A Motif Co-Occurrence Approach for Genome-Wide Prediction of Transcription-Factor-Binding Sites in *Escherichia coli*

Martha L. Bulyk,[1,2,3,4] Abigail M. McGuire,[1,2,3] Nobuhisa Masuda,[2] and George M. Church[1,2,5]

[1]*Harvard University Graduate Biophysics Program, Harvard Medical School, Boston, Massachusetts 02115, USA;*
[2]*Harvard Medical School Department of Genetics, Boston, Massachusetts 02115, USA*

Various computational approaches have been developed for predicting *cis*-regulatory DNA elements in prokaryotic genomes. We describe a novel method for predicting transcription-factor-binding sites in *Escherichia coli*. Our method takes advantage of the principle that transcription factors frequently coregulate gene expression, but without requiring prior knowledge of which groups of genes are coregulated. Using position weight matrices for 49 known transcription factors, we examined spacings between pairs of matrix hits. These pairs were assigned probabilities according to the overrepresentation of their separation distance. The functions of many open reading frames (ORFs) downstream from predicted binding sites are unknown, and may correspond to novel regulon members. For five predictions, knockouts with mutated replacements of the predicted binding sites were created in *E. coli* MG1655. Quantitative real-time PCR (RT-PCR) indicates that for each of the knockouts, at least one gene immediately downstream exhibits a statistically significant change in mRNA expression. This approach may be useful in analyzing binding sites in a variety of organisms.

[Supplemental material including detailed methods is available online at www.genome.org and http://arep.med.harvard.edu/ecoli_matrices/spacing/spacing_predictions.html.]

Although the pace of genome sequencing has been growing at an exponential pace, much still remains to be understood about how the genes in the various genomes are regulated. Even in *Escherichia coli*, probably the most well-studied model organism, the complete mechanism of transcriptional regulation of many of its genes is still unknown, despite the fact that the *E. coli* genome contains only ~240 candidate transcription factors regulating ~4300 genes in total (Blattner et al. 1997; Robison et al. 1998). Although many genes in the *E. coli* genome are grouped into separate clusters of coregulated genes (termed regulons), it is likely that many of these regulons are parts of interconnected transcriptional regulatory networks (Neidhardt 1996; Hengge-Aronis 1999). In the yeast *Saccharomyces cerevisiae*, the results of genome-wide binding analysis of 106 transcription factors indicate that more than one-third of the promoter regions that were bound by regulators were bound by two or more regulators, and that there is a highly connected network of transcriptional regulators (T. Lee et al. 2002). Furthermore, many of the genes in the *E. coli* genome are still uncharacterized (termed URFs). If a given URF (uncharacterized open reading frame) is found to be regulated by one or more transcription factors, then it is a reasonable hypothesis that they are members of the same regulon.

The presence of multiple copies of a given transcription factor's binding site motif can be used to predict candidate target genes. For example, a search of the *Drosophila melanogaster* genome for three or more optimal binding sites within a span of 400 bp for the transcription factor Dorsal resulted in the identification of two additional functional regulatory regions containing at least three Dorsal binding sites (Markstein et al. 2002). Similarly, a search of the *Drosophila* genome for overrepresented clusters of binding sites for the transcription factor Suppressor of Hairless [Su(H)] found both known and logical targets of Su(H) binding and regulation (Rebeiz et al. 2002). Another study searched the *Drosophila* genome for overrepresented clusters of binding sites for five different transcription factors important early in anterior–posterior specification in the developing embryo (Berman et al. 2002). Another recent study has searched for clusters of binding sites for muscle-specific transcription factors in the human genome (Frith et al. 2002). In addition, in *E. coli* grammatical models have been used for the identification of regulatory regions (Rosenblueth et al. 1996). Binding site matrices for more than 55 *E. coli* transcription factors are presently available (Robison et al. 1998; Thieffry et al. 1998). These data, also known as position weight matrices (PWMs), can be used to conduct searches of the genome to predict additional candidate binding sites for the particular transcription factor (Robison et al. 1998; Thieffry et al. 1998; Hughes et al. 2000). Although studies have shown that there is some interdependence of the nucleotides of transcription-factor-binding sites (Man and Stormo 2001; Bulyk et al. 2002; M.-L. Lee et al. 2002), a recent study indicates that mononucleotide PWMs are a good approximation for use in identifying high-affinity binding sites (Benos et al. 2002).

The TRANSCompel database on composite regulatory elements in eukaryotic genes provides information on composite elements, containing two closely situated binding sites for distinct transcription factors, within a particular gene and experimental results confirming cooperative action between the transcription factors (Kel-Margoulis et al. 2002b). Similarly, the Transcription Regulatory Regions Database (TRRD) contains

[3]These authors contributed equally to this work.
[4]Present address: Division of Genetics, Brigham and Women's Hospital and Harvard Medical School, Harvard Medical School New Research Building, Boston, MA 02115, USA.
[5]Corresponding author.
E-MAIL church@arep.med.harvard.edu; FAX (617) 432-7266.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.1448004.

information on genes, transcription-factor-binding sites, regulatory regions, locus control regions, and expression patterns (Kolchanov et al. 2002). Likewise, the TRANSFAC database contains information on eukaryotic transcription factors and their known binding sites (Matys et al. 2003).

There are several different algorithms presently in use for finding sequence motifs shared by sets of genes (Bailey and Elkan 1995; Grundy et al. 1996; Roth et al. 1998; van Helden et al. 1998; Hertz and Stormo 1999; Bussemaker et al. 2000; Workman and Stormo 2000; Liu et al. 2001, 2002). A large number of false positives are generated if the search matrix is not specific enough to discriminate true sites. It is more difficult to obtain a specific search matrix in prokaryotes than in eukaryotes. Because of the presence of operons in prokaryotes, there are often fewer instances of each transcription-factor-binding site within the genome. In prokaryotes, each operon would have a single upstream region containing regulatory sites, whereas in eukaryotes, each gene would have its own upstream regulatory region.

Our hypothesis is that many high-scoring false positives can be filtered out by an additional criterion: the condition that most true binding sites co-occur with a second binding site, either for the same transcription factor or a different one. The basis for this assumption is twofold: (1) a transcription factor that regulates a particular ORF often has multiple binding sites in the upstream region either for purposes of binding or for simply increasing the local concentration of that particular transcription factor; and (2) ORFs tend to be coregulated by two or more transcription factors.

There are several different algorithms presently in use for finding pairs of sequence motifs shared by sets of genes. One approach combines a search algorithm for transcription-factor-binding sites with a distance correlation function (Quandt et al. 1996; Frech and Werner 1997; Klingenhoff et al. 1999). Dyad analysis assesses the statistical significance of each possible pair of short oligonucleotides separated by a spacer of fixed length but variable sequence (van Helden et al. 2000; Li et al. 2002). Thus, this approach is well-suited for identifying binding sites for transcription factors that tend to bind as dimers, with a linker domain in the transcription factor separating the DNA binding and dimerization domains. However, this approach at present does not allow for variable spacers between two binding site motifs. This can be important either if the transcription factors have a flexible interaction with DNA, or if the transcription factors do not bind together, but rather simply coregulate highly overlapping sets of genes. A similar approach is aimed at discovering binding site motifs by modeling cooperative binding by transcription factors within a user-specified pattern length (Guha-Thakurta and Stormo 2001). A related effort aimed at identifying so-called structured motifs, composed of two ordered parts separated by a variable distance and allowing for substitution, has been applied to a set of sequences upstream of a subset of *E. coli* and *Bacillus subtilis* genes (Robin et al. 2002).

Likewise, there are several different ways that predicted transcription-factor-binding sites can be tested. One way is to knockout the predicted transcription factor and see if the mRNA levels of the ORF(s) physically downstream from the predicted binding site(s) are either up- or down-regulated. However, several secondary effects may make interpretation of such data difficult (Lee et al. 2002b). A better way to test the predicted binding sites is to mutate the predicted site itself, so that no other genes regulated directly by the transcription factor in question are immediately expected to be up- or down-regulated. Any other genes whose expression is up- or down-regulated in these mutants are then secondary effects caused by a perturbation in the expression of the gene(s) downstream from the predicted binding site.

In this paper, we describe a new approach we have developed to predict sets of two or more transcription-factor-binding sites that coregulate the downstream genes in the *E. coli* genome. We used a database of *E. coli* binding site weight matrices (Robison et al. 1998), obtained by aligning upstream regions identified from other researchers' data, such as biochemical footprinting. AlignACE, a Gibbs sampling strategy (Lawrence et al. 1993) modified for use in identifying DNA sequence motifs (Roth et al. 1998; Hughes et al. 2000), was used to perform the alignments. The Berg and von Hippel algorithm was used to construct the matrices (Berg and von Hippel 1987). The ScanACE program was used to perform the matrix searches (Hughes et al. 2000).

Pairs of candidate binding sites were then assigned probability scores, according to how overrepresented the spacing between the predicted binding sites is. Binding site substitutions were created such that MG1655 genomic DNA contained mutant versions of the predicted binding sites, rather than mere deletions of the predicted binding sites (see Fig. 1 for a summary of the binding site knockouts). Quantitative real-time PCR analysis indicates that at least one of the genes immediately downstream from each of the binding site knockouts exhibits a significant change in mRNA expression.

RESULTS

Binding Site Predictions

All instances of biochemically footprinted DNA-binding sites for 55 different *E. coli* DNA-binding proteins in the literature were assembled into a database previously (Robison et al. 1998). These sites were used to compile matrices representing the nucleotide frequencies of the natural binding sites identified in the footprinting assays. These matrices were used to search the entire *E.*

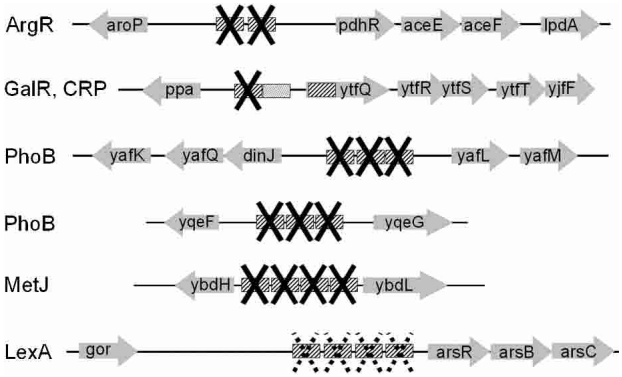

Figure 1 Summary of binding site knockouts. (Solid X) Predicted binding sites that were knocked out. (Dashed X) Knockouts of four predicted LexA-binding sites in the *gorR–arsR* IGR that were not successful, possibly because of lethality. Distances between genes are approximately to scale. In the uppermost construct, two predicted ArgR-binding sites (cross-hatched boxes) were knocked out in the *aroP–pdhR* IGR. In the construct shown below that one, only one of the two predicted GalR-binding sites (cross-hatched boxes) in the *ppa–ytfQ* IGR was knocked out; the predicted CRP site (stippled boxes) was not knocked out. The other constructs created successfully were knockout of three predicted PhoB sites (cross-hatched boxes) in the *dinJ–yafL* IGR; knockout of three predicted PhoB sites (cross-hatched boxes) in the *yqeF–yqeG* IGR; knockout of four predicted MetJ sites (cross-hatched boxes) in the *ybdH–ybdL* IGR. The predicted ArgR-binding sites are 241 bp upstream of *aroP* and 260 bp upstream of *pdhR*; the predicted PhoB sites are 103 bp upstream of *dinJ* and 73 bp upstream of *yafL*; the predicted GalR and CRP sites are 164 bp upstream of *ppa* and 129 bp upstream of *ytfQ*; the predicted MetJ sites are 18 bp upstream of *ybdH* and 58 bp upstream of *ybdL*; the predicted PhoB sites are 37 bp upstream of *yqeF* and 181 bp upstream of *yqeG*; the predicted LexA sites are 328 bp downstream of *gor* and 556 bp upstream of *arsR* (by "upstream," here we mean the distance between the predicted binding sites and the start codon of the gene).

coli genome for potential new binding sites for each of the 55 DNA-binding proteins (Robison et al. 1998). For each pairwise combination of search matrices, we calculated the spacings between each pair of matrix hits. The predicted pairs of sites were ranked according to the probability of obtaining the observed number of pairs of sites separated by that distance or similar distances, given the number expected by chance for that particular distance or range of distances (see Table 1). Pairs of sites with significant spacings are listed in Table 1 and on our Web site (http://arep.med.harvard.edu/ecoli_matrices/spacing/spacing_predictions.html).

For example, it is highly significant that there are eight pairs of ArgR-binding sites separated by 3 bp found in the genome by our search matrices. Seven of these pairs were previously described in the literature; one pair is new. This new pair of ArgR sites lies in the *aroP–pdhR* intergenic region (IGR; see Fig. 1). A complete listing of our results and predictions is found at www.genome.org and http://arep.med.harvard.edu/ecoli_matrices/spacing/spacing_predictions.html. We have not included predictions from six of the 55 search matrices in our ranked list of predictions because these six search matrices (DnaA, Hns, IHF, Lrp, RpoS, RpoD) are extremely nonspecific and yield a large number of hits in the genome.

Binding Site Knockouts

Of the top-scoring predictions, we selected five based on biological interest (see Table 1). The following genomic knockouts of five sets of predicted motifs were created in MG1655 *E. coli*: two ArgR sites in the *aroP–pdhR* IGR, a single GalR site in the *ytfQ–ppa* IGR, three PhoB sites in the *dinJ–yafL* IGR, three PhoB sites in the *yqeF–yqeG* IGR, and four MetJ sites in the *ybdH–ybdL* IGR. Duplicate knockout clones were isolated for three of these five sets of predicted sites. Additionally, a genomic knockout of two predicted LexA sites upstream of *arsR* was attempted, but because no knockouts were isolated out of 78 colonies screened, it is possible that this knockout may be lethal.

The knockouts created are not mere deletions of the predicted transcription-factor-binding sites, but, rather, they are substitutions of the most information-rich bases in the motif with those found with least frequency in a given transcription factor's footprinted binding sites (see Fig. 2). Furthermore, the replacement sequences were verified to ensure that they neither destroyed overlapping sites for other known transcription factors, nor created new potential sites.

Quantitative Real-Time PCR Assays

The results of triplicate quantitative real-time PCR assays are shown in Table 2. These data indicate that at least one of the genes immediately downstream from each of the five binding site knockouts exhibits a significant change in mRNA expression. This indicates that the predicted binding sites, which were mutated in the binding site knockout strains, are most likely real and involved in regulation of these downstream genes.

Negative-control quantitative real-time PCR assays consisted of mispairings between binding site knockout RNAs and primer/probe pairs (i.e., quantitative real-time PCR assays were performed on genes assayed in this project, but not downstream from the binding site knockouts in the particular assayed RNA). Out of 10 such mispairings, eight resulted in essentially no change (see Supplemental material available online at www.

Table 1. Predicted Sites That Were Experimentally Tested

Predicted pairs with ≥0 separation

Ranking	Site 1, site 2	Predicted regions	Probability	Separation
1	ArgR, ArgR	*aroP–pdhR*	9.1×10^{-12}	3 bp
8	LexA, LexA	*arsR*	6.7×10^{-7}	0–30 bp
9	GalR, CRP	*ppa–ytfQ*	1.6×10^{-6}	0 bp

Predictions from analysis of overlapping sites[a]

Ranking	Site 1, site 2	Predicted regions	Significance index	Separation
7	PhoB, PhoB	*dinJ–yafL, yqeF–yqeG*	2.3×10^{-12}	0 bp
14	MetJ, MetJ	*ybdH–ybdL*	6.0×10^{-12}	0 bp

Predicted sites are ranked according to the probability of obtaining the observed number of hits for the most overrepresented bin or spacing, given the number expected by chance for that particular bin or spacing ("separation"). Predictions coming from our analysis of pairs separated by ≥0 bp and predictions from our analysis of overlapping sites are treated separately (see Methods).

[a]The *yqeF–yqeG* and *dinJ–yafL* IGRs each contain three adjacent 11-bp PhoB-predicted sites; the *ybdH–ybdL* IGR contains four adjacent 8-bp predicted MetJ sites. For PhoB and MetJ, we searched the genome with a matrix consisting of two adjacent sites, because a matrix consisting of a single site is not specific enough to be useful. Thus, triplets of the sites presented above showed up in our analysis as overlapping dimers of sites (two 22-bp matrix hits overlapping by 11 bp in the case of PhoB, or two 16-bp matrix hits overlapping by 8 bp in the case of MetJ). In addition, we constructed a highly specific 33-bp matrix from all known and footprinted triplets of PhoB sites, which identifies a very small number of sites in the genome, including the known sites and the *dinJ–yafL* and *yqeF–yqeG* IGRs. We also constructed a highly specific 24-bp matrix from all known and footprinted triplets of MetJ sites, which identifies only one new IGR in the genome in addition to the known ones (the *ybdH–ybdL* IGR). This matrix actually predicts two 24-bp sites in the *ybdH–ybdL* upstream region overlapping by 8 bp (i.e., a 32-bp pattern consisting of four consecutive motif instances). The fourth 8-bp site is weak, however; it does not show up when searching with the shorter 16-bp matrix. PhoB and MetJ also both showed up as highly significant in our spacing analysis with the pattern consisting of two dimers of sites separated by 0 bp (i.e., four adjacent sites—a 44-bp pattern for PhoB with probability 4.2*e*-6, or a 32-bp pattern for MetJ with a probability 1.1*e*-7). However, only known sites contributed to these spacing patterns; thus, there are no new predictions fitting this pattern. For both PhoB and MetJ, there are more footprinted sites with triplets of adjacent sites than strings of four adjacent sites. Our predictions in the table were based on significant triplets of sites.

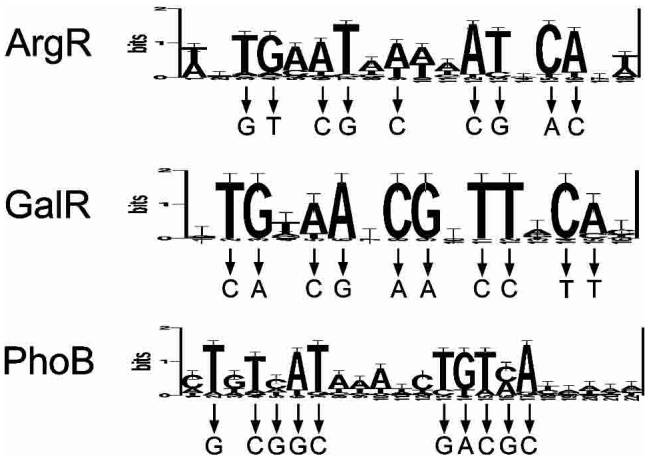

Figure 2 Design of binding site knockouts. Only the binding site substitution knockouts for ArgR, GalR, and PhoB are shown. The same strategy was followed in designing the MetJ- and LexA-binding site mutations (data not shown).

genome.org). One negative control that resulted in a significant change was *pdhR* in the strain containing MetJ-binding site knockouts in the *ybdH–ybdL* IGR. YbdH is a hypothetical oxidoreductase, and YbdL is a hypothetical aminotransferase; it is possible that misregulation of *ybdH* and/or *ybdL* might impact glycolysis and the TCA cycle, causing a change in expression of *pdhR*. The other negative control that resulted in a change was *aroP* in the strain containing PhoB binding site knockouts in the *yqeF–yqeG* IGR. The functions of these proteins are unknown; it is possible that misregulation of *yqeF* and/or *yqeG* might cause a change in expression of *aroP*.

Table 2. Results of Quantitative Real-Time PCR Assays

Transcript assayed	Strain	Expression ratio
aroP	ArgRpdhR clone 1	1.3 (0.33)
pdhR	ArgRpdhR clone 1	2.4 (0.35)
ppa	GalRytfQ clone 1	1.11 (0.45)
ytfQ	GalRytfQ clone 1	1.34 (0.23)
ybdH	MetJybdH clone 1	1.19 (0.21)
ybdH	MetJybdH clone 2	1.55 (0.21)
ybdL	MetJybdH clone 1	3.25 (0.36)
ybdL	MetJybdH clone 2	3.82 (0.40)
yqeF	PhoByqeG clone 1	0.208 (0.049)
yqeF	PhoByqeG clone 2	0.203 (0.051)
yqeG	PhoByqeG clone 1	1.35 (0.18)
yqeG	PhoByqeG clone 2	1.39 (0.26)
dinJ	PhoByafL clone 1	0.73 (0.16)
dinJ	PhoByafL clone 2	0.85 (0.22)
dinJ	PhoByafL clone 3	0.68 (0.18)
yafL	PhoByafL clone 1	3.0 (0.6)
yafL	PhoByafL clone 2	2.2 (0.5)
yafL	PhoByafL clone 3	2.2 (0.5)
pdhR	Δ*argR* clone 1	1.3 (0.43)

Ratios resulting from mean values of triplicate assays are shown, with standard deviations given in parentheses. Ratios given are relative to wild-type MG1655 grown under the same conditions as the particular knockout strain. In the notation for naming the binding site knockouts, the name of the transcription factor whose site(s) have been mutated is followed by the name of one of the genes downstream of the mutated site(s). For example, "ArgRpdhR" indicates that at least one ArgR-binding site was knocked out upstream of the *pdhR* gene.

Transcription Factor Knockouts

Another way to test our transcription-factor-binding site predictions is to knock out the predicted transcription factor and see if the mRNA levels of the ORF(s) physically downstream from the predicted binding site(s) are either up- or down-regulated. However, several secondary effects may make interpretation of such data difficult. For example, there may be genes regulated by the knocked out transcription factor that can then up- or down-regulate the genes physically downstream from the predicted binding site, and this indirect regulation may confound correct interpretation of the data (Lee et al. 2002b). Nevertheless, as a demonstration of this approach in comparison to the binding site knockout approach, we examined the effects of an *argR* knockout on *pdhR* expression (see Table 2 and Fig. 3). The results from analysis of the *argR* knockout are consistent with the results from analysis of the strain with mutated ArgR-binding sites in the *aroP–pdhR* IGR. The slight differences in expression level of *pdhR* in the *argR* knockout as compared with in the ArgR-binding sites knockout might be due to secondary effects of deleting *argR*, which is predicted to regulate other transcriptional regulators (data not shown).

Primer Extension Assays

In primer extension experiments using the ArgR-binding sites knockout, a 1.2-fold derepression of *pdhR* expression was observed (see Fig. 3). This is consistent with the 2.4-fold derepression of *pdhR* observed by quantitative real-time PCR. It is also consistent with the 1.5-fold *pdhR* derepression observed in a primer extension assay of an *argR* transcription factor knockout (data not shown).

Affymetrix Oligonucleotide Array mRNA Expression Analysis

Affymetrix oligonucleotide arrays (Affymetrix 2002) were used to perform genome-wide mRNA expression analysis of duplicate wild-type and binding site knockout strains, to identify the secondary effects of a knockout of the three PhoB sites in the *yqeF–yqeG* IGR. This binding site knockout was chosen because it resulted in the most dramatic fold change, as assayed by quantitative real-time PCR, of one of the genes immediately downstream from the mutated binding sites (i.e., fivefold decreased expression of *yqeF*). However, the genome-wide mRNA expression data indicated that only three other genes, *ybaM*, *vsr*, and *gcvP*, showed changes at least as great as *yqeF*. Because it is unclear how these three genes might fit together into a biological pathway, no

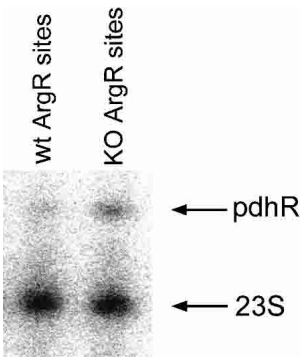

Figure 3 Primer extension assay of ArgR-binding site knockout in the *aroP–pdhR* intergenic region. The levels of *pdhR* transcript were measured, using the 23S-specific internal probe as an internal quantitation control for each RNA sample.

conclusions could be drawn regarding the possible function of *yqeF*.

DISCUSSION

Many of the binding site knockouts are upstream of uncharacterized genes or operons. Some of these URFs have important homologies for connecting them to regulons via the predicted transcription-factor-binding sites. For example, *ytfQ,* which is downstream from predicted GalR- and CRP-binding sites, shows significant homology in a BLAST search to a number of D-ribose-binding periplasmic proteins (*E*-value = 2×10^{-29}) and ribose ABC transporters from a few different prokaryotes, including homology to D-galactose-binding periplasmic proteins (*E*-value = 5×10^{-5}). Its highest scoring hit is for homology to a bifunctional carbohydrate binding and transport protein (*E*-value = 2×10^{-30}). These homologies provide further support that this URF might be regulated by transcription factors known to regulate galactose metabolism.

In addition, the results of such experiments can indicate interconnections between various regulons. For example, our data indicate that the two ArgR sites, separated by 3 bp, predicted upstream of the *pdhR–aceEF–lpd* operon are functional. ArgR is presently only known to regulate genes specifically involved in arginine biosynthesis, whereas PdhR is the repressor of the pyruvate dehydrogenase complex. However, the product of the pyruvate dehydrogenase complex, acetyl-CoA, is needed in the first step of arginine biosynthesis. This biochemical pathway information further supports our finding that the ArgR-binding sites upstream of the *pdhR–aceEF–lpd* operon are functional, and thus that the ArgR regulon is interconnected with the PdhR regulon.

Furthermore, once a predicted site has been demonstrated to be functional, that site can then be added to the set of sequences used to generate a binding site weight matrix for the given transcription factor. That refined weight matrix can then be used in a new search of the *E. coli* genome to identify a refined list of predicted binding sites. This set of genes can then be analyzed to determine how the genes are involved in a regulon and thus to further characterize the functions of these genes. Moreover, if the gene(s) physically downstream from the predicted binding site(s) have not yet been functionally characterized, then the functions of the genes affected in a secondary manner should aid in assigning the URF to a regulon.

A critical point in these experiments is selection of the proper culture conditions to permit analysis of the predicted binding site. The culture conditions used must be those that will induce expression of the transcription factor in the wild-type cells. Otherwise, if the transcription factor is not expressed, then none of the transcription factor's binding sites will be bound, and comparing wild-type versus the knockout will not provide data on the predicted binding site.

A particularly interesting finding is that despite the fact that the predicted binding sites we examined were all in divergent promoters, the immediately downstream genes in either direction were not affected equally by the binding site mutations. For example, mutating the three PhoB sites in the *yqeF–yqeG* IGR resulted in 5.0-fold down-regulation of *yqeF*, and 1.4-fold up-regulation of *yqeG*. Similarly, mutating the four MetJ sites in the *ybdH–ybdL* IGR resulted in 1.4-fold up-regulation of *ybdH* and 3.6-fold up-regulation of *ybdL*. These different changes in gene expression could not be explained by the distance between the predicted sites and the affected gene; that is, considering these two pairs of divergently transcribed genes, the genes closer to the predicted sites do not as a rule exhibit a stronger up- or down-regulation as a result of the binding site knockouts as compared with the genes that are farther away from the predicted sites. It is unclear what may be the mechanism of this differential regulation at divergent promoters. In constructing the binding site mutations, care was taken neither to disrupt overlapping transcription factor binding sites, nor to create new sites for the 55 *E. coli* transcription factors for which weight matrices have been published. Nevertheless, it is possible that some as-yet-unknown binding site, whose binding factor functions in a directional manner, or some other kind of DNA sequence element that functions in an orientation-dependent manner, was either disrupted or created in the binding site knockouts. For example, a sequence-dependent bend upstream of the rRNA promoter *P1* in *E. coli* is responsible for high promoter activity, and both the distance and angular orientation of the bent DNA is crucial for the degree of activation (Zacharias et al. 1991). Similarly, the transcription factors that bind the predicted binding sites might function in a directional manner. It is also possible that some as-yet-unknown higher-order chromosome structure occurs that results in differential expression of the genes at these divergent promoters. For example, despite extensive overlap of regulatory elements, the divergently transcribed *E. coli* genes *nrfA* and *acsP1* are regulated independently; evidence indicates that a nucleoprotein structure in this intergenic region allows these genes to be regulated independently (Browning et al. 2002).

Initial site clustering approaches that simply consider a certain number of sites within a given genomic sequence window size recently have produced some initial successes in predicting DNA regulatory elements in eukaryotic genomes (Wagner 1999; Frith et al. 2001, 2002; Pilpel et al. 2001; Berman et al. 2002; Halfon et al. 2002; Kel-Margoulis et al. 2002a; Markstein et al. 2002; Rebeiz et al. 2002). A binding site co-occurrence approach (Sudarsanam et al. 2002) that considers the spacing between transcription-factor-binding site motifs, such as the one we present here, might be useful in further improving the accuracy with which regulatory binding sites are predicted in higher eukaryotic genomes.

METHODS

Binding Site Prediction

Only those matrix hits that occurred within noncoding regions were analyzed because most experimentally confirmed binding sites for transcription factors occur in noncoding regions (of course, this could be at least in part caused by a bias in where people traditionally have looked for transcription-factor-binding sites). We used no size restrictions on noncoding regions; any nucleotide that does not code for protein is called noncoding in our analysis. All matrix hits scoring above two standard deviations below the mean of the scores of the known footprinted (input) sites (Robison et al. 1998) were used in the spacing analysis. The assumption that scores of sites follow a normal distribution appears to be valid, and the vast majority of known sites fall within two standard deviations of the mean (Robison et al. 1998). The matrix pairs were ranked according to either their most significant single spacing between 0 and 500 bp (e.g., exactly 3 bp) or their most significant spacing bin (McGuire 2000). Eight different spacing bins were examined (the bins including separation distances 0–30 bp, 30–60 bp, 60–90 bp, 0–100 bp, 100–200 bp, 200–300 bp, 300–400 bp, and 0–450 bp).

The rankings were based on the probability of obtaining the observed number of hits for the most overrepresented bin or spacing, given the number expected by chance for that particular bin or spacing. This number expected by chance was determined in the following manner:

$$E(x) = N_a \cdot N_b \cdot \pi(x - c), \tag{1}$$

where N_a and N_b are the number of hits in the genome using search matrices a and b, c is a correction factor to account for the lengths of the search matrices, and $\pi(x)$ is the probability that

two randomly chosen noncoding base pairs are separated by a distance x. $\pi(x)$ was computed by tabulating the actual frequencies of occurrence of separations between all pairwise combinations of noncoding bases in *E. coli*. $\pi(x)$ is a decreasing function of x (McGuire 2000). When we refer to the spacing between matrix hits, we are referring to the distance between the end of the first search matrix and the beginning of the second search matrix. Thus, the length of the first search matrix, c, is needed in equation 1.

The probability $P(x)$ of obtaining at least the observed number of pairs, $obs(x)$, at each spacing x between 0 and 500 bp, given $N_a \cdot N_b$ trials, where the probability of observing a pair at this spacing in a single trial is $\pi(x - c)$, was then calculated:

$$P(x) = 1 - \sum_{s=0}^{obs(x)-1} \binom{N_a \cdot N_b}{s} \cdot \pi(x-c)^s \cdot (1 - \pi(x-c))^{N_a \cdot N_b - s} \quad (2)$$

where s is an index variable in the summation.

By checking 500 different spacings, multiple hypotheses are being tested. To obtain a more reliable probability value, the probability of observing $obs(x)$ sites at any single spacing within a spacing range that includes x (i.e., 0 to x bp), $P(x)$ was summed over this range of x values. All pairs of search matrices that have a spacing x for which this adjusted value for $P(x)$ is <0.05 were saved.

Similarly, the probabilities of obtaining the observed number of hits within the eight different spacing bins was calculated:

$$P_{bin} = 1 - \sum_{s=0}^{obs(bin)-1} \binom{N_a \cdot N_b}{s} \cdot \Pi^s \cdot (1 - \Pi)^{N_a \cdot N_b - s}, \quad (3)$$

$$\Pi = \sum_{x=0}^{binsize} \pi(\mathrm{x - c}), \quad (4)$$

$$obs(bin) = \sum_{x=0}^{binsize} obs(x), \quad (5)$$

where $obs(bin)$ is the observed number of hits in that spacing bin.

In the case in which the two search matrices are identical, the number of hits expected by chance should be determined in the following manner:

$$E(x) = \frac{N_a \cdot (N_a - 1)}{2} \cdot \pi(x - c) \quad (6)$$

The equations for $P(x)$ and P_{bin} can be modified accordingly.

The matrix pairs were ranked according to their values for P_{bin} for each of the spacing bins, and all those that had values for $P_{bin}(x_{min} \ldots x_{max})$ <0.05 were saved.

In our calculation of probabilities, we assumed the presence of two independent sites. This assumption is not valid in the case of overlapping sites. However, we found it useful to calculate a "significance index" for the overlapped data in the same way as we calculated the probabilities above. These values are not comparable with the probabilities listed above because of the nonindependence of the two overlapped sites, but we found this index to be useful for comparisons within our analysis of the overlapped data.

Because most of the search matrices, even the more nonspecific ones such as DnaA, Hns, Ihf, Lrp, OmpR, Fis, NarL, TyrR, and RpoS, are biased in their distribution within the noncoding regions, the false positives can be expected likewise to be distributed nonrandomly. This nonrandomness is caused by variation in AT content for different noncoding regions in *E. coli*. A sharp dip in AT content at ~10 bp upstream of the start codon is due to the Shine-Dalgarno sequence (AGGAGG). Furthermore, AT content dips between 10–40 bp downstream of the stop codon. This dip is partially explained by the presence of BIMEs, which have a 42% AT content (the overall AT content for noncoding regions in *E. coli* is 58%). This background nonrandomness in the locations of false-positive search matrix hits results in a nonrandom distribution of spacing distances as well. We calculated the ratio of the number of observed pairs at a spacing x to the expected number of pairs at this spacing, $R = obs(x)/E(x)$. Typical values of R are ~1.5–2. A cutoff of $R = 5$ was used to exclude the majority of the nonspecific hits. The most nonspecific matrices, which produced >10,000 hits upon search of *E. coli* noncoding regions, were not included for purposes of predicting binding sites (McGuire 2000).

Strains

For cloning purposes, *E. coli* DH5α was used. The binding site knockouts were created in MG1655 *E. coli*, which was used by Blattner et al. (1997) for the determination of the *E. coli* genome sequence. The transcription factor knockouts were created in EMG2 *E. coli*, so that they could be grown competitively with a set of 46 other transcription factor knockouts created in EMG2 (Phillips 2000).

Binding Site and Transcription Factor Knockouts

For both the binding site knockouts and the transcription factor knockouts, ~500 bp of flanking DNA 5′ and 3′ of the sequences to be replaced were amplified in two separate PCR reactions (Ni and No primers were used to amplify the N-terminal flanking sequence, and Ci and Co primers were used to amplify the C-terminal flanking sequence). Restriction sites were incorporated into the No and Co primers to permit unidirectional ligation into the plasmid pKOV. The plasmid pKOV is a stuffer-containing derivative of pKO3. For the binding site knockouts, the Ni and Ci primers contained the replacement binding site sequences. For the transcription factor knockouts, the Ni and Ci primers contained a 33-bp tag sequence. The two PCR products representing the N- and C-terminal flanking regions were reamplified in a second PCR reaction using the No and Co primers before cloning into pKOV. A list of primers used in creating the transcription factor knockouts can be found in Supplemental material. The transcription factor knockouts resulted in replacement of the coding sequence from start to stop with the 33-bp tag sequence. Cointegration, resolution, and elimination of the plasmid were performed as previously described (http://arep.med.harvard.edu/labgc/pko3.html; Phillips 2000).

The binding site substitutions were created by modifying the current pKOV knockout scheme (Link et al. 1997; http://arep.med.harvard.edu/labgc/pko3.html) such that (1) no universal tag is included on the inner knockout primers; and (2) the inner primers are not exact complements of wild-type MG1655 genomic DNA, but rather contain mutant versions of the predicted binding sites. Care was taken neither to disrupt overlapping transcription factor binding sites, nor to create new sites for the 55 *E. coli* transcription factors for which weight matrices have been published. For example, the nucleotide substitutions created in the three predicted PhoB binding sites in the *dinJ–yafL* IGR were carefully chosen so as to maximally disrupt the PhoB motif while minimizing disruptive mutation of the two overlapping ArcA sites. A list of primers used in creating the binding site knockouts can be found in Supplemental material.

Chloramphenicol-sensitive colonies were tested by PCR. For the transcription factor knockouts, the Co and No primers were used in PCR. Because these primers flanked the gene, the size of the PCR product indicated whether the template was from a wild-type or deletion strain. For the binding site knockouts, analytical PCRs were performed using the Co primer in conjunction with a primer representing either the wild-type or mutant binding site, in two separate PCR reactions. These primers were designed to be complementary to either the wild-type binding site sequence or the mutant binding site sequence. See Supplemental material for a listing of these primer sequences.

Binding site knockouts were verified by sequencing. See Supplemental material for a listing of the sequencing primers.

PCR

All PCRs were performed essentially as previously described (http://twod.med.harvard.edu/labgc/estep/longPCR_protocol.html).

Media and Culture Conditions

Duplicate wild-type and binding site knockout strains were grown under the following conditions (Neidhardt et al. 1974): (1)

knockout ArgR-binding sites: 37°C; 1× M9 minimal, 0.4% glucose, 1 mM arginine; to ~0.3 OD_{600}; (2) knockout GalR-binding sites: 37°C; 1× M9 minimal, 0.4% glucose; to ~0.35 OD_{600}; (3) knockout MetJ-binding sites: 37°C; 1× M9 minimal, 0.4% glucose, 1 mM methionine; to ~0.5 OD_{600}; (4) knockout PhoB-binding sites: 37°C; 1× MOPS, 0.4% glucose, 0.066 mM K_2HPO_4 (phosphate-limited); to ~0.25 OD_{600}. These conditions were chosen to induce expression of the transcription factor whose predicted sites were knocked out.

Duplicate wild-type and transcription factor knockout strains were grown under the conditions listed below. The *argR* knockout strains were grown at 37°C in M9 minimal medium with 11 mM (0.2%) glucose, 0.5% casamino acids, 1 mM arginine (Charlier et al. 1992; Tian et al. 1994), to ~0.82 OD_{600}. For the primer extension assays, these cultures were grown to 0.7–0.9 OD_{600}. These conditions were chosen to induce expression in wild-type cells of the knocked out transcription factor.

RNA Isolation and Purification

Bacterial lysis and isolation of crude total RNA was achieved using hot acid phenol (Mangan et al. 1997). Briefly, cultures were pelleted by centrifugation at 5000*g*. Cell pellets were stored at −80°C. Cells were resuspended in acidic phenol:chloroform, 5:1 solution (pH 4.5; Ambion), prewarmed at 65°C. RNA was extracted three times with acidic phenol:chloroform by addition of acidic phenol:chloroform, vortexing, incubating for 3 min at 65°C, incubating for 3 min on ice, then centrifuging for 5 min. The RNA was purified once with chloroform at room temperature, ethanol-precipitated, and resuspended in DEPC-treated H_2O. RNA samples were quantified at A_{260} and A_{280} using a spectrophotometer.

Primer Extension Analysis

Primer extension analysis was performed as described before (Sambrook et al. 1989). Briefly, 10 µg of total cellular RNA, [γ-^{32}P]-end-labeled gene-specific probe, and [γ-^{32}P]-end-labeled 23S-specific probe were heated for 90 min at 65°C in hybridization buffer and then slowly cooled down to allow for specific annealing of the probes. The 23S-specific probe served as an internal quantitation control for each RNA sample. Afterward, a mix of AMV reverse transcriptase, actinomycin D, and dNTPs in reverse transcription buffer was added at 42°C and the reaction was allowed to proceed for 1 h. The reaction was then incubated with RNase I for 15 min at 37°C. The products were purified by phenol:chloroform extraction and ethanol precipitation, and then run on denaturing acrylamide gels. The gels were scanned on a Molecular Dynamics Storm PhosphorImager and quantified using ImageQuant software. Sequences of the primers used as probes can be found in Supplemental material.

Real-Time RT-PCR Primers and Fluorogenic Probes

PCR primers and fluorogenic probes for the genes of interest were designed using DNASTAR PrimerSelect software. The *rrlH* ribosomal RNA gene was used as the internal control gene against which all other genes were normalized. Probes were selected such that their T_ms were ~7°–10°C higher than the matching primer pair. The dual-labeled fluorogenic probes contained an FAM report dye covalently attached at the 5′-end and a BHQ1 quencher dye covalently attached at the 3′-end. These probes were synthesized and HPLC-purified by BioSearch Technologies, Inc. A listing of primers and fluorogenic probes used in the quantitative real-time PCR assays can be found in Supplemental material.

Real-Time RT-PCR Amplification

RT-PCR reactions were carried out in iCycler IQ Real-Time Detection Systems (Bio-Rad). SuperScript One-step RT-PCR with Platinum Taq kits (Invitrogen) were used for all RT-PCR amplification in a total volume of 50 µL, which contained 200 ng of total RNA, 5 mM $MgSO_4$, 500 nM forward and reverse primers, and 200 nM fluorogenic probe. RT-PCR amplification for each RNA sample was performed in triplicate wells. One "no RT" (without reverse transcriptase) control for each RNA sample and one "no RNA" (substituted RNA with dH_2O) control for each primer and probe set were also performed. The one-step RT-PCR condition is as follows: 15 min at 50°C and 5 min at 95°C, followed by a total of 45 two-temperature cycles (15 sec at 95°C and 1 min at 60°C). Relative gene expression data analysis was carried out with the standard curve method (Heid et al. 1996; Winer et al. 1999).

mRNA Expression Analysis Using Affymetrix Oligonucleotide Arrays

Genome-wide mRNA expression analysis using Affymetrix GeneChip oligonucleotide arrays was performed essentially as described previously. Briefly, to enrich for mRNA, reverse transcriptase and primers specific to 16S and 23S rRNA were used to synthesize cDNAs from total RNA. Then, rRNAs were removed by incubation with RNase H, which specifically digests rRNA within an RNA:DNA hybrid. The cDNAs were then removed by DNase I digestion, and the enriched mRNA was then purified on QIAGEN RNeasy columns. The purified, enriched mRNA was fragmented by heat and ion-mediated hydrolysis, and labeled at the 5′-ends with [γ-S]ATP using T4 polynucleotide kinase. The thiolated RNA was then labeled with PEO-iodoacetyl-biotin, and hybridized to the chip. After washing, the chip was stained with streptavidin, followed by staining with biotin-conjugated anti(streptavidin) antibody, and then finally by phycoerythrin-conjugated streptavidin.

ACKNOWLEDGMENTS

We thank Dereth Phillips, Xiaohua Huang, Vasudeo Badarinarayana, Aimee Dudley, Doug Selinger, Martin Steffen, Rey Sequerra, and Jennifer C. Lee for technical assistance. We also thank Dereth Phillips, Jason Hughes, Pete Estep, Tzachi Pilpel, and other members of the Church Lab for helpful discussion. M.L.B. was partially supported by an NSF Graduate Fellowship. A.M.M. was a Howard Hughes Predoctoral Fellow. This work was supported in part by a grant from the Office of Naval Research (N00014-99-1-0783).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Affymetrix, Inc. 2002. *Affymetrix GeneChip Expression Analysis Technical Manual*. Affymetrix, Inc., Santa Clara, CA.

Bailey, T. and Elkan, C. 1995. The value of prior knowledge in discovering motifs with MEME. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **3:** 21–29.

Benos, P., Bulyk, M., and Stormo, G. 2002. Additivity in protein–DNA interactions: How good an approximation is it? *Nucleic Acids Res.* **30:** 4442–4451.

Berg, O. and von Hippel, P. 1987. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.* **193:** 723–750.

Berman, B.P., Nibu, Y., Pfeiffer, B.D., Tomancak, P., Celniker, S.E., Levine, M., Rubin, G.M., and Eisen, M.B. 2002. Exploiting transcription factor binding site clustering to identify *cis*-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc. Natl. Acad. Sci.* **99:** 757–762.

Blattner, F., Plunkett III, G., Bloch, C., Perna, N., Burland, V., Riley, M., Collado-Vides, J., Glasner, C., Rode, G., Mayhew, J., et al. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* **277:** 1453–1474.

Browning, D., Beatty, C., Wolfe, A., Cole, J., and Busby, S. 2002. Independent regulation of the divergent *Escherichia coli nrfA* and *acsP1* promoters by a nucleoprotein assembly at a shared regulatory region. *Mol. Microbiol.* **43:** 687–701.

Bulyk, M., Johnson, P., and Church, G. 2002. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res.* **30:** 1255–1261.

Bussemaker, H., Li, H., and Siggia, E. 2000. Building a dictionary for genomes: Identification of presumptive regulatory sites by statistical analysis. *Proc. Natl. Acad. Sci.* **97:** 10096–10100.

Charlier, D., Roovers, M., Vliet, F.V., Boyen, A., Cumin, R., Nakamura, Y., Glansdorff, N., and Pierard, A. 1992. Arginine regulon of *Escherichia coli* K-12: A study of repressor–operator interactions and

of in vitro binding affinities versus in vivo repression. *J. Mol. Biol.* **226:** 367–386.
Frech, K. and Werner, T. 1997. Specific modelling of regulatory units in DNA sequences. *Pac. Symp. Biocomput.* 151–162.
Frith, M., Hansen, U., and Weng, Z. 2001. Detection of *cis*-element clusters in higher eukaryotic DNA. *Bioinformatics* **17:** 878–889.
Frith, M., Spouge, J., Hansen, U., and Weng, Z. 2002. Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. *Nucleic Acids Res.* **30:** 3214–3224.
Grundy, W., Bailey, T., and Elkan, C. 1996. ParaMEME: A parallel implementation and a web interface for a DNA and protein motif discovery tool. *CABIOS* **12:** 303–310.
GuhaThakurta, D. and Stormo, G. 2001. Identifying target sites for cooperatively binding factors. *Bioinformatics* **17:** 608–621.
Halfon, M., Grad, Y., Church, G., and Michelson, A. 2002. Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model. *Genome Res.* **12:** 1019–1028.
Heid, C., Stevens, J., Livak, K., and Williams, P. 1996. Real time quantitative PCR. *Genome Res.* **6:** 986–994.
Hengge-Aronis, R. 1999. Interplay of global regulators and cell physiology in the general stress response of *Escherichia coli*. *Curr. Opin. Microbiol.* **2:** 148–152.
Hertz, G. and Stormo, G. 1999. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15:** 563–577.
Hughes, J., Estep, P., Tavazoie, S., and Church, G. 2000. Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.* **296:** 1205–1214.
Kel-Margoulis, O., Ivanova, T., Wingender, E., and Kel, A. 2002a. Automatic annotation of genomic regulatory sequences by searching for composite clusters. *Pac. Symp. Biocomput.* 187–198.
Kel-Margoulis, O., Kel, A., Reuter, I., Deineko, I., and Wingender, E. 2002b. TRANSCompel: A database on composite regulatory elements in eukaryotic genes. *Nucleic Acids Res.* **30:** 332–334.
Klingenhoff, A., Frech, K., Quandt, K., and Werner, T. 1999. Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity. *Bioinformatics* **15:** 180–186.
Kolchanov, N., Ignatieva, E., Ananko, E., Podkolodnaya, O., Stepanenko, I., Merkulova, T., Pozdnyakov, M., Podkolodny, N., Naumochkin, A., and Romashchenko, A. 2002. Transcription Regulatory Regions Database (TRRD): Its status in 2002. *Nucleic Acids Res.* **30:** 312–317.
Lawrence, C., Altschul, S., Boguski, M., Liu, J., Neuwald, A., and Wootton, J. 1993. Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science* **262:** 208–214.
Lee, M.-L., Bulyk, M., Whitmore, G., and Church, G. 2002. A statistical model for investigating binding probabilities of DNA nucleotide sequences using microarrays. *Biometrics* **58:** 981–988.
Lee, T., Rinaldi, N., Robert, F., Odom, D., Bar-Joseph, Z., Gerber, G., Hannett, N., Harbison, C., Thompson, C., Simon, I., et al. 2002. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298:** 799–804.
Li, H., Rhodius, V., Gross, C., and Siggia, E. 2002. Identification of the binding sites of regulatory proteins in bacterial genomes. *Proc. Natl. Acad. Sci.* **99:** 11772–11777.
Link, A., Phillips, D., and Church, G. 1997. Methods for generating precise deletions and insertions in the genome of wild-type *Escherichia coli*: Application to open reading frame characterization. *J. Bacteriol.* **179:** 6228–6237.
Liu, X., Brutlag, D., and Liu, J. 2001. BioProspector: Discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.* 127–138.
———. 2002. An algorithm for finding protein–DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat. Biotechnol.* **20:** 835–839.
Man, T.K. and Stormo, G.D. 2001. Non-independence of Mnt repressor–operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Res.* **29:** 2471–2478.
Mangan, J.A., Sole, K.M., Mitchison, D.A., and Butcher, P.D. 1997. An effective method of RNA extraction from bacteria refractory to disruption, including mycobacteria. *Nucleic Acids Res.* **25:** 675–677.
Markstein, M., Markstein, P., Markstein, V., and Levine, M. 2002. Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the *Drosophila* embryo. *Proc. Natl. Acad. Sci.* **99:** 763–768.
Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A., Kel-Margoulis, O., et al. 2003. TRANSFAC: Transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* **31:** 374–378.
McGuire, A. 2000. "Computational studies of transcriptional regulation in prokaryotes." Ph.D. thesis, Harvard University, Cambridge.
Neidhardt, F. 1996. *Escherichia coli and Salmonella: Cellular and molecular biology*. American Society for Microbiology, Washington, DC.
Neidhardt, F., Bloch, P., and Smith, D. 1974. Culture medium for enterobacteria. *J. Bacteriol.* **119:** 736–747.
Phillips, D. 2000. "Competitive growth analysis of *E. coli* in-frame deletion mutants across a spectrum of environmental conditions." Ph.D. thesis, Harvard University, Cambridge.
Pilpel, Y., Sudarsanam, P., and Church, G. 2001. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.* **29:** 153–159.
Quandt, K., Grote, K., and Werner, T. 1996. GenomeInspector: A new approach to detect correlation patterns of elements on genomic sequences. *Comput. Appl. Biosci.* **12:** 405–413.
Rebeiz, M., Reeves, N., and Posakony, J. 2002. SCORE: A computational approach to the identification of *cis*-regulatory modules and target genes in whole-genome sequence data. *Proc. Natl. Acad. Sci.* **99:** 9888–9893.
Robin, S., Daudin, J., Richard, H., Sagot, M., and Schbath, S. 2002. Occurrence probability of structured motifs in random sequences. *J. Comput. Biol.* **9:** 761–773.
Robison, K., McGuire, A.M., and Church, G.M. 1998. A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome. *J. Mol. Biol.* **284:** 241–254.
Rosenblueth, D.A., Thieffry, D., Huerta, A.M., Salgado, H., and Collado-Vides, J. 1996. Syntactic recognition of regulatory regions in *Escherichia coli*. *Comput. Appl. Biosci.* **12:** 415–422.
Roth, F.P., Hughes, J.D., Estep, P.W., and Church, G.M. 1998. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotech.* **16:** 939–945.
Sambrook, J., Fritsch, E., and Maniatis, T. 1989. *Molecular cloning: A laboratory manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
Sudarsanam, P., Pilpel, Y., and Church, G. 2002. Genome-wide co-occurrence of promoter elements reveals a *cis*-regulatory cassette of rRNA transcription motifs in *Saccharomyces cerevisiae*. *Genome Res.* **12:** 1723–1731.
Thieffry, D., Salgado, H., Huerta, A., and Collado-Vides, J. 1998. Prediction of transcriptional regulatory sites in the complete genome sequence of *Escherichia coli* K-12. *Bioinformatics* **14:** 391–400.
Tian, G., Lim, D., Oppenheim, J.D., and Maas, W.K. 1994. Explanation for different types of regulation of arginine biosynthesis in *Escherichia coli* B and *Escherichia coli* K12 caused by a difference between their arginine repressors. *J. Mol. Biol.* **235:** 221–230.
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.
van Helden, J., Rios, A., and Collado-Vides, J. 2000. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res.* **28:** 1808–1818.
Wagner, A. 1999. Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics* **15:** 776–784.
Winer, J., Jung, C., Shackel, I., and Williams, P. 1999. Development and validation of real-time quantitative reverse transcriptase-polymerase chain reaction for monitoring gene expression in cardiac myocytes in vitro. *Anal. Biochem.* **270:** 41–49.
Workman, C. and Stormo, G. 2000. ANN-Spec: A method for discovering transcription factor binding sites with improved specificity. *Pac. Symp. Biocomput.* 467–478.
Zacharias, M., Theissen, G., Bradaczek, C., and Wagner, R. 1991. Analysis of sequence elements important for the synthesis and control of ribosomal RNA in *E. coli*. *Biochimie* **73:** 699–712.

WEB SITE REFERENCES

http://arep.med.harvard.edu/ecoli_matrices/spacing/spacing_predictions.html; Web site contains tab-delimited files containing predictions based on individual spacings, and separately based on spacing bins.
http://arep.med.harvard.edu/labgc/pko3.html; Descriptions of the gene replacement vectors pKO3 and pKOV.
http://twod.med.harvard.edu/labgc/estep/longPCR_protocol.html; Descriptions of the PCR conditions and protocols used in this project.

Received April 21, 2003; accepted in revised form November 5, 2003.