number of methylation targets by 16-fold over Dam (CG versus GATC). By using methylation-sensitive endonucleases that contain 5′CG within their recognition sequence (i.e., Aci I, Hinp I, HpaII, and MaeII), one can, in theory, assay the methylation status of a substantially greater number of sites than with Dam alone.

Expression of M.SssI in *E. coli* increases the rate of C to T mutations, but has no other obvious deleterious effects[10]. We were able to get almost complete methylation of the *E. coli* chromosomal DNA by expressing M.SssI from pMSI, a high copy plasmid with a colE1 origin of replication. We used *E. coli* strain ER1821 as host because it lacks McrA, McrBC, and Mrr activity that restricts DNA modified by the SssI DNA methyltransferase[11]. After purifying and digesting the in vivo methylated DNA with HpaII (HpaII recognizes 5′CCGG and cuts if the internal cytosine is unmethylated), the fragments were displayed and identified as in the Dam methylane protection experiment, except that Sau3A I was used as the second endonuclease. Of the 10 most prominent bands, the majority of the sequences (8/10) were found to match sequences on the high-copy expression plasmid pMSI. This was expected because compared with the single copy chromosome, the plasmids spend more time in the hemi- and unmethylated state throughout a generation time. Of the two sequences that matched *E. coli* chromosomal DNA, one fell within the putative promoter of the cryptic *ascFB* operon where the repressor *ascG* is thought to bind[12] (Fig. 3). The second chromosomal site fell 17 bp downstream of the start codon for *galE*, the first gene of the *galETK* operon. The protected cytosine that we identified within *galE* is 1 bp away from a known operator element within the structural gene[13].

Quantitative analysis of protection patterns. This approach allows for a quantitative comparison of protection profiles across various growth conditions via densitometric measurements of band intensities from Phosphorimager exposures of polyacrylamide gels. We used the percent-protection from methylation for one of these sites, *mtlA* (determined by MboI digestion and Southern blot analysis) as an internal calibration for determining the percent-protection of other sites. We quantitated the extent of methylase-protection (Fig. 4) for the sites identified in the comparison of *crp-* and *crp+* strains grown in log and stationary phase (Fig. 2).

This parallel nature of the approach allowed us to inquire about the more global organization of protection mechanisms. To this end, we applied a mathematical approach to cluster protected sites into protection groups. The members of such protection groups likely interact with the same DNA-binding protein. This approach provides insights into the function of novel genes by grouping them with genes of known function based on a common pattern of transcriptional regulation. A similar approach was used to group together cancer chemotherapy drugs and their targets based on activity and expression patterns[14].

We used a multivariate statistical approach to construct a matrix of correlations between protected sites[15]. Our dataset consisted of observation vectors, whose elements were the protection values of each upstream region in a particular experimental condition. In the Dam methylase-protection experiment (Fig. 2), the four observation vectors form the columns of a 4 × 15 observation matrix (15 protected sites in the four different growth/strain combinations).

The dynamic range of percent-protection under the four different conditions was highly variable for the 15 genes (Fig. 4). The percent protection of *yjdG* ranged from 0–3%, and for *b1776* it ranged from 16–32%. This variability most likely reflects differences in binding affinity for the different DNA-binding proteins and their targets. To make the analysis insensitive to the absolute level of protection, and sensitive to the degree of change, we trans-

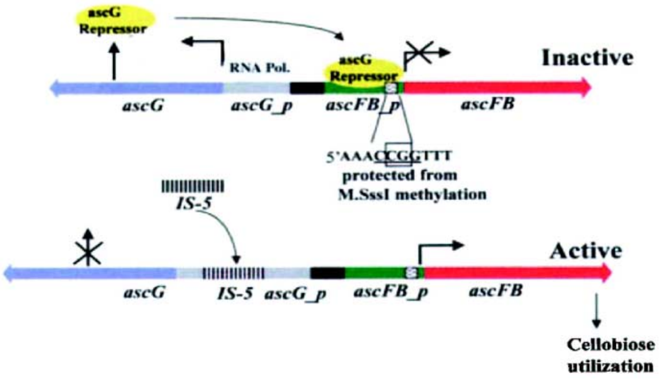

Figure 3. M.SssI protected site identified within the cryptic *ascFB* operon. The protected CG fell within the promoter of *ascFB* where the ascG repressor is thought to bind[12]. Normally, the product of the divergently transcribed *ascG* represses trasncription from the *ascFB* promoter. Upon the insertion of an IS element within the promoter of *ascG*, this gene is inactivated. In the abscence of the ascG repressor, the transcription of the *ascFB* operon proceeds leading to the utilization of sugars such as cellobiose.

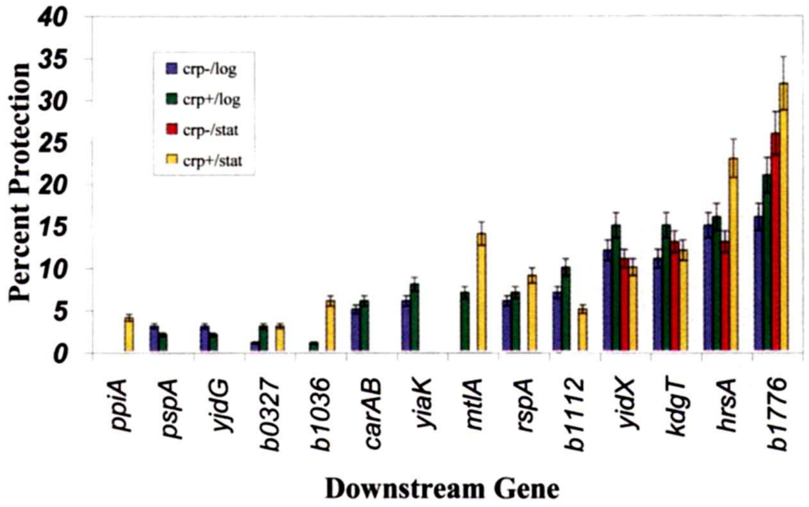

Figure 4. Densitometric measurements of percent methylation-protection. The level of methylase-protection is quantitated for the sites identified in the isogenic *crp-* and *crp+* strains of *E. coli* grown to log or stationary phase.

formed the observation matrix such that each site's protection had the same variance across the four different conditions. The transformation was achieved by subtracting the protection level of each site from the mean protection level of that site and dividing by the square root of the variance. Thus, the elements of the transformed observation matrix have unit variance and zero mean. If X_{ij} is an element of the original n x m observation matrix X, the transformed matrix Y, has elements Y_{ij}:

$$Y_{ij} = \frac{X_{ij} - \langle X_i \rangle}{\left[\frac{1}{n}\sum_{j=1}^{n}\left(X_{ij} - \langle X_i \rangle\right)^2\right]^{1/2}}$$

where

$$\langle X_i \rangle = \frac{1}{n}\sum_{j=1}^{n} X_{ij}$$

The next step was to use the unit-variance observation matrix Y to construct a correlation matrix C:

$$C = \frac{1}{n} Y.Y^T$$

Y^T is the transpose of matrix Y. The elements C_{ij} of C, an $m \times m$ matrix, give a measure of correlation of protection between the ith and jth protected sites. The correlation coefficient C_{ij} can range from -1 to +1, where positive values denote correlated and negative values anticorrelated protections. If C_{ij} is a large positive value, then protections of the sites corresponding to i and j tend to correlate with each other across the conditions that were used to construct the correlation matrix. Conversely, if C_{ij} is a large negative value then the protections of these sites tend to be anticorrelated with respect to each other. For example, protected sites that bind the same regulatory protein are expected to have a large positive C_{ij}.

We constructed a color-coded correlation matrix (correlogram) for the protection data (Fig. 5) showing positive and negative correlations. By using a hierarchical clustering algorithm[15], we permuted the rows of the observation matrix in order to group together highly correlated protections into clusters. In this way, "protection groups" emerged as square blocks of highly correlated protected sites along the diagonal of the matrix, surrounded by anticorrelated or slightly correlated sites away from the diagonal. The diagonal of the matrix consists of bright red elements because, by definition, each protection is fully correlated with itself.

The protected sites cluster into five groups. These groups represent protections that change in the same manner across the four different condition/strain combinations (Fig. 5). For example, the members of a group consisting of *b1112, YiaK, carAB*, and *b0327* are only protected during the log phase of growth (Fig. 2). The

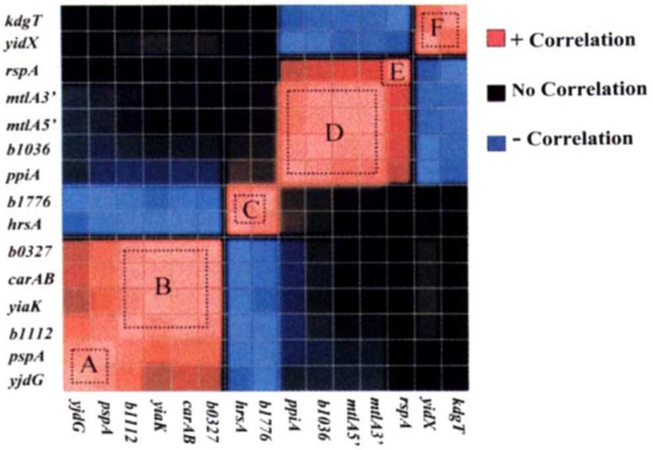

Figure 5. Correlogram. Color-coded, clustered correlation matrix obtained for the protected sites identified in the experiment shown in figure 2. Red: positive correlations; blue; negative correlations. The saturation of each color represents the strength of the correlation. The clustered blocks of red (A through E) form protection groups whose members are protected in the same manner across the different conditions. The off-diagonal blue elements represent protections that are anticorrelated with respect to each other. The dark elements are correlations that are small in absolute value.

members of a group, consisting of *ppiA*, *b1036*, *mtlA5′*, and *mtlA3′* form a cluster as their protection is dependent on Crp. This is consistent with the role of Crp in the regulation of *ppiA*[16] and *mtlA*[17], and the strong match of the protected site upstream of *b1036* to the Crp binding consensus (Table 1). To further determine the role of Crp in the protection upstream of this group of genes, we compared the protection patterns of *crp+* cells grown with and without glucose in Luria-Bertani medium. Through catabolite repression[8], the protection should decrease in cells grown in the presence of glucose. Indeed, comparison of otherwise identical cultures of *crp+* cells showed that the protection was only observed in cells that were grown without supplemental glucose. The protection of *ppiA* and *b1036* are presumably the most strongly correlated within their group because they require stationary phase in addition to Crp for protection. This may reflect the interaction of a stationary-phase specific element with Crp as a requirement for optimal binding to the sites upstream of *ppiA* and *b1036*.

Discussion

Uncovering the mechanisms by which cells establish and maintain specific patterns of gene expression is central to understanding fundamental processes such as cellular differentiation and adaptation. The global monitoring of DNA-protein interactions allows a systems level approach to understanding the connectivity and dynamics of gene regulatory networks that underlie the diverse repertoire of gene expression patterns.

We described a whole-genome, in vivo approach for identifying, displaying, and quantitating many DNA-protein interactions simultaneously. This approach identifies in vivo DNA-protein interactions with minimal perturbations to physiology; the screening methodology obviates a priori knowledge of binding sites and/or DNA-binding proteins; and the technique allows the simultaneous comparison of protein occupancy at numerous loci simultaneously. Of the 25 protected sites studied, 24 were found within 5′ noncoding regions, and in the case of *galE*, the only exception, the M.SssI methylase protected site lies 1 bp outside of a previously identified operator element, which was the first such motif described within the coding region of a gene[13]. The small percentage of noncoding sequences in *E. coli*[18] (11%), and the skewed abundance of GATCs within coding regions[19] makes the probability of observing this distribution of protected sites less than one in 10^{27}.

The protected sites show even more specific bias within noncoding regions. Five of the protected sites (*carAB, pspA, cdd, ppiA,* and *mtlA)* were found to lie within independently determined protein binding sites by DNaseI footprinting[16,17,20–22]. Another eight of the sites (*fep, gcd, flhD, yffE, gut, proP, ascFB,* and *galE*) agree with independent evidence for protein binding on the basis of genetic and sequence data[12,13,23–28]. The remaining 12 sites, most of which fall upstream of ORFs of unknown function, have not been previously studied and should provide interesting subjects for subsequent studies of DNA-protein interactions and transcriptional regulation. In the case of the protected sites upstream of *b1036, b0327, yjdG, b1112, yiaK,* and *rpsA*, our dataset suggests possible physiologic roles for these genes. For example, protections upstream of *b0327, yjdG, b1112,* and *yiaK* are observed exclusively during logarithmic growth phase and are probably involved in the activation or repression of these genes during log phase growth.

The protection upstream of *pspA* and the unknown gene *yjdG* are highly correlated (Fig. 5). The *pspA* gene codes for the phage-shock protein A, which is synthesized in response to a variety of stressful conditions such as heat-shock, osmotic shock, and ethanol treatment[29]. Dnase I footprinting has shown that the integration host factor (IHF) binds to the sequence we found protected upstream of *pspA*[29]. Furthermore, the protected sequence upstream of *pspA* closely matches the IHF DNA-binding consensus[30]. It is tempting to postulate that the high level of protection-correlation between *pspA* and *yjdG* indicates regulation of *yjdG* by IHF. Indeed, the protected site upstream of *yjdG* also matches the IHF DNA-binding consensus (Table 1).

The increased target resolution possible with foreign methyltransferases, together with novel approaches to display protected sites such as probing high-density DNA arrays (DNA-chips) will allow the identification and quantitation of many hundreds of DNA-protein interactions simultaneously. The mathematical/visualization approaches presented here can also be applied to the massive amounts of raw data stemming from other genome-wide techniques such as whole-genome transcriptional profiling[31,32]. The methylase-protection approach should also be applicable to studying chromatin structure in eukaryotic genomes. The level of methylation at specific loci of *Saccharomyces cerevisiae* depends on the transcriptional state and chromosomal location of the gene, as telomeric and transcriptionally repressed loci seem to be refractory to Dam modification[33]. The recent development of *S. cerevisiae* gene expression oligonucleotide arrays[31] together with available strains of *S. cerevisiae* expressing the Dam and SssI methyltransferases should allow the global, high-resolution study of chromatin structure and its influence on the expression of all the genes in a eukaryotic organism.

Experimental protocol

Enzymes, buffers, and media. All restriction enzymes, T4 DNA ligase, and T4 polynucleotide kinase and their corresponding buffers were purchased from the New England Biolabs (NEB; Beverly, MA). AMV Reverse Transcriptase and Taq DNA polymerase were purchased from Boehringer Mannheim (Indianapolis, IN). LB medium is 1% bactotryptone, 0.5% yeast extract, and 1% NaCl.

Bacterial strains and plasmids. Three strains of *E. coli* were used: *crp*+ (CA 8000/CGSC#6026: *thi-1*, *rel A1*, *spo T1*, λ-) and *crp*- (CA 8445-1/CGSC#7043: *thi-1*, *rel A1*, *spo T1*, λ-, *rpsL 136*, Δ*crp-45*) were kindly provided by B. Bachmann of the Yale University (New Haven, CT) *E. coli* Genetic Stock Center. ER 1821 [F- e14- (McrA-) *endA1 supE44 thi-1 relA1? rfbD1? spoT1?* Δ(*mcrC-mrr)114::IS10*], a derivative of MM294 was obtained from NEB. Plasmids pUC19 and pCAL7 (harboring the M.SssI gene) were also obtained from NEB.

Display of in vivo protected sites. Genomic DNA was extracted from *E. coli* cells expressing native or foreign DNA methyltransferases grown in batch cultures[34]. This DNA was isolated, purified, and digested with the appropriate methylation-sensitive endonuclease (5 µg genomic DNA, 20 units of DpnII in a final volume of 20 µl for 2 h) leaving a 5′ overhang. A radioactively labeled linker with a cohesive end was then constructed by annealing two partially complementary oligos (DPNL: 5′-CTTTTTTTTTTTTCGTTCGAGCT-CACGTAGATGTC and DPNS: 5′ P-GATCGACATCTACGTGAGCTC-GAACG) and primer-extending with [α^{32}P] dATP's using AMV Reverse Transcriptase. The poly-T tail of DPNL acts as template for polymerization of [α^{32}P] dATP by AMV Reverse Transcriptase (3 pmol of each oligonucleotide; 3 µl of 5X AMV Reverse Transcriptase buffer; 5 µl of [α^{32}P] dATP [6000 Ci/mmoles]; 1 µl of AMV Reverse Transcriptase [25 U/µl] in a volume of 15 µl was incubated at 37°C for 20 min). The reaction was then chased with 1 µl of cold 100 mM dATP for 10 min at 37°C and subsequently inactivated by incubation at 70°C for 15 min. The resulting labeled linker contains multiple ^{32}P-labeled nucleotides at one end, a middle 21 nucleotide region used as a primer site for subsequent PCR amplification, and a 5′ overhang at the other end used for specific ligation to fragments generated by the methylation-sensitive endonuclease. The labeled linkers are ligated to the ends of the genomic fragments. One microgram of DpnII digested *E. coli* chromosomal DNA, 1 µl (0.2 pmol) of the radioactive linker, 1.5 µl of 10X ligase buffer, and 0.3 µl of T4 DNA ligase (2000 U/µl) in a total volume of 15 µl were incubated for 2 h at 16°C and subsequently heat inactivated at 65°C for 20 min. To be able to resolve these fragments on a gel, the linker-ligated fragments were then cut with a four-base recognition endonuclease (0.5 µl of NEB #2 buffer, 0.5 µl of MspI endonuclease [100 U/µl], and 5 µl of dH_2O added to the ligation mixture above and incubated at 37°C for 1.5 h). After heat-inactivation of the digestion product at 65°C for 20 min and addition of 5 µl of 6X loading dye, 12 µl were loaded onto a 6% nondenaturing polyacrylamide gel (360 × 260 ×

1 mm). The samples were electrophoresed at 4.4 V/cm for 12–14 h. The gel was then used to expose a Phosphorimager screen for 5 h and then scanned to quantitate the intensity of protection bands. These intensities were quantified using a Molecular Dynamics (Sunnyvale, CA) Phosphorimager scanner and ImageQuant version 1.1 software (Molecular Dynamics). The gel was also used to expose an x-ray film for 24 h. This was done to create a template for excising fragments of interest from the gel.

Identifying protected sites of interest. Once the protection pattern was visualized, a specific protected site was identified by excision of the corresponding band from the polyacrylamide gel using the x-ray autoradiogram template. Fragments of interest are cut out of the gel, crushed, and the DNA is eluted into a high-salt buffer (0.5 M ammonium acetate, 1 mM EDTA pH 8.0). The DNA is ethanol-precipitated and resuspended in 15 µl of dH_2O. A Y-shaped, partially double-stranded linker (MSPY) is constructed by annealing two oligos (MSPY5: 5′-ACTACGCACCGGACGAGACGTAGCGTC, and MSPY3: 5′ P-CGGACGCTACGTCCGTGTTGTCGGTCCTG). The Y-shaped linker allows for specific amplification of the methylase protected fragments[6]. At one end, the linker has a 5′CG overhang complementary to the overhang generated by the second endonuclease, an 11 bp double-stranded middle region and a 16 bp region of noncomplementarity at the other end, giving rise to its schematized Y-shape. This linker was ligated to the cohesive ends generated by the MspI endonuclease (2 pmol of the MSPY linker, 5 µl [10 ng] of the gel-isolated DNA, 1 µl of 10X T4 DNA ligase buffer, 0.2 µl T4 DNA ligase [2000 U/µl] in a volume of 10 µl incubated at 16°C for 12 h). The ligation mixture was heat-inactivated and 1 µl was used as template in a PCR reaction containing primers P1 (5′-CGTTCGAGCTCACGTAGATGTC) and P2 (5′-ACTACGCACCGGACGAGACGT) using Taq polymerase. The PCR reactions reproducibly generated fragments of the expected size, which were agarose gel-purified and cycle-sequenced using primers P1 and P2 on an ABI automated sequencer. For the majority of sites, only one of the two endonuclease-generated fragments was identified. This was largely due to the small predicted size of the second digestion fragment, which in most cases would be obscured by the high-intensity signal from the radioactively labeled linker-dimers at the bottom of the gel. In one notable case (*mtlA*), the 3′ generated fragment migrated at appoximately 340 bp, which was different from that expected from sequence data (260 bp). This may reflect decreased mobility of the 3′ fragment in the native polyacrylamide gel due to the predicted bend of 85.32 degrees across 70% of this fragment (GenBank: U15409|ECU15409).

Constructing the pMSI expression plasmid. Primers MS_5 (5′-ACAAAAGCTGCTCGCCTGCAGGATGAGCAAAGTAGAAAATAAAACAAA) and MS_3 (5′-GCCCTATATACCGGTACCTTAACCTCCAATTTTATCTATAATCGCTTC), containing restriction sites for PstI and KpnI respectively, were used to PCR amplify the M.SssI gene from pCAL7. The PCR product was digested with PstI/KpnI and ligated into the pUC19 (a plasmid with a colE1 origin of replication) multiple cloning site, downstream of the Lac promoter. *E. coli* strain ER1821 was transformed with the ligation mixture and colonies expressing M.SssI were isolated by assaying the cytosine methylation level of the chromosomal DNA using HpaII digestions. Plasmid pMSI, bearing a functional M.SssI gene was then isolated and used to transform ER1821 strains for subsequent methylation-protection experiments.

Acknowledgments

We thank Rich Chin for his important contributions to this work and Pete Estep, Dereth Phillips, Fritz Roth, Benjamin Gewurz, Vamsi Mootha, and Sohail Tavazoie for many enlightening discussions and critical readings of the manuscript. This work was supported by DOE grant DE-FG02-87ER60565.

1. Cartwright, I.L. and Kelly, S.E. 1991. Probing the nature of chromosomal DNA-protein contacts by in vivo footprinting. *Biotechniques* **11**:188–203.
2. Barras, F. and Marinus, M.G. 1989. The Great GATC: DNA methylation in *E. coli*. *Trends Genet.* **5**:139–143.
3. Rinquist, S. and Smith, C.L. 1992. The *Escherichia coli* chromosome contains specific, unmethylated *dam* and *dcm* sites. *Proc. Natl. Acad. Sci. USA* **89**:4539–4543.
4. Wang, M.X. and Church, G.M. 1992. A whole genome approach to in vivo DNA-protein interactions in *E. coli*. *Nature* **360**:606–610.
5. Hale, W.B., van der Woude M.W., and Low D.A. 1994. Analysis of Nonmethylated GATC sites in the *Escherichia coli* chromosome and identification of sites that are differentially methylated in response to environmental stimuli. *J. Bact.* **176**:3438–3441.
6. Prashar, Y. and Weissman S.M. 1996. Analysis of differential gene expression by display of 3′ end restriction fragments of cDNAs. *Proc. Natl. Acad. Sci. USA* **93**:659–663.
7. Sabourin, D. and Beckwith, J. 1975. Deletion of *Escherichia coli* crp gene. *J. Bacteriol.* **122**:338–340.
8. Kolb, A., Busby, X., Buc, H., Garges, S., and Adhya, S. 1993. Transcriptional regulation by cAMP and its receptor protein. *Annu. Rev. Biochem.* **62**:749–795.
9. Renbaum, P., Abrahamove, D., Fainsod, A., Wilson, G.G., Rottem, S., and Razin, A. 1990. Cloning, characterization, and expression in *Escherichia coli* of the gene coding for the CpG DNA methylase from *Spiroplasma sp.* strain MQ1(M.SssI). *Nucleic Acids Res.* **18**:1145–1152.
10. Zhang, X. and Mathews, C.K. 1994. Effect of DNA cytosine methylation upon deamination-induced mutagenesis in a natural target sequence in duplex DNA. *J. Biol. Chem.* **269**:7066–7069.
11. Kelleher, J. and Raleigh, E.A. 1991. A novel activity in Escherichia coli K-12 that directs restriction of DNA modified at CG dinucleotides. *J. Bacteriol.* **173**:5220–5223.
12. Hall, B.G. and Xu, L. 1992. Nucleotide sequence, function, activation, and evolution of the cryptic *asc* operon of *Escherichia coli* K12. *Mol. Biol. Evol.* **9**:688–706.
13. Irani, M.H., Orosz, L.M., and Adhya, S. 1983. A control element within a structural gene: The *gal* operon of *Escherichia coli*. *Cell* **32**:783–788.
14. Weinstein, J.N., Myers, T.G., O'Connor, P.M., Friend, S.H., Fornace, A.J. Jr., Kohn K.W. et al. 1997. An information-intensive approach to the molecular pharmacology of cancer. *Science* **275**:343–349.
15. Kendall, M. 1980. *Multivariate analysis*. Macmillan, New York.
16. Norregaard-Madsen, M., Mygind, B., Pedersen, R., Valentin-Hansen, P., and Sogaard-Anderson, L. 1994. The gene encoding the periplasmic cyclophilin homologue, PPIase A, in *Escherichia coli*, is expressed from four promoters, three of which are activated by the cAMP-CRP complex and negatively regulated by the CytR repressor. *Mol. Microbiol.* **14**:989–997.
17. Ramseier, T. M. and Saier, M.H. Jr. 1995. cAMP-cAMP receptor protein complex: five binding sites in the control region of the *Escherichia coli* mannitol operon. *Microbiology* **141**:1901–1907.
18. Blattner, F.R., Plunkett, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M. et al. 1997. The complete sequence of *Eschericia coli* K-12. *Science* **277**:1453–1462.
19. Barras, F. and Marinus, M.G. 1988. Arrangement of Dam methylation sites (GATC) in the *Escherichia coli* chromosome. *Nucleic Acids Res.* **16**:9821–9838.
20. Charlier, D., Gigot, D., Huysveld, N., Roovers, M., Pierard, A., and Glansdorff, N. 1995. Pyrimidine regulation of the *Escherichia coli* and *Salmonella typhimurium carAB* operon: CarP and Intergration Host Factor (IHF) modulate the methylation status of a GATC site present in the control region. *J. Mol. Biol.* **250**:383–391.
21. Jovanovic, G., Weiner, L., and Model, P. 1996. Identification, nucleotide sequence, and characterization of PspF, the transcriptional activator of the *Escherichia coli* stress-induced *psp* operon. *J. Bacteriol.* **178**:1936–1945.
22. Holst, B., Sogaard-Anderson, L., Pedersen, H., and Valentin-Hansen, P. 1992. The cAMP-CRP/CytR nucleoprotein complex in *Escherichia coli*: two pairs of closely linked binding sites for the cAMP-CRP activator complex are involved in combinatorial regulation of the cdd promoter. *EMBO J.* **11**:3635–3643.
23. Shea, C.M. and McIntosh, M.A. 1991. Nucleotide sequence and genetic organization of the ferric enterobactin transport system: homology to other periplasmic binding protein-dependent systems in *Escherichia coli*. *Mol. Microbiol.* **5**:1415–1428.
24. Yamada, M., Asaoka, S., Saier, M.H. Jr., and Yamada, Y. 1993. Characterization of the *gcd* gene from *Escherichia coli* K-12 W3110 and regulation of its expression. *J. Bacteriol.* **175**:568–571.
25. Kutsukake, K., Ohya, Y., and Iino, T. 1989. Transcriptional analysis of the flagellar regulon of *Salmonella typhimurium*. *J. Bacteriol.* **177**:741–747.
26. Andrews, S.C., Harrison, P.M., and Guest, J.R. 1991. A molecular analysis of the 53.4 minute of the Escherichia coli linkage map. *J. Gen. Microbiol.* **137**:361–367.
27. Yamada, M., Saier, M.H. Jr., and Yamada, Y. 1988. Positive and negative regulators for glucitol (gut) operon expression in *Escherichia coli*. *J. Mol. Biol.* **203**:569–583.
28. Xu, J. and Johnson, R.C. 1995. Fis activates the RpoS-dependent stationary-phase expression of *proP* in *Escherichia coli*. *J. Bacteriol.* **177**:5222–5231.
29. Weiner, L., Brissette, J.L., Ramani, N., and Model, P. 1995. Analysis of the proteins and *cis*-acting elements regulating the stress-induced phage shock protein operon. *Nucleic Acid Res.* **23**:2030–2036.
30. Goodrich, J. A., Schwartz, M.L., and McClure W.R. 1990. Searching for and predicting the activity of sites for DNA binding proteins: compilation and analysis of the binding sites for Escherichia coli integration host factor (IHF). *Nucleic Acids Res.* **18**:4993–5000.
31. Lockhart, D.J., Dong, H.L., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S. et al. 1996. Expression monitoring by hybridizatiion to high-density oligonucleotide arrays. *Nature Biotechnology* **14**:1675–1680.
32. DeRisi, J.L, Iyer, V.R., and Brown, P.O. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**:680–686.
33. Gottshcling, D.E. 1992. Telomere-proximal DNA in Saccharomyces cerevisiae is refractory to methyltransferase activity in vivo. *Proc. Natl. Acad. Sci. USA* **89**:4062–4065.
34. Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.G., Smith, J.A. et al. 1989. Current protocols in molecular biology. John Wiley and Sons, New York.