Article No. mb982160 *J. Mol. Biol.* (1998) **284**, 241–254

JMB

A Comprehensive Library of DNA-binding Site Matrices for 55 Proteins Applied to the Complete *Escherichia coli* K-12 Genome

Keith Robison[1], Abigail Manson McGuire[2] and George M. Church[1,2]*

[1]*Department of Genetics and* [2]*Graduate Program in Biophysics, Warren Alpert Building, Room 513, Harvard Medical School, 200 Longwood Avenue, Boston MA 02115, USA*

**Corresponding author*

A major mode of gene regulation occurs *via* the binding of specific proteins to specific DNA sequences. The availability of complete bacterial genome sequences offers an unprecedented opportunity to describe networks of such interactions by correlating existing experimental data with computational predictions. Of the 240 candidate *Escherichia coli* DNA-binding proteins, about 55 have DNA-binding sites identified by DNA footprinting. We used these sites to construct recognition matrices, which we used to search for additional binding sites in the *E. coli* genomic sequence. Many of these matrices show a strong preference for non-coding DNA. Discrepancies are identified between matrices derived from natural sites and those derived from SELEX (Systematic Evolution of Ligands by Exponential enrichment) experiments. We have constructed a database of these proteins and binding sites, called DPInteract (available at http://arep.med.harvard.edu/dpinteract).

Keywords: bioinformatics; DNA-binding proteins; matrix search; SELEX; footprinting

Introduction

Sequence-specific DNA-binding proteins perform a multitude of roles in a living cell and regulate a variety of processes including transcription. *Escherichia coli* contains at least 240 proteins that are known or predicted to be DNA-binding proteins (Robison, 1997). Known binding sites for a DNA-binding protein can be used to identify additional sites for that protein, and thereby identify further genes regulated by that protein (Wasserman & Fickett, 1998; Tronche *et al*., 1997; Fondrat & Kalogeropoulos, 1996; Goodrich *et al*., 1990; Lewis *et al*., 1994; Ramseier *et al*., 1995; Stormo, 1990; Verbeek *et al*., 1990).

A number of approaches have been used to search for additional sites, including searches using consensus sequences, and searches using position weight matrices. Fondrat & Kalogeropoulos (1996) used a precise set of rules and constraints together with a degenerate consensus pattern to search for binding sites for five yeast regulatory proteins on *Saccharomyces cerevisiae* chromosome III. A number of studies have used position weight matrices to characterize the distribution of bases at each position in the recognition sequence. In a recent matrix search study using an alignment of 21 DNA sequences recognized by the liver-specific transcription factor HNF-1, 52 out of the 54 high-scoring sites tested experimentally were found to bind HNF-1 *in vitro* (Tronche *et al*., 1997).

Several approaches for performing matrix searches have been proposed. We used the log transformation described by Berg & von Hippel because scores from this method have been shown to correlate with *in vitro* binding constants (Berg & von Hippel, 1988; Cui *et al*., 1995). A search method employing the Berg & von Hippel function scaled by an index of information content (Schneider *et al*., 1986) has shown promise for promoter recognition (O'Neill, 1989). In addition, neural network approaches have been used for promoter recognition (Grahn *et al*., 1994; Horton & Kanehisa, 1992). However, neural nets work best when trained and tested on large example sets, and for most DNA-binding proteins relatively few (<20) example sites are known. A different approach is the grammatical implementation of Collado-Vides (Rosenblueth *et al*., 1996). This

Present address: K. Robison, Millennium Pharmaceuticals Inc. 238 Main Street, Cambridge, MA 02138, USA

E-mail address of the corresponding author: Church@rascal.med.harvard.edu

0022–2836/98/470241–14 $30.00/0

approach restricts false positives by including additional information on biological properties of promoter regions, such as location and spacing between elements. This technique has been used with a combined weight matrix and string search strategy to predict binding sites for 56 transcriptional regulatory proteins in *E. coli* (Blattner *et al.*, 1997; Thieffrey *et al.*, 1998).

In this study, we were interested in performing an exhaustive matrix search for each motif so that we can study the entire distribution of sites in the genome and their spacing patterns. We systematically applied this approach to the complete *E. coli* genome sequence. From the literature, we identified 55 proteins for which footprinted DNA binding sites have been determined and built search matrices using these sites. We calibrated these search matrices by a variety of approaches, including determining the statistical distribution of the scores of the set of known sites, the distribution of scores on the complete *E. coli* genome, and the ability of the search matrices to distinguish between coding and non-coding DNA.

Results and Discussion

E. coli matrix searches

Several previous studies have reported the results of searches for binding sites of particular proteins in *E. coli*, including LexA (Lewis *et al.*, 1994), FruR (Ramseier *et al.*, 1995), Fis (Verbeek *et al.*, 1990), and Lrp (Cui *et al.*, 1995). We have created a library of 61 search matrices for 55 different *E. coli* DNA-binding proteins. Each matrix is constructed from natural sites identified by DNA-footprinting assays, based on data recorded in the DPInteract database (Robison, 1997), except for the RpoD matrix, which is supplemented by initiation data.

Our matrix search technique is intended to study spacing patterns between binding-site elements. For DNA-binding proteins with large data sets, for which spacing patterns have already been established, we can include multiple elements in the search. For binding sites containing two elements separated by a variable spacing (such as the −10 and −35 elements of the rpoD binding site), we created one matrix for each spacing class containing more than three known examples. Five different RpoD matrices were built based on the promoter compilation reported by Lisser & Margalit (1993), corresponding to the 15-19 nucleotide spacing classes of *E. coli* RpoD promoters. Two different RpoH and RpoS matrices were constructed, representing two promoter spacing classes for each (Gross, 1996; Wise *et al.*, 1996).

To score sites in the matrix searches, we chose the method of Berg and von Hippel (1987). This method uses a statistical-mechanical selection model to predict the affinity of a given DNA sequence based on the sequence statistics of the known footprinted sites. The following equation is used to obtain the score E for a given sequence:

$$E = \sum_{l=0}^{M} \ln\left[\frac{n_{lB} + 0.5}{n_{l0} + 0.5}\right] \quad (1)$$

M is the length of the binding site motif, B is the base at position l within the motif, n_{lB} is the number of occurrences of base B at position l in the footprinted input sites, and n_{l0} is the number of the occurrences of the most common base at position l in the footprinted input sites. Because the GC content of the *E. coli* genome is close to 50%, correction terms to account for the background energy due to genome composition are very close to zero.

We calibrated each matrix by measuring the mean (μ_i) and standard deviation (σ_i) for the set of footprinted input sites used to construct the matrix. A high score corresponds to a high-affinity site with a close match to the consensus, while a low score corresponds to a poor match to the consensus sequence. We used two standard deviations below the mean, $\mu_i - 2\sigma_i$, as the cutoff for further searches in the *E. coli* genomic sequence.

Each matrix was then used to score every possible sequence window in the *E. coli* genome and the mean (μ_g) and standard deviation (σ_g) of all possible genomic sites was computed. Figure 1 summarizes the ability of the binding-site search matrices to distinguish known binding sites from the remainder of the *E. coli* genome. A high genomic Z-score indicates a specific matrix for which the scores of the inputs are significantly higher than the scores of random sites in the genome. A lower genomic Z-score indicates a less specific search matrix, which probably has a greater number of false positive hits in the genome.

For each matrix, Table 1 lists the number of inputs, the %GC content, the number of sites in the genome scoring above μ_i, the number of sites scoring above the less restrictive cutoff $\mu_i - 2\sigma_i$, as well as the fraction found in non-coding regions for each of the two cutoffs. Listings of the locations of these sites in the genome can be found on our web site at http://arep.med.harvard.edu/ecoli_matrices.

It should be noted that if all input sites are a perfect consensus, they will all have a score of zero and hence show no deviation ($\mu_i = 0$ and $\sigma_i = 0$). This is guaranteed to occur if only two examples are available, as each position will either have a single nucleotide represented or two equally frequent nucleotides. It can occur also if only a few examples of a site are known. The small-sample variance in each individual score calculation (σ_ε^2) can be calculated by using the equation:

$$\sigma_\varepsilon^2 = \frac{1}{n_{l0} + 0.5} + \frac{1}{n_{lB} + 0.5} \quad (2)$$

(Berg & von Hippel, 1988). This value is large for

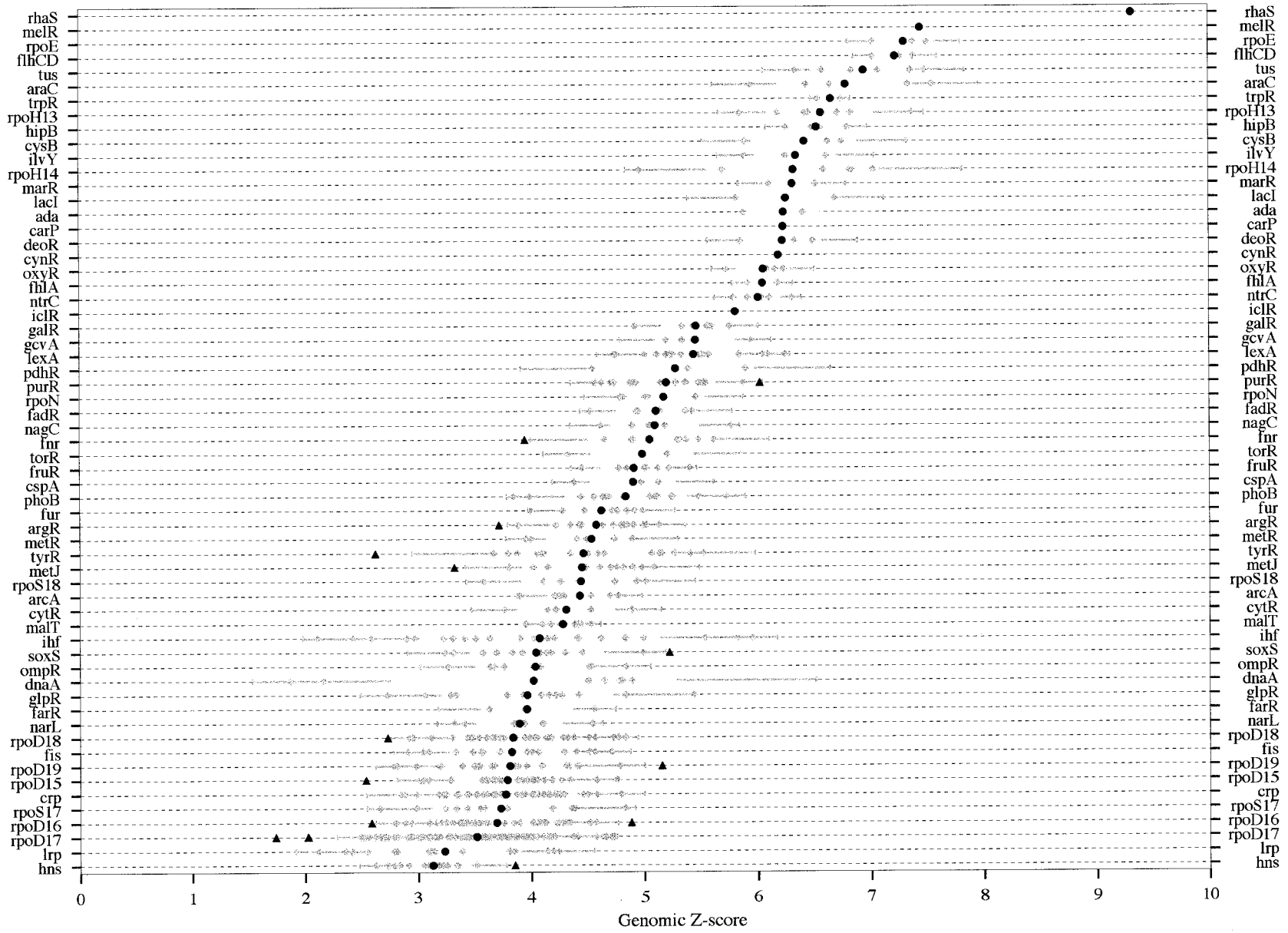

Figure 1. Matrix search summary. Each matrix was used to generate a score for every possible sequence window in the *E. coli* genomic sequence. The mean (μ_g) and standard deviation (σ_g) of this distribution were used to calculate a genomic *Z*-score ($Z_g(\mu_i) = (\mu_i - \mu_g)/\sigma_g$) for the mean of the footprinted input sites (*x*-axis). $Z_g(\mu_i)$ (indicated by a filled circle) shows where the mean of the inputs (μ_i) lies on the distribution of all possible genomic sites; i.e. how many standard deviations μ_i falls above μ_g. Scores for the footprinted input sites used to build the matrix were normalized with this *Z*-distribution (gray diamonds). The gray stripes depict the $-2\sigma_i$ to $-\sigma_i$ and $+\sigma_i$ to $+2\sigma_i$ ranges of the input site score distribution. Outliers (listed in Table 2) are plotted with filled triangles Sample values for the ranges of small-sample standard deviations (σ_e) for the calculations of the individual scores for the input sites (scaled by the value of the genomic standard deviation): carP (1.77), rhaS (1.68), melR (1.30), purR (0.37–0.43), argR (0.34–0.42), rpoD17 (0.34–0.37), lexA (0.32–0.41), crp (0.26–0.32).

Table 1. Matrix search summary

DNA-binding protein	Search matrix		Sites scoring > μ_i[c]		Sites scoring > $\mu_i - 2\sigma_i$[d]	
	No. sites[a]	%CG[b]	Count	% Non-coding[e]	Count	% Non-coding[e]
Ada	3	38.7	2	100	3	100
AraC	6	40.3	3	100	6	100
ArcA	12	20.0	137	47	1028	38
ArgR	17	22.9	19	76	592	54
CarP	2	46.0	2	100	2	100
Crp	49	37.0	220	81	9097	34
CspA	4	41.3	4	75	109	14
CynR	2	33.3	2	100	2	100
CysB	3	34.2	2	100	3	100
CytR	5	37.8	78	28	3240	23
DeoR	3	22.9	3	33	3	40
DnaA	8	36.7	191	32	548,602	18
FadR	7	47.0	9	73	65	43
FarR	4	20.0	456	38	6526	29
FhlA	3	42.0	1	100	3	100
Fis	19	38.2	205	52	20,228	38
FlhCD	3	53.8	2	100	3	100
Fnr	9	25.8	6	86	445	39
FruR	12	42.7	8	79	113	27
Fur	9	23.4	36	89	501	53
GalR	7	43.8	5	100	17	58
GcvA	4	22.5	3	67	21	71
GlpR	13	39.6	67	26	48,064	19
HipB	4	40.9	2	100	4	100
Hns	15	31.5	5756	28	63,559	23
IclR	2	26.7	2	33	3	33
Ihf	26	32.6	468	81	176,488	47
IlvY	2	32.4	1	100	2	100
LacI	3	44.4	1	100	5	40
LexA	18	33.9	10	100	60	65
Lrp	18	28.9	3287	56	238,622	38
MalT	10	61.0	173	22	637	15
MarR	2	37.5	2	100	2	100
MelR	2	33.3	2	100	2	100
MetJ	15	40.1	16	70	1924	21
MetR	8	33.3	36	37	1497	23
NagC	6	22.5	3	67	121	60
NarL	9	34.6	375	28	8358	25
NtrC	6	51.9	3	100	6	100
OmpR	9	26.7	131	53	11,927	41
OxyR	4	42.3	3	100	4	100
PdhR	3	33.3	2	100	669	26
PhoB	15	29.7	7	100	599	43
PurR	23	45.3	6	82	131	36
RhaS	2	50.0	2	100	2	100
RpoD15	27	37.4	1082	47	32,905	35
RpoD16	48	34.9	945	50	45,334	35
RpoD17	116	37.3	3138	51	138,293	30
RpoD18	34	38.0	394	50	31,666	32
RpoD19	25	38.2	877	43	50,286	30
RpoE	3	38.9	2	100	3	100
RpoH13	8	42.1	4	100	10	100
RpoH14	7	44.3	5	100	11	64
RpoN	7	40.6	2	100	45	36
RpoS17	15	43.0	353	38	46,389	23
RpoS18	7	45.2	14	33	2027	22
SoxS	14	40.8	71	49	15,242	26
TorR	4	37.5	16	44	1047	14
TrpR	4	38.5	3	100	4	100
Tus	6	26.8	3	100	6	83
TyrR	17	33.2	24	64	13,724	28

[a] The number of input sites used to construct the matrix.

[b] %GC content of the matrix.

[c] The number of sites found, and the fraction of these in non-coding regions, above a cutoff set at the mean of the known site scores. Sites with high scores in both the forward and reverse directions are counted only once.

[d] The number of sites found, and the fraction of these in non-coding regions, above a cutoff set at two standard deviations below the mean of the known site scores.

[e] A site was considered to be in a non-coding region if greater than 10% of the bases in the site are contained within a non-coding region.

the matrices with few input sites. (See Figure 1 for a sampling of the range of σ_ε values).

Aberrant sites

To check for anomalous or incorrect sites and sequencing inconsistencies, we checked that all of the *E. coli* inputs were found also as outputs in the genome search. By this method, we detected 39 sequencing inconsistencies between the known footprinted sites and the *E. coli* genome. These inconsistencies are listed on our web site.

We used a cutoff of two standard deviations below (i.e. lower affinity than) the mean score of the inputs for our searches. This assumption of a normal distribution of scores appears to be valid, but a few of the known sites score outside this distribution. Table 2 lists the known sites scoring outside two standard deviations ($\mu_i \pm 2\sigma_i$) of the input mean. RpoD has more input sites than the other matrices in our study and has sites scoring both above $\mu_i + 2\sigma_i$ (sites with unusually high affinity), and below $\mu_i - 2\sigma_i$ (sites with low match to the consensus).

Hns, PurR and SoxS have footprinted sites scoring greater than $\mu_i + 2\sigma_i$. Our analysis predicts that these will be strong or unusual binding sites. The purR autoregulatory site is unusual because it is a two-operator system (consisting of O1 and O2), while every other known gene regulated by PurR has only one operator. Our analysis predicts that O1 is the unusually high-affinity site, in agreement with the experimental observation that PurR binds non-cooperatively to O1 and O2 with a sixfold higher affinity for O1 (Rolfes & Zalkin, 1990).

Table 2. Known sites scoring outside two standard deviations of the known site mean

Matrix[a]	Z_i[b]	Gene[c]
A. *Sites scoring* $< \mu_i - 2\sigma_i$ *(low-affinity sites)*		
TyrR	2.4	tyrB (2)
ArgR (1)	2.1	argR (2)
rpoD15	2.6	narG
rpoD16	2.1	tnaA
rpoD17	2.4	dnaN(5)
rpoD17	2.9	melA
rpoD17	2.4	pnp
rpoD18	2.1	nupG
Fnr	2.1	ndh(2)
MetJ	2.2	metF(5)
B. *Sites scoring* $> \mu_i + 2\sigma_i$ *(high-affinity sites)*		
Hns	−2.3	hns (3)
PurR	−2.0	purR (2)
SoxS	−2.1	micF (2)
rpoD16	−2.2	recA
rpoD19	−2.3	glnL

[a] The search matrix.
[b] Number of standard deviations away from the mean of the known sites, $Z_i(E) = (E - \mu_i)/\sigma_i$, where E is the score for that site.
[c] Identification of the gene from the literature. Numbers in parentheses indicate that the site is one of several upstream of the given target.

Fnr, MetJ, TyrR and ArgR have sites scoring lower than $\mu_i - 2\sigma_i$. Each of these sites is one of multiple sites upstream of the regulated gene in question. Our analysis predicts that these are weak or unusual sites. The autoregulatory argR site is unusual, because the two cooperatively binding ARG boxes are separated by only two base-pairs, whereas in all other known ArgR-regulated genes, the two ARG boxes are separated by three base-pairs (Berg, 1988b). TyrR-mediated repression occurs by cooperative binding at a pair of adjacent TyrR sites with unequal affinity for TyrR. In agreement with experimental observation, our analysis predicts a lower score for the lower-affinity member of the pair of sites for all five TyrR-regulated genes that fit these characteristics, including tyrB (Pittard & Davidson, 1991).

Non-coding *versus* coding discrimination by binding-site matrices

A salient feature of microbial genomes is the dense packing of genetic elements. Greater than 88.6% of the *E. coli* chromosome encodes proteins or stable RNAs (Blattner *et al.*, 1997). While some DNA-binding sites are found in protein-coding regions, most are located in 5′ non-coding sequences. Many of our matrices show a strong tendency for high-scoring sites to be located in non-coding regions (Table 1). Sites scoring greater than μ_i are more likely to be located in non-coding regions than sites scoring greater than $\mu_i - 2\sigma_i$, implying that there are a greater number of false positives in the less restrictive score range between $\mu_i - 2\sigma_i$ and μ_i.

The matrices could be recognizing a trivial difference between coding and non-coding regions. For example, non-coding regions tend to have a low %GC content (Staden, 1984). Two lines of evidence suggest that %GC content cannot fully

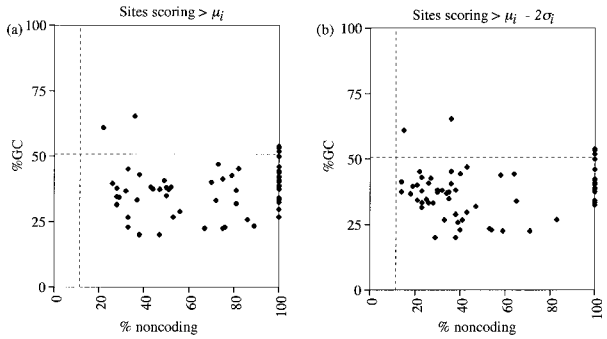

Figure 2. Non-coding preference *versus* matrix %GC content. The percent of sites found in the *E. coli* genome for each matrix which are located in non-coding regions is plotted *versus* the %GC content of the matrix for sites scoring higher than (a) μ_i, and (b) $\mu_i - 2\sigma_i$. The horizontal broken line marks the overall %GC content of the *E. coli* genome (50.8%). The vertical gray line marks the overall percentage non-coding of the *E. coli* genome (11.4%).

explain our results. First, there is only a weak correlation between the %GC content of our matrices and their coding/non-coding discrimination (Figure 2).

Second, shuffling the columns of a matrix, which maintains the %GC content, greatly reduces its coding/non-coding discrimination. We have performed matrix searches for ten different permutations of the crp and lexA search matrices (Table 3). For lexA, column shuffling greatly reduces the number of high-scoring, specific sites found, as well as their non-coding discrimination. For Crp, column shuffling decreases the number of sites scoring higher than μ_i, as well as their non-coding discrimination, but increases the number of sites scoring higher than $\mu_i - 2\sigma_i$. These sites are not specific sites, however (discussed below).

Figures 3 and 4 illustrate the intrinsic preference of the DNA-binding protein matrices for non-coding regions due to their low %GC content. In the non-specific region (low genomic *Z*-scores), there is a general upward slope of the curve for Crp, LexA, ArgR and Fis. Due to their low %GC content, these matrices have a slight intrinsic preference for non-coding regions even in the non-specific score range. Pep is a matrix based on a DNA motif from a protein-coding region, as described in the Materials and Methods. For low genomic *Z*-scores, the curve shows a clear downward slope (Figure 3(d)). Due to its high %GC content (57.5%), the matrix has a small intrinsic preference for coding regions, even in the nonspecific score range. Among the low-scoring, specific sites seen at high *Z*-scores, none of the sites scoring greater than μ_i is found in a non-coding region, and only 2% of those scoring greater than $\mu_i - 2\sigma_i$ are found in non-coding regions for the Pep matrix.

Comparing the two shuffled sites in Figure 4 (Crp shuffle and LexA shuffle) to the corresponding matrices (Crp and LexA), the upward slope of the curve at low *Z*-scores is essentially unchanged, because the shuffled matrices have the same %GC content and these are non-specific sites. However, there is a large reduction in the number of high-scoring sites with high non-coding percentage, between and to the right of the vertical broken lines. This is especially visible for lexA, where there is a large number of high-scoring sites found 100% within non-coding regions only before the columns are shuffled. This is particularly evident in the overall averages, quoted at the bottom of each Figure. Therefore, shuffling the columns of a matrix does not reduce the preference for non-coding regions of low-scoring, non-specific sites, but greatly reduces the preference for high-scoring, specific sites.

Comparison of matrix searches with actual binding-site abundance

A key question is how many binding sites for a particular protein exist in the *E. coli* genome. One

Table 3. Pep and shuffled control matrices

Matrix	Search matrix		Sites scoring > μ_i[c]		Sites scoring > $\mu_i - 2\sigma_i$[d]	
	No. of sites[a]	%CG[b]	Count	% Non-coding[e]	Count	% Non-coding[e]
Pep	10	57.5	9	0	65	2
Crp	49	37.0	220	81	9097	34
Crp shuffle 1	49	37.0	220	28	38,534	26
2			209	30	42,007	25
3			233	35	40,970	25
4			325	28	43,077	26
5			161	42	36,188	28
6			148	33	1506	33
7			113	29	32,918	28
8			123	32	34,985	27
9			235	35	38,522	26
10			113	24	32,836	26
LexA	18	33.9	10	100	60	65
LexA shuffle 1	18	33.9	0	0	9	22
2			0	0	36	28
3			2	50	49	29
4			0	0	21	24
5			1	100	25	32
6			0	0	40	43
7			1	100	39	23
8			0	0	25	36
9			0	0	39	21
10			1	0	85	16

[a] The number of sites used to construct the matrix.
[b] %GC content of the matrix.
[c] The number of sites found, and the fraction of these in non-coding regions, above a cutoff set at the mean of the known site scores. Sites with high scores in both the forward and reverse directions are only counted once.
[d] The number of sites found, and the fraction of these in non-coding regions, above a cutoff set at two standard deviations below the mean of the known site scores.
[e] A site was considered to be in a non-coding region if greater than 10% of the bases in the site are contained within a non-coding region.

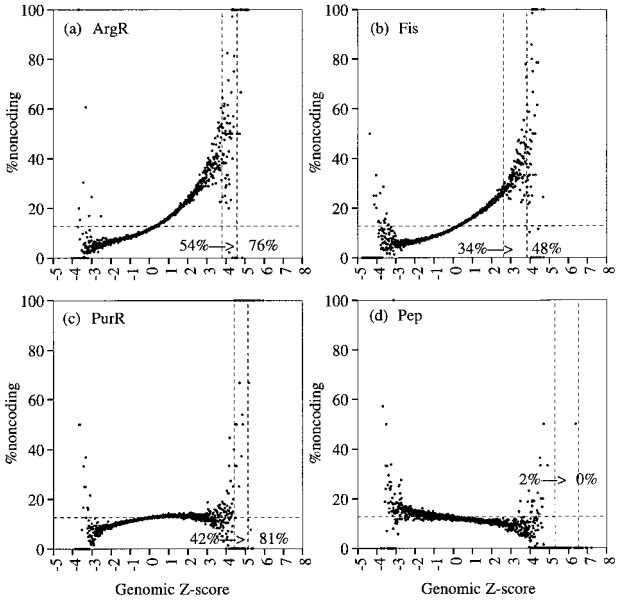

Figure 3. Non-coding preferences for four example matrices. Percentage non-coding *versus* genomic *Z*-score ($Z_g(E) = (E - \mu_g)/\sigma_g$, where E is the score). The percentage non-coding was calculated by computing the fraction of non-coding bases in each site and taking an average over all genomic sites scoring within Z_g increments of 0.01. The horizontal broken line marks the overall fraction of the *E. coli* genome that is non-coding (11.4%). The two vertical lines, from left to right, mark the *Z*-scores of the two cutoffs used, $Z_g(\mu_i)$ and $Z_g(\mu_i - 2\sigma_i)$. This gives an indication of where the footprinted sites fall on the distribution of scores of all possible sequence windows in the genome. Sites to the left of the broken lines are non-specific; sites to the right of the broken lines are high-scoring, specific sites. The numbers at the bottom indicate the percentage non-coding averaged for all sites scoring higher than $\mu_i - 2\sigma_i$ and μ_i (Tables 2 and 3). (a) ArgR; (b) Fis; (c) PurR; (d) Pep control matrix, based on a protein-coding motif from ABC-type transporters.

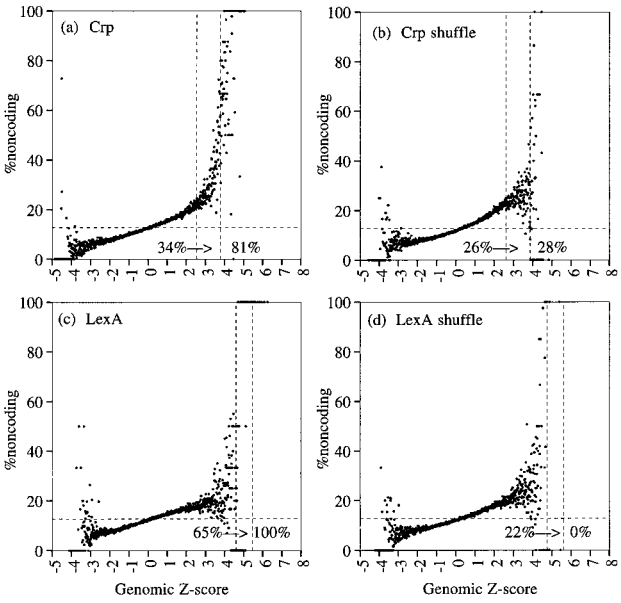

Figure 4. Non-coding preference for shuffled controls. Same notation as Figure 2. (a) Crp; (b) shuffled Crp matrix. (c) LexA; (d) shuffled LexA matrix.

approach for establishing an upper bound is to determine the number of protein molecules in the cell, although it is believed that DNA-binding proteins are present in excess of the number of biologically relevant sites (Berg, 1978, 1988a). We have identified such values from the published literature and from a 2-D electrophoresis survey in our laboratory (Link *et al.*, 1997) for 16 of the proteins in our study (Table 4). Comparison of our predictions to the observed protein abundances suggests that many of our matrices are underspecific; many more sites are predicted than are expected from protein abundance levels. There is no significant correlation between our predictions and the observed abundances. However, there is also no consistency in the conditions under which the experimental measurements in Table 4 were performed.

Looking at the vertical broken lines on the distribution of sites in Figures 3 and 4 gives an idea of the significance of sites scoring above the two cutoffs, and the degree of overestimation of the number of sites in the genome. For both the Crp matrix and the shuffled Crp matrix (Figure 4), there is a large number of sites scoring between μ_i and $\mu_i - 2\sigma_i$ (between the two broken lines), and the percentage non-coding is approximately the same for both the Crp site and the shuffled Crp site. This indicates that these are non-specific sites and the number of hits within this range quoted in Table 1 (9097) is a clear overestimate. The number of sites with a high non-coding fraction that disappear in the shuffled Crp matrix is a better estimate of the number of specific sites: 81% of the sites scoring lower than the mean are non-coding for Crp, whereas only 28% are non-coding for the shuffled Crp matrix.

For some matrices, the cutoff at $\mu_i - 2\sigma_i$ is a better measure of the number of specific sites than for other matrices. For LexA, there is a large difference in the number of sites with a high-percentage non-coding fraction between the two broken lines in Figure 4. This is clear also from looking at the overall averages (100% non-coding for lexA *versus* 0% non-coding for the shuffled lexA matrix for a cutoff set at μ_i, 65% non-coding for lexA *versus* 22% non-coding for the shuffled lexA matrix for a cutoff set at $\mu_i - 2\sigma_i$).

A number of our matrices are clearly not specific. Factors contributing to the number of hits include the length of the input sites and the number of sites used to build the matrix. Matrices with long input sites tend to have high specificity. An example is RhaS, which has an unusually high genomic *Z*-score in Figure 1. To construct this matrix, we used two examples of a motif contain-

Table 4. *E. coli* DNA-binding protein abundances

Protein	Copies/cell[a]	Reference	Sites > μ_i[b]	Sites > $\mu_i - 2\sigma_i$[c]
ArcA	200	Link *et al.* (1997)	4	397
ArgR	330-510	Maas (1994)	19	592
Crp	1300	Anderson *et al.* (1971)	220	9097
DnaA	330	Hansen *et al.* (1991)	191	548,602
FhlA	360	Hopper *et al.* (1994)	1	31
Fis	100–50,000	Ball *et al.* (1992)	205	20,228
Hns	800	Link *et al.* (1997)	5756	63,559
Ihf	17,000–34,000	Ditto *et al.* (1994)	468	176,488
LacI	10	Gilbert & Muller-Hill (1966)	1	5
LexA	200–4000	Dri & Moreau (1994)	10	60
Lrp	6000	Willins *et al.* (1991)	3287	238,622
RpoD	500–700	Jishage & Ishihama (1995)	6436	298,484
RpoH	650	Straus *et al.* (1987)	9	21
RpoN	110	Jishage *et al.* (1996)	4	237
RpoS	170–230	Jishage & Ishihama (1995)	371	48,284
TrpR	120–375	Gunsalus *et al.* (1986)	3	4

[a] Monomer concentration.
[b] The number of sites scoring above a cutoff set at the mean of the known site scores.
[c] The number of sites scoring above a cutoff set at two standard deviations below the mean of the known site scores.

ing two 17 bp half-sites and the intervening 16 bp spacer, for a total motif length of 50 bp. When we align the four 17 bp conserved half-sites, ignoring the 16 bp spacer, and use this as our search matrix, we obtain a much lower *Z*-score and a much larger number of sites scoring above the cutoff.

Matrices with short input sites tend to have lower specificity and thus obtain a large number of hits. This is an issue for several of the proteins in our study that recognize short DNA motifs, such as DnaA. Many of these proteins tend to have multiple binding sites in one promoter region. If two or more sites are separated by a consistent spacing, then we can construct our matrices to include more than one motif, thus increasing the specificity of the matrix. For example, our matrix for ompR includes two tandemly repeated motifs, and many of our matrices contain two palindromic half-sites separated by a spacer.

However, if the motifs are separated by variable spacing, two different approaches can be used. If there are several distinct spacing classes, each spacing class can be treated separately. We have done this for several classes of spacing between promoter elements for rpoD, rpoS and rpoH. If spacing varies widely, this approach is impossible and we must use a single motif as the search element. After performing the search, we can examine the spacing between pairs of short motifs to filter out candidate binding sites with appropriate spacing. We can also look at the spacing between sites for pairs of different, interacting DNA-binding proteins. In a recent study of regulatory proteins in skeletal muscle, large numbers of false positives were filtered by looking for clusters of sites, since experimental data suggest that muscle expression requires multiple binding sites for these factors in close proximity (Wasserman & Fickett, 1998).

Several of our matrices with large numbers of hits are simply proteins that bind to multiple short motifs separated by variable spacings (0–100 base-pairs) and/or with variable orientations. This includes DnaA, FarR, GlpR, MalT, Lrp, MetR, NarL, TyrR, TorR and TreR. Since we use only one motif in our search, there is not sufficient specificity encoded in the matrix. We are effectively searching for only a portion of the full binding site. However, if we filtered out only those pairs of sites that are in proximity, we would obtain a much smaller set of candidate sites.

Proteins such as Hns and Ihf are believed to have loose sequence specificity. These proteins have a role in maintaining chromosome structure and DNA supercoiling. Their known binding sites are not highly conserved and thus we obtain many hits in our matrix search. Several other proteins, such as Crp and Lrp, have a more ''global'' regulatory role and a larger number of specific binding sites in the genome is expected.

Comparison of natural and SELEX sites

An increasingly common means for determining the binding site of a DNA-binding protein is to use cycles of binding and amplification to extract sites from pools of random sequence, a technique called SELEX, or Systematic Evolution of Ligands by Exponential enrichment (Tuerk & Gold, 1990). Such analyses have been reported for eight DNA-binding proteins from *E. coli*: FadR (Gui *et al.*, 1996), FruR (Negre *et al.*, 1996), IclR (Pan *et al.*, 1996), Lrp (Cui *et al.*, 1995), MetJ (He *et al.*, 1996), OmpR (Harlocker *et al.*, 1995), OxyR (Toledano *et al.*, 1994) and TrpR (Czernik *et al.*, 1994).

An important question is whether the results of SELEX experiments are consistent with natural sites. We have approached this by examining aligned information content curves (Schneider *et al.*, 1986) and sequence logos (Schneider & Stephens, 1990) for natural and SELEX data for these eight proteins (Figures 5 and 6).

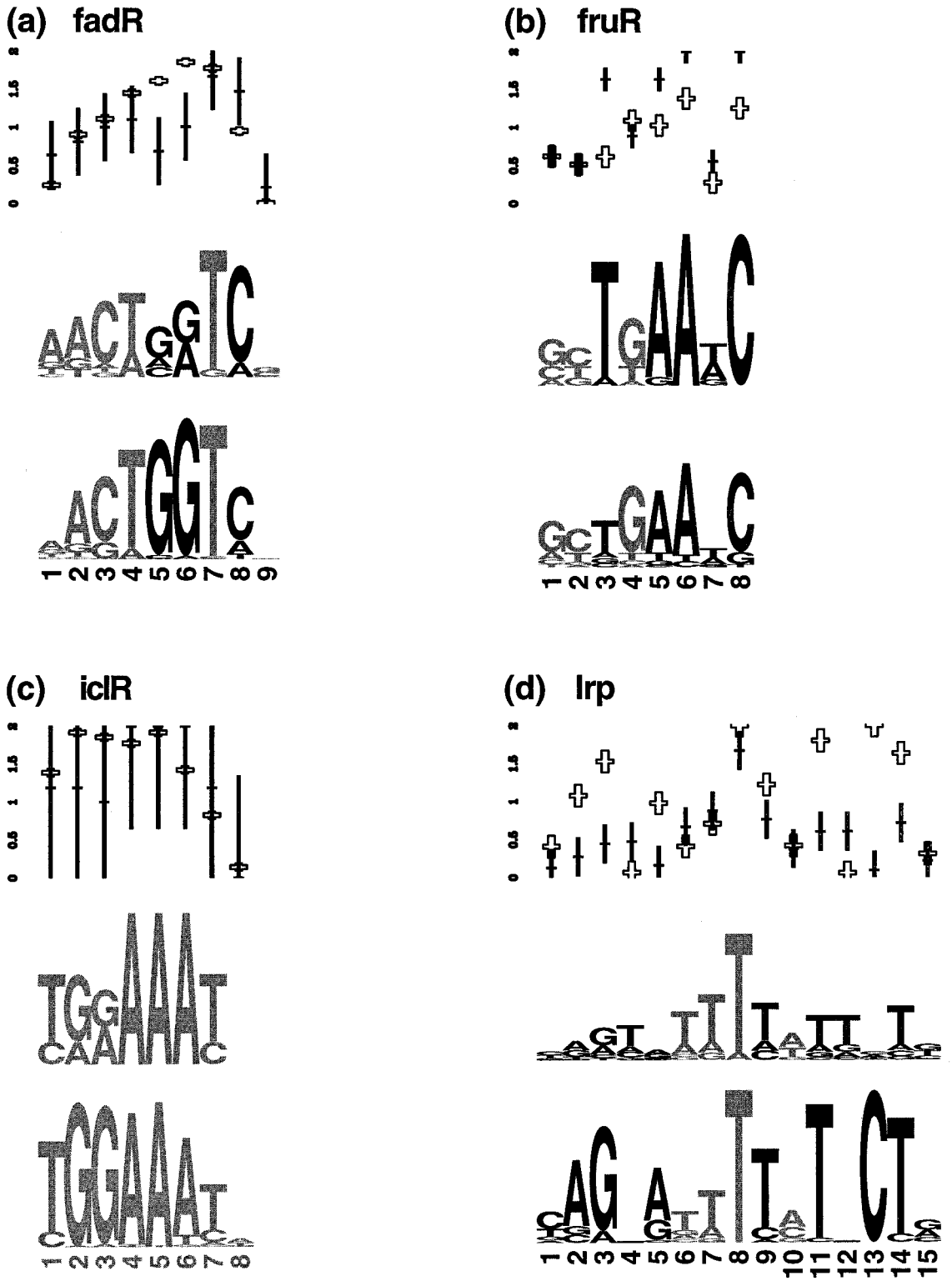

Figure 5. Comparison of information content of natural and SELEX sites. For each dataset, the top panel shows the information content curves (Schneider *et al.*, 1986) for the natural site (filled bars) and SELEX (open bars) datasets, the middle panel shows the sequence logo (Schneider & Stephens, 1990) for the natural sites, and the lower panel shows the sequence logo for the SELEX data. The horizontal portion of each bar marks the measured information content, and the vertical bars mark the estimate for the two standard deviation confidence interval (Schneider *et al.*, 1986). In order to enable comparison, the sample size correction is not applied to the information content calculation. The sequence logos are printed in black at positions where the information content measure for the SELEX data is outside the two standard deviation estimate for the information content of the natural sites. (a) FadR; (b) FruR; (c) IclR; (d) Lrp.

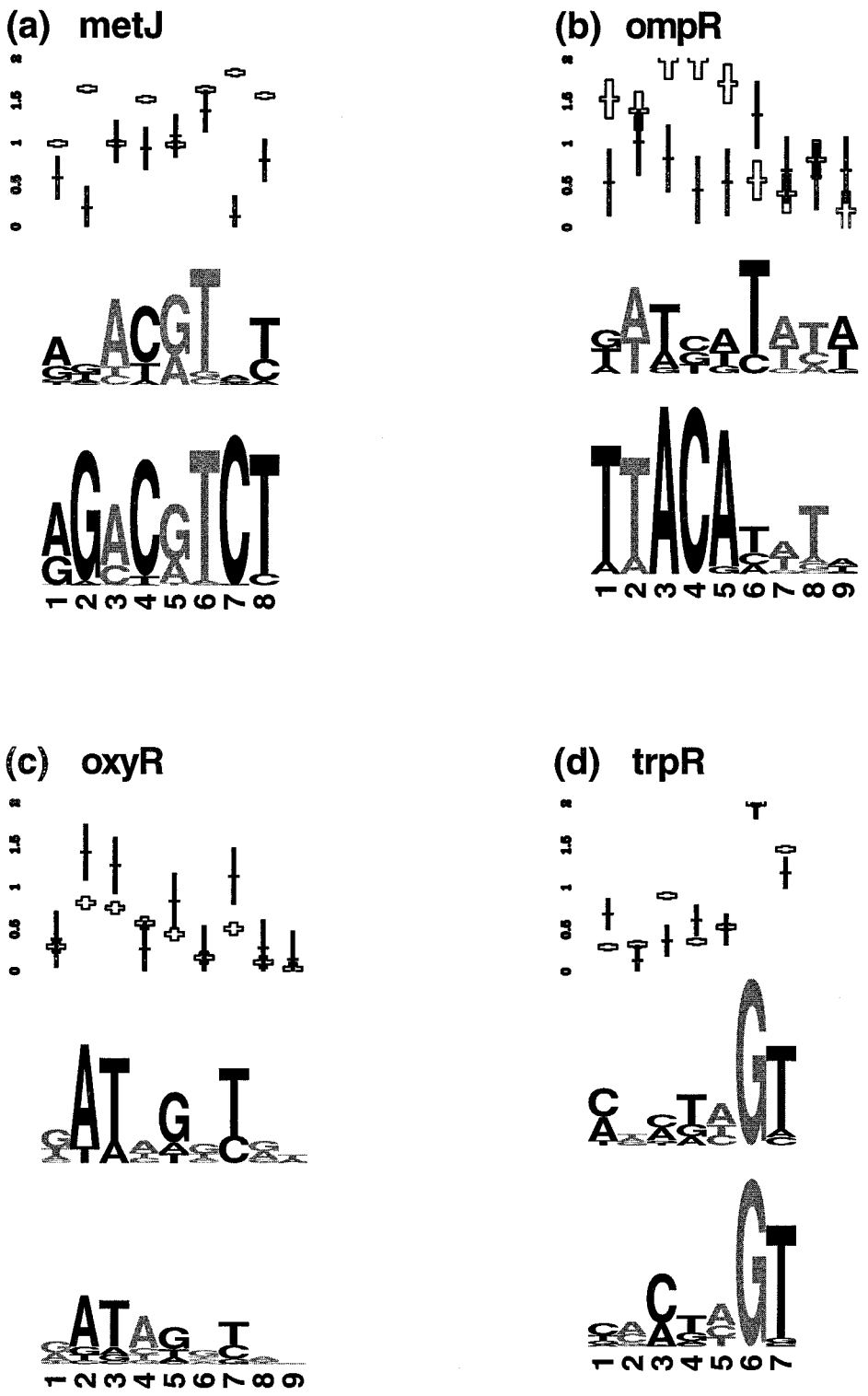

Figure 6. Comparison of information content of natural and SELEX sites. Same notation as Figure 5. (a) MetJ; (b) OmpR; (c) OxyR; (d) TrpR.

For IclR (Figure 5(c)), only two natural sites are available, and even after symmetrizing the matrix the imprecision in the information estimate (Schneider *et al*., 1986) is too great to allow comparison. Significant deviations are observed for each of the other seven proteins (Figures 5 and 6). While in most cases the difference is only in the magnitude of the bases, in some cases the ordering of bases is significantly different. As noted by Wild *et al*. (1996), natural MetJ sites contain a C at position 7 far less frequently than would be expected from the SELEX data, and is in fact slightly less frequent than A (Figure 6(a)).

Even more striking is the difference for the OmpR data; with strong deviations at the first three positions (Figure 6(b)). Natural sites contain a G at position 1 in 40% of cases, but only T and A were observed here in the SELEX experiment. The most prominent bases at positions 2 and 3 are A and T, respectively, for the natural ompR sites, and T and A, respectively, for the SELEX sites. An absolute preference for C at position 4 of the SELEX data is not maintained in the natural sequences.

Striking differences are also observed between the natural and SELEX data for Lrp. An imperfect palindromic symmetry with the consensus cAG-at—at-CTg is clearly present in the SELEX sites and is highlighted by the prominent AG and CT dinucleotides. A similar prominent palindromic component is not observed for the natural sites. Furthermore, the absolute selection for C at position 13 of the site contrasts with the weak conservation of position 13 in the natural sites (Figure 5(d)).

Why are so many discrepancies observed between the natural binding-site data and SELEX data? One possibility is an artifact due to the design of the SELEX experiment. While most SELEX experiments are based on an oligonucleotide pool containing a large randomized region flanked by constant but non-binding sequences, the OmpR experiment used random sequence anchored by a partial binding site (Harlocker *et al*., 1995). This constant region may have influenced the selection of bases. Another possibility is that the set of available natural sites is itself biased and is not a representative sample. For example, a sample biased towards high-affinity sites might underestimate the variability in binding sites.

Another likely possibility is that the SELEX data faithfully represent the binding preferences of the protein in isolation, but that the biological sites are under additional constraints. For example, functional phage T7 promoters selected from random sequence lack a sequence element found in natural T7 promoters; presumably this is the binding-site for a second protein (Schneider & Stormo, 1989). The discrepancies between SELEX and natural binding site data for MetJ have been interpreted as the result of evolutionary pressure to avoid binding by TrpR and constraints imposed by the tandem repeat nature of natural MetJ sites (Wild *et al*., 1996).

Such potential cross-talk is likely to be common. Many of these proteins belong to large families, with multiple members of the family present in *E. coli*. For example, FruR is one of 13 proteins in *E. coli* K12 belonging to the LacI family. Many LacI family members have closely related binding specificities (Lehming *et al*., 1990; Schumacher *et al*., 1994). Such selection need not be only against cross-talk; some regulatory systems may use such cross-regulation. For example, the *E. coli* proteins MarA, Rob and SoxS all target genes involved in the oxidative stress response and bind to closely related DNA sequences. However, some targets respond differently from these proteins (Ariza *et al*., 1995; Jair *et al*., 1995; Li & Demple, 1996).

Another form of potential cross-talk is between *E. coli* methylation systems and DNA binding proteins. The methylases Dam and Dcm methylate the sequences GATC and CCWGG, respectively (Palmer & Marinus, 1994). Some genomic sites with these sequences are unmethylated, presumably due to a bound DNA-binding protein blocking access by the cognate methylase (Hale *et al*., 1994; Ringquist & Smith, 1992; Wang & Church, 1992). Conversely, the methylation status of DNA-binding sites can affect the binding of proteins to these sites (Bolker & Kahmann, 1989; Braun & Wright, 1986; Charlier *et al*., 1995; van der Woude *et al*., 1992; Yin *et al*., 1988).

Examination of sequence logos for the two sets of FadR sites (Figure 5(a)) reveals that positions 5 through 8 of the half-site could easily specify a Dam site. A at position 6 is almost as common as G in the natural sites, whereas the G at position 6 is nearly invariant in the SELEX data. Indeed, two of the 14 natural half-sites (from two different sites) contain GATC in these positions. Similarly, the logo for the natural OmpR sites (Figure 6(b)) suggests that GATC could occur at positions 1 through 4 or 4 through 7. One of each such GATC sites is observed in a natural site. In contrast, the SELEX data are incompatible with a GATC site at either location (though one could be expected at positions 5 through 8 at a frequency of about one in 5000).

Conclusions

We have built a collection of search matrices from alignments of available experimental binding-site data for *E. coli* DNA-binding proteins. Calibration of these matrices against the sites used to build the matrices assists in identifying anomalous sites, which may be misaligned, incorrect, or unusual sites. These matrices show a sharp preference for the minority of *E. coli* DNA that does not encode proteins. This preference appears to be due to more than just a preference for base composition, as there is little correlation between non-coding preference and matrix %GC content, and because shuffling the columns of a matrix reduces its preference for non-coding regions.

The comparison of the number of sites predicted by our matrices *versus* the known abundance of these DNA-binding proteins suggests that our matrices are often underspecific, as the matrices predict far more sites than could be bound by protein. However, the paucity of published cellular abundance values for *E. coli* DNA-binding proteins prevents exhaustive analysis.

Our analysis of the available SELEX data suggests that such data should be treated with caution, as it may give a distorted view of the binding specificity of a DNA-binding protein. However,

comparisons between natural and SELEX-derived sites may both reveal such discrepancies and suggest the additional selective pressures that shape natural sites.

Materials & Methods

Matrix construction

Binding sites identified by biochemical footprinting or SELEX were used to construct two separate sets of binding-site matrices. Sites were obtained from the DPInteract database (http://arep.med.harvard.edu/dpinteract) and other databases (Huerta *et al*., 1998). The sites were aligned either based on published alignments, using the CLUSTALW multiple alignment algorithm (Higgins *et al*., 1991), or using the Gibbs sampler algorithm (Lawrence *et al*., 1993) to identify conserved motifs. The frequency of each base at each position in the site was used to build a matrix. We used only sites from *E. coli*. To determine the length of the sites used to construct the matrices, we chose only the conserved regions out of each set of aligned footprinted sites.

Control matrices

We constructed a control matrix based on a motif from a protein-coding region. The Pep control matrix was based on a motif from ABC-type transporters (Prosite entry PDOC00185). An amino acid sequence corresponding to this motif (GAGKSTLL) was back-translated using an *E. coli* codon usage table and the BACKTRANSLATE program in the GCG package (Devereux *et al*., 1984). This sequence was then used to search the *E. coli* dataset with BLASTN, and the top ten matches were used to generate the Pep matrix.

Since non-coding regions tend to have lower %GC content than non-coding regions, we also generated another set of control matrices that account for this, by ''shuffling'' the columns from one of the actual binding-site search matrices. This maintains the %GC content. We generated shuffled versions of the crp and lexA matrices. A vector with the integers from 1 to the motif width was created, and then shuffled randomly. This vector was then used to shuffle each of the sites in the matrix. If the first position of the shuffling vector contained ''ten'', then the first nucleotide of site A would become the tenth nucleotide of the shuffled site A, the first nucleotide of site B would become the tenth nucleotide of the shuffled site B, etc. These sites were then treated as before. The result of this is to generate a matrix that looks like a columnwise shuffle of the original matrix. The positive control score distribution for the shuffled matrix will look exactly like that of the original matrix, as each shuffled site has a score with the shuffled matrix identical with the corresponding original site with the original matrix.

Searches over the *E. coli* genome

For searches, we used the program ScanACE (Roth *et al*., 1998), which is available on our web site (http://arep.med.harvard.edu). This program uses the log transformation described by Berg & von Hippel (1987). Both strands of the genome are searched. Near-symmetric sites with high scores in both the forward and reverse direction are counted only once in our analysis, and the higher of the two scores is used. The *E. coli* sequence was obtained from GenBank entry U00096.

Acknowledgments

The authors thank Mark Poritz, John Aach, Martha Bulyk, Dereth Phillips, Laura Richterich, Fritz Roth, Jason Hughes, Ed Lin and Peter deWulf for help and discussions. A.M.M. is a Howard Hughes Medical Institute Predoctoral Fellow. This work was funded by DOE grant DE-FG02-87ER60565 and the Lipper Foundation.

References

Anderson, W. B., Schneider, A. B., Emmer, M., Perlman, R. L. & Pastan, I. (1971). Purification of and properties of the cyclic adenosine 3′-5′-monophosphate receptor protein which mediates cyclic adenosine 3′-5′ monophosphate-dependent gene transcription in *Escherichia coli*. *J. Biol. Chem.* **246**, 5929–5937.

Ariza, R. R., Li, Z., Ringstad, N. & Demple, B. (1995). Activation of multiple-antibiotic resistance and binding of stress-inducible promoters by *Escherichia coli* Rob protein. *J. Bacteriol.* **177**, 1655–1661.

Ball, C. A., Osuna, R., Ferguson, K. C. & Johnson, R. C. (1992). Dramatic changes in Fis levels upon nutrient upshift in *Escherichia coli*. *J. Bacteriol.* **174**, 8043–8056.

Berg, O. G. (1978). A model for the statistical fluctuations of protein numbers in a microbial population. *J. Theoret. Biol.* **71**, 587–603.

Berg, O. G. (1988a). Selection of DNA binding sites by regulatory proteins. Functional specificity and pseudosite competition. *J. Biomol. Struct. Dynam.* **6**, 275–97.

Berg, O. G. (1988b). Selection of DNA binding sites by regulatory proteins: the LexA protein and the arginine repressor use different strategies for functional specificity. *Nucl. Acids Res.* **16**, 5089–5105.

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

Berg, O. G. & von Hippel, P. H. (1988). Selection of DNA binding sites by regulatory proteins. II. The binding specificity of cyclic AMP receptor protein to recognition sites. *J. Mol. Biol.* **200**, 709–723.

Blattner, R. F., Plunkett, G. I., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A. & Rose, D. J. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1462.

Bolker, M. & Kahmann, R. (1989). The *Escherichia coli* regulatory protein OxyR discriminates between methylated and unmethylated states of the phage Mu mom promoter. *EMBO J.* **8**, 2403–2410.

Braun, R. E. & Wright, A. (1986). DNA methylation differentially enhances the expression of one of the two *E. coli dnaA* promoters *in vivo* and *in vitro*. *Mol. Gen. Genet.* **202**, 246–250.

Charlier, D., Hassanzadeh, Gh G., Kholti, A., Gigot, D., Pierard, A. & Glansdorff, N. (1995). carP, involved in pyrimidine regulation of the *Escherichia coli* carbamoylphosphate synthetase operon encodes a sequence-specific DNA-binding protein identical to

XerB and PepA, also required for resolution of ColEI multimers. *J. Mol. Biol.* **250**, 392–406.
Cui, Y., Wang, Q., Stormo, G. D. & Calvo, J. M. (1995). A consensus sequence for binding of Lrp to DNA. *J. Bacteriol.* **177**, 4872–4880.
Czernik, P. J., Shin, D. S. & Hurlburt, B. K. (1994). Functional selection and characterization of DNA binding sites for trp repressor of *Escherichia coli*. *J. Biol. Chem.* **269**, 27869–27875.
Devereux, J., Haeberli, P. & Smithies, O. (1984). A comprehensive set of sequence analysis programs for the VAX. *Nucl. Acids Res.* **12**, 387–395.
Ditto, M. D., Roberts, D. & Weisberg, R. A. (1994). Growth phase variation of integration host factor level in *Escherichia coli*. *J. Bacteriol.* **176**, 3738–3748.
Dri, A. M. & Moreau, P. L. (1994). Control of the LexA regulon by pH; evidence for a reversible inactivation of the LexA repressor during the growth cycle of *Escherichia coli*. *Mol. Microbiol.* **12**, 621–629.
Fondrat, C. & Kalogeropoulos, A. (1996). Approaching the function of new genes by detection of their potential upstream activation sequences in *Saccharomyces cerevisiae*: application to chromosome III. *Comput. Appl. Biosci.* **12**, 363–374.
Gilbert, W. & Muller-Hill, B. (1966). Isolation of the lac repressor. *Proc. Natonl Acad. Sci. USA*, **56**, 1891–1898.
Goodrich, J. A., Schwartz, M. L. & 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). *Nucl. Acids Res.* **18**, 4993–5000.
Grahn, A. M., Bamford, J. K., O'Neill, M. C. & Bamford, D. H. (1994). Functional organization of the bacteriophage PRD1 genome. *J. Bacteriol.* **176**, 3062–3068.
Gross, C. A. (1996). Function and regulation of the heat shock proteins. In Escherichia coli *and* Salmonella: *Molecular and Cellular Biology* (Neidhardt, F. C., ed.), 2nd edit., vol. 1, ASM Press, Washington, DC.
Gui, L., Sunnarborg, A. & LaPorte, D. C. (1996). Regulated expression of a repressor protein: FadR activates IclR. *J. Bacteriol.* **178**, 4704–4709.
Gunsalus, R. P., Miguel, A. G. & Gunsalus, G. L. (1986). Intracellular Trp repressor levels in *Escherichia coli*. *J. Bacteriol.* **167**, 272–278.
Hale, W. B., van der Woude, M. W. & 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. Bacteriol.* **176**, 3438–3441.
Hansen, F. G., Atlung, T., Braun, R. E., Wright, A., Hughes, P. & Kohiyama, M. (1991). Initiator (DnaA) protein concentration as a function of growth rate in *Escherichia coli* and *Salmonella typhimurium*. *J. Bacteriol.* **173**, 5194–5199.
Harlocker, S. L., Bergstrom, L. & Inouye, M. (1995). Tandem binding of six OmpR proteins to the ompF upstream regulatory sequence of *Escherichia coli*. *J. Biol. Chem.* **270**, 26849–26856.
He, Y. Y., Stockley, P. G. & Gold, L. (1996). *In vitro* evolution of the DNA binding sites of *Escherichia coli* methionine repressor, MetJ. *J. Mol. Biol.* **255**, 55–56.
Higgins, D. G., Bleasby, A. J. & Fuchs, R. (1991). CLUSTALV: improved software for multiple sequence alignment. *Comput. Appl. Biosci.* **8**, 189–191.
Hopper, S., Babst, M., Schlensog, V., Fischer, H. M., Hennecke, H. & Bock, A. (1994). Regulated expression *in vitro* of genes coding for formate hydrogenlyase components of *Escherichia coli*. *J. Biol. Chem.* **269**, 19597–19604.
Horton, P. B. & Kanehisa, M. (1992). An assessment of neural network and statistical approaches for prediction of *E. coli* promoter sites. *Nucl. Acids Res.* **20**, 4331–4338.
Huerta, A. M., Salgado, H., Thieffry, D. & Collado-Vides, J. (1998). RegulonDB: a database on transcriptional regulation in *Escherichia coli*. *Nucl. Acids Res.* **26**, 55–59.
Jair, K. W., Martin, R. G., Rosner, J. L., Fujita, N., Ishihama, A. & Wolf, R. E. (1995). Purification and regulatory properties of MarA protein, a transcriptional activator of *Escherichia coli* multiple antibiotic and superoxide resistance promoters. *J. Bacteriol.* **177**, 7100–7104.
Jishage, M. & Ishihama, A. (1995). Regulation of RNA polymerase sigma subunit synthesis in *Escherichia coli*: intracellular levels of sigma 70 and sigma 38. *J. Bacteriol.* **177**, 6832–6835.
Jishage, M., Iwata, A., Ueda, S. & Ishihama, A. (1996). Regulation of RNA polymerase sigma subunit synthesis in *Escherichia coli*: intracellular levels of four species of sigma subunit under various growth conditions. *J. Bacteriol.* **178**, 5447–5451.
Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F. & Wootton, J. C. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.
Lehming, N., Sartorius, J., Kisters-Woike, B., von Wilcken-Bergmann, B. & Muller-Hill, B. (1990). Mutant lac repressors with new specificities hint at rules for protein-DNA recognition. *EMBO J.* **9**, 615–621.
Lewis, L. K., Harlow, G. R., Gregg-Jolly, L. A. & Mount, D. W. (1994). Identification of high affinity binding sites for LexA which define new DNA damage-inducible genes in *Escherichia coli*. *J. Mol. Biol.* **241**, 507–523.
Li, Z. & Demple, B. (1996). Sequence specificity for DNA binding by *Escherichia coli* SoxS and Rob proteins. *Mol. Microbiol.* **20**, 937–945.
Link, A., Robison, K. & Church, G. M. (1997). Comparing the predicted and observed properties of proteins encoded in the genome of *Escherichia coli*. *Electrophoresis*, **18**, 1259–1313.
Lisser, S. & Margalit, H. (1993). Compilation of *E. coli* mRNA promoter sequences. *Nucl. Acids Res.* **21**, 1507–1516.
Maas, W. K. (1994). The arginine repressor of *Escherichia coli*. *Microbiol. Rev.* **58**, 631–640.
Negre, D., Bondon-Bidaud, C., Geourjon, C., Deleage, G., Cozzone, A. J. & Cortay, J. C. (1996). Definition of a consensus DNA-binding site for the *Escherichia coli* pleiotropic regulatory protein, FruR. *Mol. Microbiol.* **21**, 257–266.
O'Neill, M. C. (1989). Consensus methods for finding and ranking DNA binding sites: application to *E. coli* promoters. *J. Mol. Biol.* **207**, 301–310.
Palmer, B. R. & Marinus, M. G. (1994). The dam and dcm strains of *Escherichia coli* – a review. *Gene*, **143**, 1–12.
Pan, B., Unnikrishnan, I. & LaPorte, D. C. (1996). The binding site of the IclR repressor protein overlaps the promoter of aceBAK. *J. Bacteriol.* **178**, 3982–3984.

Pittard, A. J. & Davidson, B. E. (1991). TyrR protein of *Escherichia coli* and its role as respressor and activator. *Mol. Microbiol.* **5**, 1595–1592.

Ramseier, T. M., Bledig, S., Michotey, V., Feghali, R. & Saier, M. H., Jr (1995). The global regulatory protein FruR modulates the direction of carbon flow in *Escherichia coli*. *Mol. Microbiol.* **16**, 1157–1169.

Ringquist, S. & Smith, C. L. (1992). The *Escherichia coli* chromosome contains specific, unmethylated dam and dcm sites. *Proc. Natl. Acad. Sci. USA*, **89**, 4539–4543.

Robison, K. (1997). Whole genome computational analyses of DNA-protein recognition networks, PhD thesis, Harvard University.

Rolfes, R. J. & Zalkin, H. (1990). Autoregulation of *Escherichia coli* purR requires two control sites downstream of the promoter. *J. Bacteriol.* **172**, 5758–5766.

Rosenblueth, D. A., Thieffry, D., Huerta, A. M., H., S. & 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. & Church, G. M. (1998). Revealing regulons by whole-genome expression monitoring and upstream sequence alignment. *Nature Biotechnol.* In the press.

Schneider, T. D. & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequence. *Nucl. Acids Res.* **18**, 6097–6100.

Schneider, T. D. & Stormo, G. D. (1989). Excess information at bacteriophage T7 genomic promoters detected by a random cloning technique. *Nucl. Acids Res.* **17**, 659–674.

Schneider, T. D., Stormo, G. D., Gold, L. & Ehrenfeucht, A. (1986). Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* **188**, 415–431.

Schumacher, M. A., Choi, K. Y., Zalkin, H. & Brennan, R. G. (1994). Crystal structure of lacI member, purR, bound to DNA – minor groove binding by alpha helices. *Science*, **266**, 763–770.

Staden, R. (1984). Measurements of the effects that coding for a protein has on a DNA sequence and their use for finding genes. *Nucl. Acids Res.* **12**, 551–567.

Stormo, G. D. (1990). Consensus patterns in DNA. *Methods Enzymol.* **183**, 211–21.

Straus, D. B., Walter, W. A. & Gross, C. A. (1987). The heat shock response of *E. coli* is regulated by changes in the concentration of sigma 32. *Nature*, **329**, 348–351.

Thieffry, D., Salgado, H., Huerta, A. M. & Collado-Vides, J. (1998). Prediction of transcriptional regulatory sites in the complete genome sequence of *Escherichia coli* K-12. *Bioinformatics*, **14**, 391–400.

Toledano, M. B., Kullik, I., Trinh, F., Baird, P. T., Schneider, T. D. & Storz, G. (1994). Redox-dependent shift of OxyR-DNA contacts along an extended DNA-binding site: a mechanism for differential promoter selection. *Cell*, **78**, 897–909.

Tronche, F., Ringeisen, F., Blumenfeld, M., Yaniv, M. & Pontoglio, M. (1997). Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome. *J. Mol. Biol.* **266**, 231–245.

Tuerk, C. & Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, **249**, 505–510.

van der Woude, M. W., Braaten, B. A. & Low, D. A. (1992). Evidence for global regulatory control of pilus expression in *Escherichia coli* by Lrp and DNA methylation: model building based on analysis of pap. *Mol. Microbiol.* **6**, 2429–2435.

Verbeek, H., Nilsson, L., Baliko, G. & Bosch, L. (1990). Potential binding sites of the trans-activator FIS are present upstream of all rRNA operons and of many but not all tRNA operons. *Biochim. Biophys. Acta*, **1050**, 302–306.

Wang, M. X. & Church, G. M. (1992). A whole genome approach to *in vivo* DNA-protein interactions in *E. coli*. *Nature*, **360**, 606–610.

Wasserman, W. W. & Fickett, J. W. (1998). Identification of regulatory regions which confer muscle-specific gene regulation. *J. Mol. Biol.* **278**, 167–181.

Wild, C. M., McNally, T., Phillips, S. E. V. & Stockley, P. G. (1996). Effects of systematic variation of the minimal *Escherichia coli* met consensus operator site: *in vivo* and *in vitro* met repressor binding. *Mol. Microbiol.* **21**, 1125–1135.

Willins, D. A., Ryan, C. W., Platko, J. V. & Calvo, J. M. (1991). Characterization of Lrp, and *Escherichia coli* regulatory protein that mediates a global response to leucine. *J. Biol. Chem.* **266**, 10768–10774.

Wise, A., Brems, P., Ramkrishnan, V. & Villarejo, M. (1996). Sequences in the −35 region of *Escherichia coli* rpoS-dependent genes promote transcription by E sigma S. *J. Bacteriol.* **178**, 2785–2793.

Yin, J. C., Krebs, M. P. & Reznikoff, W. S. (1988). Effect of dam methylation on Tn*5* transposition. *J. Mol. Biol.* **199**, 35–45.

Edited by R. Ebright

(Received 28 May 1998; received in revised form 4 August 1998; accepted 24 August 1998)