4994–5002 *Nucleic Acids Research, 1997, Vol. 25, No. 24*

Information analysis of Fis binding sites

Paul N. Hengen[1], Stacy L. Bartram[1,2,+], Lisa E. Stewart[1] and Thomas D. Schneider[1,*]

[1]Laboratory of Mathematical Biology, National Cancer Institute, Frederick Cancer Research and Development Center, PO Box B, Building 469, Room 144, Frederick, MD 21702-1201, USA and [2]Middletown High School, 200 High Street, Middletown, MD 21769, USA

Received August 25, 1997; Revised and Accepted October 30, 1997

ABSTRACT

Originally discovered in the bacteriophage Mu DNA inversion system *gin*, Fis (Factor for Inversion Stimulation) regulates many genetic systems. To determine the base frequency conservation required for Fis to locate its binding sites, we collected a set of 60 experimentally defined wild-type Fis DNA binding sequences. The sequence logo for Fis binding sites showed the significance and likely kinds of base contacts, and these are consistent with available experimental data. Scanning with an information theory based weight matrix within *fis*, *nrd*, *tgt*/*sec* and *gin* revealed Fis sites not previously identified, but for which there are published footprinting and biochemical data. DNA mobility shift experiments showed that a site predicted to be 11 bases from the proximal *Salmonella typhimurium hin* site and a site predicted to be 7 bases from the proximal P1 *cin* site are bound by Fis *in vitro*. Two predicted sites separated by 11 bp found within the *nrd* promoter region, and one in the *tgt*/*sec* promoter, were also confirmed by gel shift analysis. A sequence in *aldB* previously reported to be a Fis site, for which information theory predicts no site, did not shift. These results demonstrate that information analysis is useful for predicting Fis DNA binding.

INTRODUCTION

Fis is a pleiotropic DNA-bending protein that enhances site-specific recombination, controls DNA replication, and regulates transcription of a number of genes in *Escherichia coli* and *Salmonella typhimurium* (1–4). Fis is composed of two 98 amino acid polypeptides, with each polypeptide having four α helices, A–D. Homodimers of Fis bind to and deflect DNA from 40° to 90° (3,5–7). Mutational analyses suggest that the N-terminal portion of the Fis monomer containing the A helix is necessary for recombination, while the C-terminal portion containing the D helix is thought to be involved in DNA binding (8,9). However, X-ray crystal structures of Fis reveal that the D helices appear to be too close together for Fis to fit into two successive major grooves on straight B-form DNA, suggesting that the DNA bends to accommodate Fis (6,10,11), that Fis is flexible, or that Fis binds in a completely unanticipated manner.

Although Fis binds to precise sequences according to footprint data, it is often noted that the Fis binding site has a poorly defined consensus sequence (3,6,8–25). Consensus sequences are often used to locate binding sites, but it is widely known that this does not work well, especially in the case of Fis (24–26). Since a consensus is constructed by selecting the most frequent base or bases at every position across a binding site, creating a consensus sequence throws out important information about the observed frequency of bases in the binding site.

In contrast to consensus sequences, information theory provides quantitative models for binding sites. These models are represented by the sequence logo, a graphical method that retains most of the subtleties in sequence data (27–30). Even a glance at a sequence logo often reveals the possible nature of specific base contacts, which side of a base pair is likely to face the protein, and whether or not the DNA is distorted away from B-form (30,31).

Because of its importance in a variety of genetic systems and because many binding sites were already well defined, interaction of Fis with DNA was an attractive candidate for a thorough information analysis. In addition to the information content measure (32) and the sequence logo (29), we used a new method, 'individual information' (R_i), that defines the information content of individual binding sites (33) and displays the results graphically as a 'sequence walker' (34). These methods have an advantage over other methods in that training is not required to obtain a quantitative binding site model, and only examples of functional sites are used to construct the model. Using well-defined biochemical data helps to ensure that the models are realistic. Our analysis of Fis binding sites and their surrounding sequences revealed many previously unidentified sites adjacent to known ones, and experiments demonstrated that some of the predicted sequences are indeed bound by Fis *in vitro*.

MATERIALS AND METHODS

Nucleotide sequences and binding sites

Fis sites identified by footprinting, gel shift or mutational data were gathered from the GenBank accession numbers, coordinates and orientations shown in Figure 1. The exact alignment of the sites was confirmed by maximizing the information content (54).

A few of the sites previously footprinted were not included. Site II at coordinate 238 of the *aldB* promoter has an information content of –5.3 bits, which strongly implies that it is not a binding site. It was noted in Xu and Johnson (53) that this site is the weakest of all sites and that it gives effects on the DNase I footprint only at

*To whom correspondence should be addressed. Tel: +1 301 846 5581; Fax: +1 301 846 5598; Email: toms@ncifcrf.gov

+Present address: Hood College, 401 Rosemont Avenue, Frederick, MD 21701-8575, USA

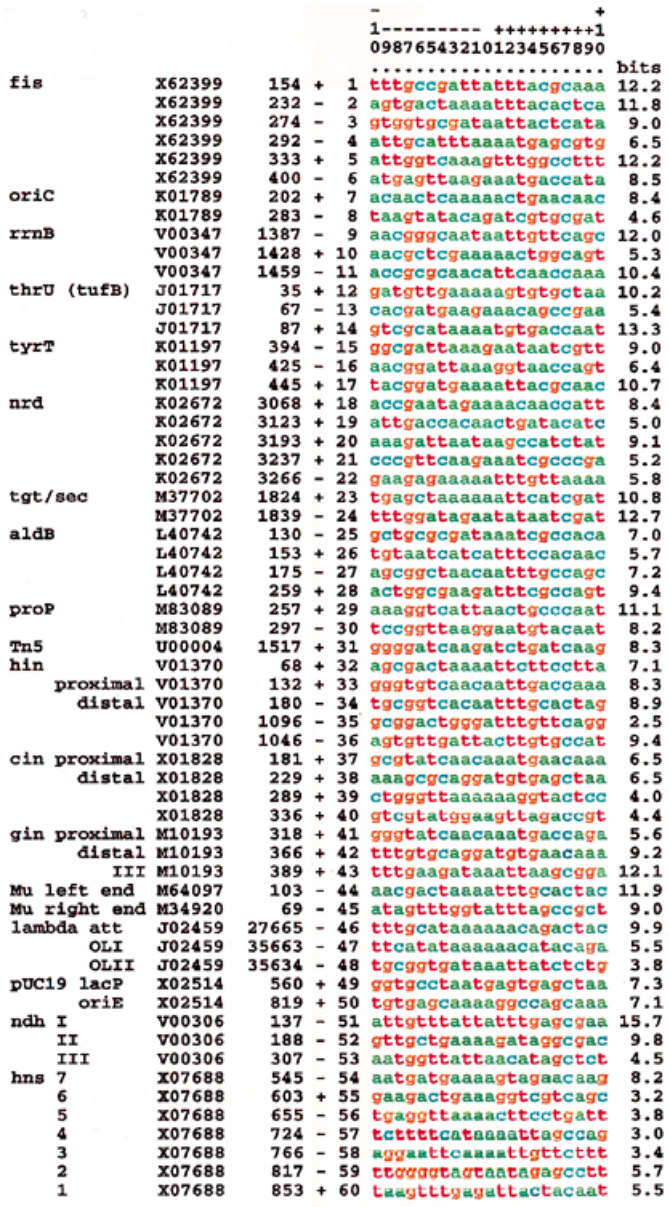

Figure 1. Aligned listing of Fis sites. Sixty Fis binding sites oriented so that there is an A or G in the center were listed by the Alist program. The numbers in the bar on the top are read vertically and give the position in the binding site, running from –10 to +10. The name, GenBank accession number, coordinate, orientation relative to the GenBank entry, number, sequence and individual information content R_i (bits) are given on each line. Each base has been assigned a standard color (a: green; c: blue; g: orange; t: red) throughout this paper. Sequence numbers 47–60 are the new set of 14 sites referred to in ref. 33. Footprint data and sequences were obtained from the following: *fis* (18,19); *oriC* (16,42,71); *rrnB* (14,72,73); *thrU* (*tufB*) (74,75); *tyrT* (21,48); *nrd* (44); *tgt/sec* (45); *aldB* (53); *proP* (76); Tn*5* (41); *hin* (12,77); *cin* (49); *gin* (9,78,79); Mu left end (80,81); Mu right end (80); *lambda att* (36); *lambda* O_LI, O_LII, pUC19 *lacP*, *oriE* (25); *ndh* (55); *hns* (56).

the highest concentration of Fis. Since, according to the model built from the 60 sites listed in Figure 1, the probability that site II is a naturally occurring Fis binding site is $<4.6 \times 10^{-6}$, and there were no better nearby sites, we did not include it in our model. The last *cin* site at 336 was listed in Finkel and Johnson (cin site #4) (4) as being at coordinate 348 on X01828. This had a –0.9 bit site which was on the edge of the DNase I footprint shown in figure 5 of ref. 49. Since there is a 4.4 bit site at coordinate 336 which fits the DNase I footprint exactly, we used it instead.

In the regions around the first 46 sites (Fig. 1) the information theory weight matrix model consistently revealed Fis binding sites for which experimental footprinting data already exist, but which had been missed by searches using a consensus sequence. At this point of our collection, we began using the first 46 sites to help locate the remaining sites (33). We examined the R6K γ origin with our current model and found multiple overlapping Fis sites, only some of which correspond to the available footprint data (26). We believe that further experimental analysis of these sites is warranted before they are included in our model. We used coordinate +48 instead of +51 for *ndh* site III because it was the closest match (55). The model revealed seven sites that correspond to the DNase I hypersensitive locations on *hns* footprints, so we used these (56).

Sequence analysis programs

Delila system programs were used for handling sequences and information calculations (29,32–34,43,57,58). Figures were generated automatically from raw GenBank data using Delila and UNIX script programs. Further information is available on the World Wide Web at http://www-lmmb.ncifcrf.gov/~toms/ .

Design of Fis binding experiments

In designing the sequences of Figure 6 we chose the *hin* site of *S.typhimurium* (12) because it is well characterized and the binding site prediction is clear. We chose 32 bases of the *hin* sequence because according to the information-theory based search (Fig. 5) this region contains two overlapping Fis sites, one of which is the Fis site proximal to the recombination junction *hixL* (12). We added five bases of natural DNA sequence on each end—half a twist of DNA—to be sure we were not missing important components, although this region does not show up significantly in the sequence logo. Beyond these ends we added *Eco*RI and *Hin*dIII overhangs. We created three other sequences using the anti-consensus of the Fis weight matrix to destroy the proximal site, the newly identified 'medial' site, or both sites. The anti-consensus sequence is the sequence that should bind Fis the worst (33). It is predicted from the number of bases at each position or the $R_{iw}(b,l)$ matrix by noting which bases appear least frequently at each position of the site or which give the lowest weight. In ambiguous cases we chose C or G when possible because these appear rarely in the logo (Fig. 2). (Note: the anti-consensus sequence in the early model we used had C at –5 rather than G.)

These sequences and their complements were synthesized (The Midland Certified Reagent Co., Midland, TX, USA) so that when annealed they provide sticky *Eco*RI and *Hin*dIII ends. Annealed oligos were ligated into plasmid pTS385 (59) which had been digested with *Eco*RI and *Hin*dIII, and transformed into *E.coli* DH5α (60) as previously described (61). Transformants were selected on LB media containing 50 μg/ml kanamycin and 50 μg/ml of ampicillin. When necessary, we transformed *E.coli* BL21/DE3 (62) and selected them on the same media containing 1 mM IPTG. We knew from previous experiments that the parental plasmid pTS385 is conditionally lethal to this strain because a strong T7 promoter is positioned between the *Eco*RI and *Hin*dIII sites (59). Induction of the chromosomally imbedded T7 RNA polymerase gene with IPTG thus provided a strong selection for recombinant

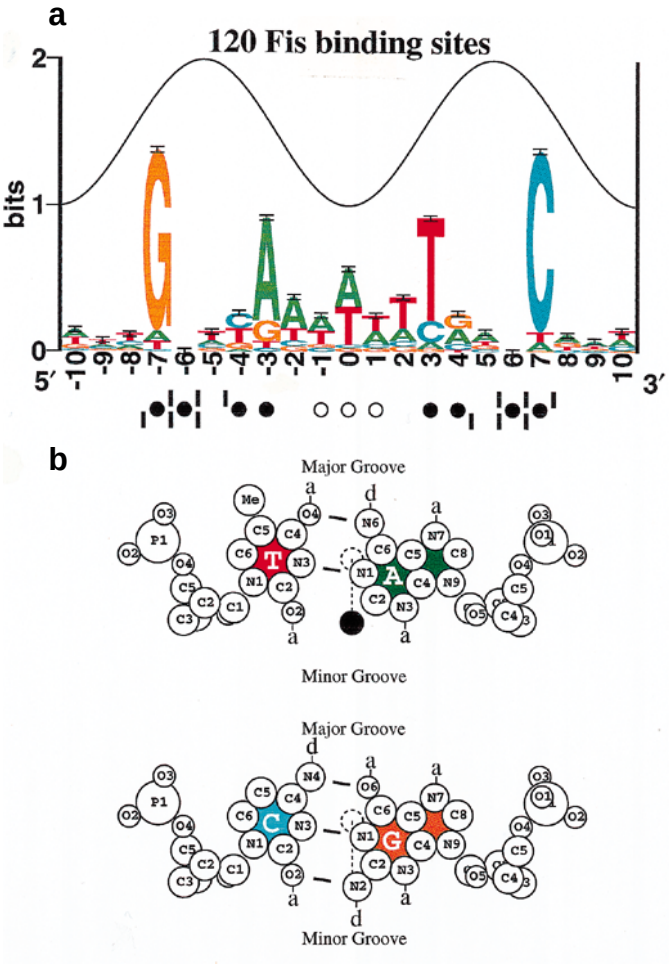

Figure 2. Sequence logo of Fis binding sites and DNA base pair structure. (**a**) Sequence logo for 60 experimentally defined Fis binding sequences and their complements, for a total of 120 bases per position. This models Fis as binding symmetrically. The color scheme is defined in Figure 1 (29). Sequence conservation is measured in bits and shown by the height of a stack of letters. The relative frequencies of bases is given by their relative heights. The total sequence conservation, found by adding the stack heights together, is $R_{sequence}$ = 7.86 ± 0.27 bits per site in the range from –10 to +10. The crest of the sine wave represents the major groove of B-form DNA facing Fis (30,31,35). Methylated guanines that interfere with Fis binding are indicated by filled circles (●) and methylated adenines that interfere with Fis binding are indicated by open circles (○) (12). DNase I hypersensitive phosphates due to the presence of Fis observed on footprints (9,12,14,18,21,41,42,45,48,49,53,72,76,77,80) are marked by vertical bars (|) for the top and bottom strands of the DNA, where the top strand (5′→3′) is the one shown in the logo. The hypersensitive locations correspond to those noted by previous workers (3,4). (**b**) Planar view of DNA base pairs showing possible contacts, d: hydrogen bond donor, a: hydrogen bond acceptor. The black circle (●) represents a probe approaching the DNA from the minor groove that would occlude C-G base pairs, but which would allow either A-T or T-A base pairs. The dashed line represents the dyad axis of the DNA. B-DNA coordinates were from ref. 82 and the radii of atoms from ref. 83. The dashed circle in the center of each base pair is 1 Å in diameter and also represents the helical rotational axis. Bars between the bases represent hydrogen bonds.

plasmids containing the intended insert DNA, eliminating all but a few wild-type pTS385 plasmids from the lot of transformants in BL21/DE3. The resultant plasmids were screened by restriction analysis and PCR amplification using primers flanking the inserted DNA pTS37f1 5′-ACATTTCCCCGAAAAGTGC-3′ and pTS37r1 5′-CGGAACACGTAGAAAGCCA-3′. When recombinants were identified, plasmid DNA was transformed into and maintained within *E.coli* DH5α. The sequence between the *Eco*RI and *Hin*dIII sites was then confirmed by dideoxy sequencing with an ABI model 373A automated sequencer (63).

For gel mobility shifts (64,65), we used Fis protein cloned and purified from *E.coli* obtained as a gift from R.Johnson (66). Plasmid DNA from the four *hin* clones was purified by the method of Birnboim and Doly (67) or Hengen (68), linearized with *Eco*RI, end-filled with biotin-16-dUTP using the Klenow fragment of *E.coli* DNA polymerase I, and digested with *Bgl*II, which cleaves 368 bp from the *Eco*RI site. The smaller fragment was purified away from the vector by electrophoresis and then digested with *Hin*dIII.

Binding assays were accomplished by incubating DNA at ~1 nM with various concentrations of Fis protein as 2-fold serial dilutions beginning with 1000 nM, and at room temperature for 15 min in 25 mM Tris–HCl pH 7.6, 80 mM NaCl, 1 mM EDTA, 2 mM DTT, 100 µg/ml acetylated bovine serum albumin and 125 µg/ml calf thymus DNA (12). Gel shift analysis was done by separation of the different species on an 8.0% polyacrylamide gel in 1× TBE. The DNA was electro-transferred onto Tropilon-Plus™ positively charged nylon membrane and detected using a Southern-Light™ chemiluminescent kit (69) using the CSPD® substrate (CAS = 142849-53-4, Tropix, Inc., Bedford, MA, USA).

For experiments designed to confirm Fis binding, chosen Fis sites were engineered as single-stranded DNAs such that the DNA loops back onto itself by forming a hairpin turn structure centered at 5′-GCGAAGC-3′ (70). The natural Fis sequence from –10 to +10 was used and an *Eco*RI restriction site added on the open end. The DNA sequence was further checked for any extraneous Fis sites above 0 bits by using walkers (34).

DNA oligos for the *cin* external site (5′-GGAATTCTATGTAC-GCGTATCAACAAATGCGAAGCATTTGTTGATACGCGTAC-ATAGAATTCC-3′), *aldB*II (5′-GGAATTCGCTGTCGTAAAGC-TGTTACCGGCGAAGCCGGTAACAGCTTTACGACAGCGA-ATTCC-3′), *tgt*/*sec* at –73 (5′-GGAATTCTGAGCTAAAAAATT-CATCGATGCGAAGCATCGATGAATTTTTTAGCTCAGAATT-CC-3′), *nrd*F1 (5′-GGAATTCTTCGCTTATATATTGACCACAG-CGAAGCTGTGGTCAATATATAAGCGAAGAATTCC-3′), and *nrd*F2 (5′-GGAATTCATTGACCACAACTGATACATCGCGA-AGCGATGTATCAGTTGTGGTCAATGAATTCC-3′) were synthesized with biotin on the 5′ end and gel purified. Gel shift assays were performed essentially as before, except that the DNA was diluted to 1 nM in Fis binding buffer, heated to 90°C for 10 min, and slowly cooled to room temperature. The self-annealing property obviated the need for further manipulation of the DNA.

RESULTS AND DISCUSSION

Interpretation of the Fis sequence logo

To create a model for Fis binding, 60 Fis binding sites for which DNA footprinting or mutational data are available were collected (Fig. 1). Each footprint was compared to the prediction from information scans and walkers to be sure that the location of the binding sites matched the final model. By this method, the locations of several sites given in previous compilations were corrected (see Materials and Methods).

Since *E.coli* and *S.typhimurium* Fis proteins have identical amino acid sequences (3,24) and no modifications that we are aware of, we used binding sites from both species. This would be

inappropriate if *in vivo* DNA binding conditions are different between the two species. Because Fis binds as a dimer (6,10), the sequences and their complements were aligned to produce a sequence logo (29). Analysis of such logos can reveal interaction details that are otherwise obscured by a consensus (31). From the sequences, the sequence logo (Fig. 2a) the structure of DNA base pairs (Fig. 2b), and molecular modeling, the following observations were made:

(i) The correlation between sequence conservation peaks at ±7 and ±3 and a 10.6 base spacing (shown by the sine wave in the figure) suggests that Fis makes contacts in two consecutive major grooves (31). Further, the information content at ±7 is above one bit, which also suggests major groove binding (30,35). This is consistent with protection data showing that the methylation of the major groove N7 of G at ±7 interferes with Fis binding (12), with Fis binding that protects against DMS methylation at ±7 (36), and with hydroxyl radical footprints (37).

(ii) As seen on the logo, if an A is substituted for the majority G at –7 (or the complementary G at +7), a possible G-O6 contact would be lost while a G-N7 contact could be retained as A-N7. On the other hand, if T is substituted, a G-N7 contact would be lost but a G-O6 contact could be replaced by T-O4. This is consistent with the observed frequency of bases at position –7 (and it's complement at +7) for which G>A~T>C. That the frequency of As and Ts are nearly the same at this position suggests this A-N7 contact is energetically equivalent to a T-O4 contact. Similar contacts appear at ±15 in OxyR binding sites (31) and at +6/–7 in CRP (30). The conservation can be explained by direct contacts or indirect through-water bonds.

(iii) At position ±6 there is no observed sequence conservation, yet methylation of a G in the major groove at that position interferes with Fis binding (12). This suggests that Fis passes close to the base in that region but does not make a specific contact.

(iv) At –4 and +3 the logo shows conservation of Cs or Ts, while at positions –3 and +4 the logo shows the complementary As or Gs. Because N7 is the only contact common to both A and G in the major groove (Fig. 2b), this observation suggests that all four positions have N7 contacts (38). These contacts are consistent with DMS interference experiments (12). The relative heights of the letters reveal a 4.8-fold preference for A over G at –3 (T over C at +3), suggesting other direct contacts in the major groove or DNA bending effects.

(v) At positions 0, ±1 and ±2 there is an A-T region where Fis most likely faces the minor groove. Since A is as frequent as T but C and G are allowed at low frequency, this preference could be caused by a series of protein probes that sterically interfere with the N2 of G in the minor groove (Fig. 2b) (30). Consistent with this, methylation of A at N3 in the minor groove at positions 0 and ±1 interferes with Fis binding (12,39).

(vi) The –4 to +4 central region of the Fis logo can be interpreted in a different way. We constructed a three-dimensional model of Fis-DNA binding predicted from the logo and probable contact points (see http://www-lmmb.ncifcrf.gov/~toms/fismodels/ for details). Mutations at Arg 85 and Lys 91 of Fis alter its ability to bind DNA (8,9), and molecular dynamics docking of Fis with DNA supports the notion that these residues contact the DNA (40). When we compressed a Fis binding site in the minor groove from –2 to +2 to account for the A-T region, kinked the DNA at ±3.5 and ±7.5 to create a bend at pyrimidine–purine pairs, and aligned Fis so that Arg 85 contacts G ±7 and Lys 91 contacts phosphate ±1.5, an unavoidable gap appeared that prevents direct contact between Fis and bases –4 to +4. Because our detailed model incorporates all the features observed in the sequence logo, and shows the same gap observed by others (5,6,10,11,40), the entire conservation from –4 to +4 might be accounted for by indirect contacts instead of direct contacts. Direct contacts represent physical contact between the protein and the DNA, while indirect 'contacts' are those in which there is no direct contact but instead the structure of the DNA is distorted, indirectly leading to sequence conservation. Since molecular modeling is not entirely reliable, both the direct and the indirect binding modes are plausible and further experimental work would be required to distinguish between them. However, these two binding modes are not exclusive since it is possible that Fis can flex enough to bind to straight DNA using direct contacts to all of the bases. Subsequently, the DNA could bend using the bending properties of the central bases.

(vii) The sequence logo shows that Fis sites easily accommodate the Dam methylase site 5′-GATC-3′ at ±4 through ±7, suggesting that under some circumstances Fis binding may be controlled by methylation. A Fis site in Tn*5* is only bound when overlapping GATCs are unmethylated (41). The only other occurrence of GATC within a Fis site in our list (Fig. 1) is at 0 to +3 of *oriC* Fis 283, so there may be a connection between Fis and this feature of the origin of DNA replication, as suggested previously (42).

(viii) It is possible that the bases at one position of a Fis site are correlated to those in another position. For example, an A at –3 might only appear when there is an A at –2, but not when there is a T at –2. This would make the sequence logo an incomplete model because these are not displayed. The Diana program (43) shows only faint correlations between –20 and –19 (0.14 bits; $P < 1 \times 10^{-7}$ given the background of correlations from –20 to +20 of –0.02 ± 0.03 bits) and between –2 and –1 (0.12 bits; $P < 1 \times 10^{-6}$) and their complements. As these values are within the error of the total sequence conservation (±0.27 bits), there is little or no missing sequence conservation in the sequence logo model.

Searching specific sequences with the Fis individual information weight matrix model

To model the base preferences of Fis, we computed a weight matrix from:

$$R_{iw}(b,l) = 2 + \log_2 f(b,l) - e(n) \quad \text{(bits per base)} \qquad \mathbf{1}$$

where $f(b,l)$ is the frequency of each base b at position l in the aligned binding site sequences and their complements, and $e(n)$ is a sample size correction factor for the $n = 120$ sequences used to create $f(b,l)$ (33). $R_{iw}(b,l)$ values range between $-\infty$ and 2 bits. To evaluate a DNA sequence, the bases of the sequence are aligned with the matrix entries and the $R_{iw}(b,l)$ values corresponding to each base are added together to produce the total R_i value. This measure has several advantages over other methods. First, the scale is in bits, which are easy units to think about and which allow direct comparison to many other systems. Second, by adding the weights together for various positions in a particular binding site, we get the total 'individual information' (R_i) for that site. Third, the average R_i for all of the binding sites used to create the $R_{iw}(b,l)$ matrix is the average information content, $R_{sequence}$ (32). This is the same as the area under the sequence logo. Fourth, unlike a neural network that needs to be cyclically trained and requires both sites and non-sites, the matrix can be created immediately using only proven sites as examples. This avoids the danger of training

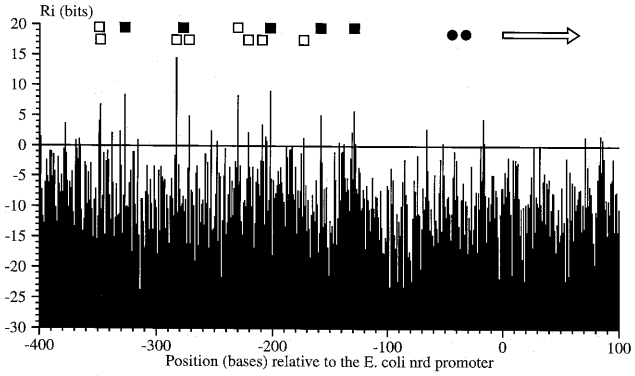

Figure 3. Individual information scan of the *E.coli nrd* promoter and surrounding region produced by programs Scan and DNAplot. The position of the zero base of the Fis weight matrix on the sequence is given on the abscissa, while the individual information for the sequence surrounding each position from –10 to +10 is given on the ordinate. This is computed at each position by adding together the weights that correspond to the sequence around the zero base. The DNA sequence is from GenBank accession K02672 (84). Transcription begins at position 0 (GenBank coordinate 3395) and proceeds to the right (arrow). Fifteen potential Fis sites ($R_i \geq 2$ bits) were located in the region from –400 to +100 relative to the start of transcription. Five Fis sites, indicated by filled squares (■), were identified by Augustin *et al.* (44) to be in the ranges: –328 to –310 (probably site –327), –285 to –268 (probably site –272 since it is the closest to the center of this range), –204 to –187 (probably site –202), –160 to –142 (probably site –158), and –139 to –122 (probably site –129) relative to the start of transcription. These five are listed in Figure 1. Additional sites that were located by the Scan program and are visible on the footprinting data of Augustin *et al.* (their fig. 3, lanes 4 and 5) but not previously described, are indicated by open squares (□). They are at positions: –349, –348, –283, –230, –221, –209 and –173. The two DnaA sites found by Augustin *et al.* are at –44 and –32 and indicated by filled circles (●).

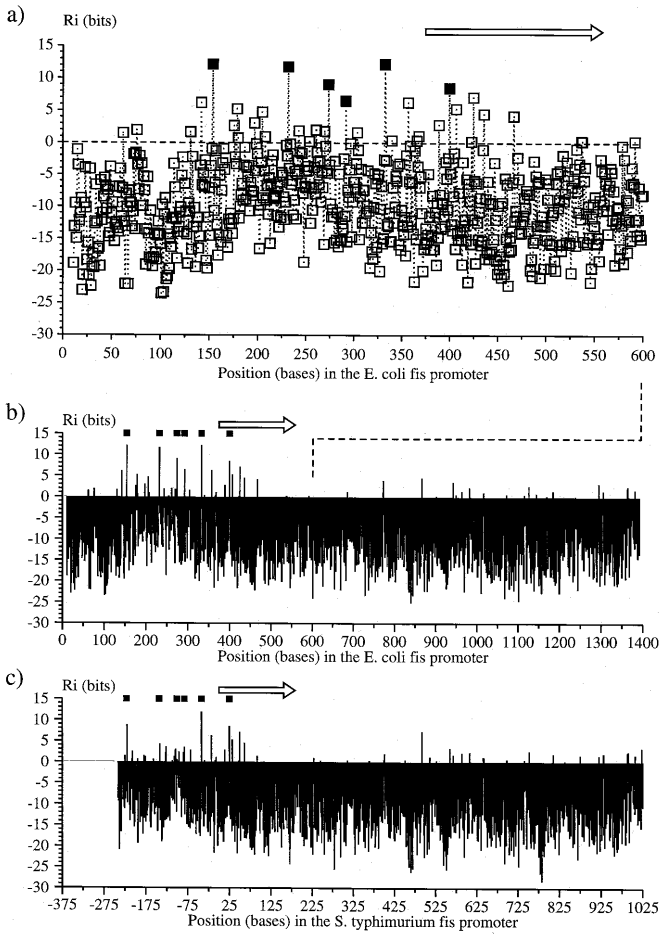

Figure 4. Individual information scan of *fis* promoters. (**a**) Detailed scans of the *E.coli fis* promoter produced by programs Scan and Xyplo. The six previously identified Fis sites are marked with filled squares (■). Predicted sites are represented as open squares (□) above the zero line. Transcription begins at base 375 and proceeds to the right (arrow). The sequence is from GenBank accession X62399 (19) [see also accession M95784 (18)]. (**b**) A larger region of the same *E.coli* sequence graphed by DNAplot shows clustering of potential Fis sites around the promoter but not further downstream. The six previously identified Fis sites are marked with filled squares (■). The dashed line indicates the corresponding parts of the figure. (**c**) The corresponding DNAplot for the *S.typhimurium fis* promoter. The sequence from –49 to +94 around the promoter is identical to the *E.coli* sequence. This can be seen by the corresponding peaks. The sequence is from GenBank accession U03101 (24). This plot differs from Figure 3 in that the individual information scores are drawn as lines from zero bits up or down rather than from the bottom up. This is set by using a switch within the DNAplot parameter file.

against unknown functional sites, and therefore was critical for obtaining the results presented here. Fifth, functional binding sites have positive R_i values, within the error of the method, allowing one to make predictions. Finally, unlike consensus sequences which destroy the available sequence data by arbitrarily rounding the frequencies up or down, the individual information method uses the base frequences directly and so it preserves subtleties in the data.

Validation of the individual information scanning method

We used individual information (33) to study Fis binding sites throughout the *E.coli* genome and at several specific loci. Although the theoretical cutoff for distinguishing sites from non-sites is 0 bits, we often used a conservative 2 bit cutoff to define Fis sites because our previous experience showed that sites between 0 and 2 bits can bind Fis (data not shown). When comparing the output from the Scan, DNAplot, MakeWalker and Lister programs to previously reported footprinting data, we consistently found sites which were seen as DNase I protected regions.

For example, by using a degenerate consensus pattern, previous workers found five Fis binding sites upstream of the transcriptional start site of the *nrd* operon of *E.coli* (44). When we scanned for potential Fis binding sites, several more sites were identified (Fig. 3). These were confirmed to be bona fide sites since Cu-phenanthroline footprinting of this region had already been done by Augustin *et al.* Their data (Fig. 3, lanes 4 and 5) correspond well with our predictions even though none of these additional sites were used in the $R_{iw}(b,l)$ model. In another case, in addition to the site found at position –58 of the *tgt/sec* promoter (45), an information scan shows a second strong site at –73 (34). Both *tgt/sec* sites were included in our model because they are supported by footprinting, gel shift and *in vivo* transcriptional assays. Although Fis has a poor consensus sequence, theoretically it can bind precisely (46,47), and indeed footprints reveal concise binding on well separated sites. Complex footprints appear to be imprecise binding if one uses a consensus sequence. Often the protected genetic regions can be dissected into their components by using individual information tools, so that the data is interpreted as representing overlapping sites.

Using sequence conservation to infer the number of Fis sites in *E.coli*

The total number of Fis sites in the *E.coli* genome is not known, so the information needed to locate those sites ($R_{frequency}$) cannot be calculated (32). However, the total sequence conservation at the binding sites is 7.86 ± 0.27 bits (Fig. 2), which suggests that there is one site roughly every $2^{7.86\pm0.27}$ = 232 ± 43 bases or an average of 4.7 ± 0.9 sites at each of the ~4289 genes of the entire 4 638 858 bp *E.coli* genome (GenBank accession no. U00096, version of 16-JAN-1997). It also implies that ~20 000 ± 3700 Fis molecules would be needed to fill all Fis sites on a single chromosome. Using the method of individual information we scanned the genome and found 68 552 Fis sites with >2 bits of sequence conservation. These estimates are comparable to the number of Fis molecules per cell, which ranges from close to zero in stationary cells to between 25 000 and 50 000 Fis dimers per cell during the transition to exponential growth or an increase in nutrients (18). Thus, almost every Fis site could be filled by one Fis dimer under those growth conditions.

A cluster of Fis sites at the *fis* promoter

Fis is an autoregulatory protein with six strong binding sites and a number of lower-affinity sites near its promoter (18,19). A scan of the *E.coli fis* promoter shows up to 12 additional sites (≥2 bits) in the immediate region of the promoter, but few downstream (Fig. 4a and b). Presumably the additional sites correspond to the weaker sites noted by Ball *et al.* (18).

In a recent study of the corresponding region of DNA from the *S.typhimurium fis* promoter (24), the authors noted that Fis sites are highly degenerate and so they could not predict which sites of the *E.coli fis* promoter region are also present in *S.typhimurium*. They used DNase I footprinting to determine the locations of Fis sites and found that the sites upstream of –49 were weak in *S.typhimurium* relative to those in *E.coli*. Figure 4c demonstrates that this result is predicted by information analysis. Notably, at high concentrations of Fis the weaker sites can be observed by footprinting (24).

In transcriptional activation of stable RNA promoters by Fis, the Fis sites are immediately upstream of the promoter on one face of the DNA and cover a region of 50 bp (48). Repression at the *fis* promoter is different because the Fis sites are spread over 350 bp and are also 90 bp downstream of the start of transcription (Fig. 4). Fis levels increase dramatically after nutritional upshift (18) and under these conditions many of these sites should be occupied simultaneously. Because Fis bends DNA when it binds, the multiple DNA contortions might exclude RNA polymerase and silence transcriptional initiation. As levels of Fis protein decrease in the cell, the physical blockage would be relieved and transcription could proceed again.

Fis sites at recombinational enhancers

Fis sites have been identified on recombinational enhancers (3,4,12,13,49,50). In the *S.typhimurium hin* region there are two Fis sites that are proximal and distal to the *hix*L recombination site. An information scan of this region shows a third potential Fis site located 11 bp [~1 helical turn of DNA, 10.6 bp (51,52)] to the right of and overlapping the proximal site (Fig. 5, top left). We call this site the ‘medial’ site; it is 37 bp (~3.5 helical turns) to the left

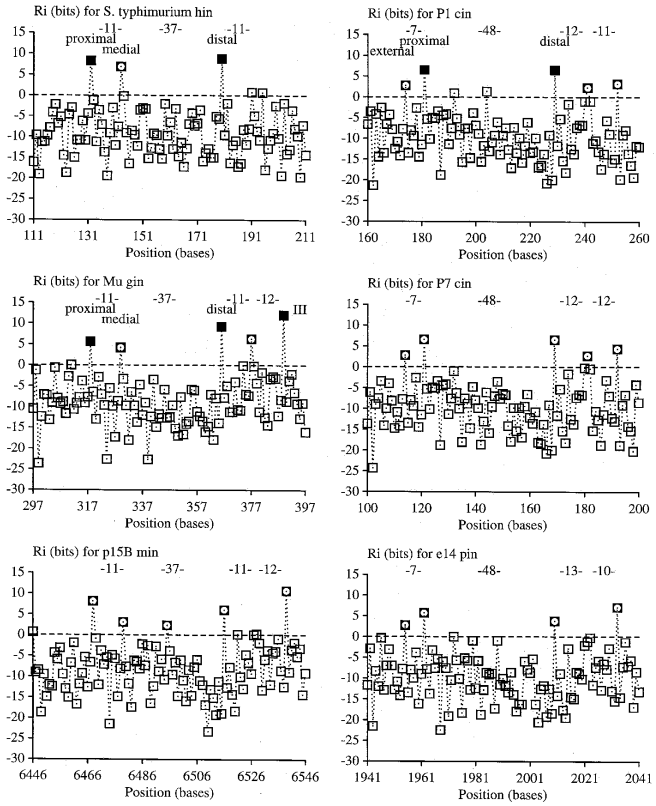

Figure 5. Individual information scans of inversion regions. Symbols are the same as in Figure 4. Previously identified Fis sites are marked with filled squares (■) and named as in refs 3,4. The proposed Fis sites with $R_i \geq 2$ bits are marked with a circle inside a square. Spacing between sites is indicated by numbers surrounded by dashes. Note that the spacing between proximal and distal sites is always 48 bases (11 + 37 on the left three graphs) (2). GenBank accession numbers: *hin*, V01370; *gin*, M10193; *min*, X62121; P1 *cin*, X01828; P7 *cin*, X07724; *pin*, X01805.

of the distal site. The same structure is found in bacteriophage Mu *gin* and p15B *min* enhancers (Fig. 5, left three graphs).

In the bacteriophage P1 *cin*, bacteriophage P7 *cin*, and *E.coli e14 pin* enhancers, a potential overlapping site occurs 7 bp (~1/2 helical turn) to the left of the previously identified proximal site (Fig. 5, right three graphs). Since this potential site is outside the region between the proximal and distal sites, we named it the ‘external’ site.

At recombinational enhancer proximal Fis sites, when a potential new Fis site is found on the right, it is 11 bases away while when a potential new Fis site is found on the left, it is 7 bases away. It is not known whether this correlation is coincidental. We also observed that potential Fis sites corresponding in location to site III in *gin* (9) appear in all other enhancers scanned except *hin* and that in three cases a weaker potential site falls exactly between the distal site and the one corresponding to site III with spacings of 11–12 bp.

Predicted Fis sites are bound *in vitro*

Scanning the Fis $R_{iw}(b,l)$ model across DNA inversion regions reveals pairs of Fis sites spaced either 7 or 11 bases apart (Fig. 5).

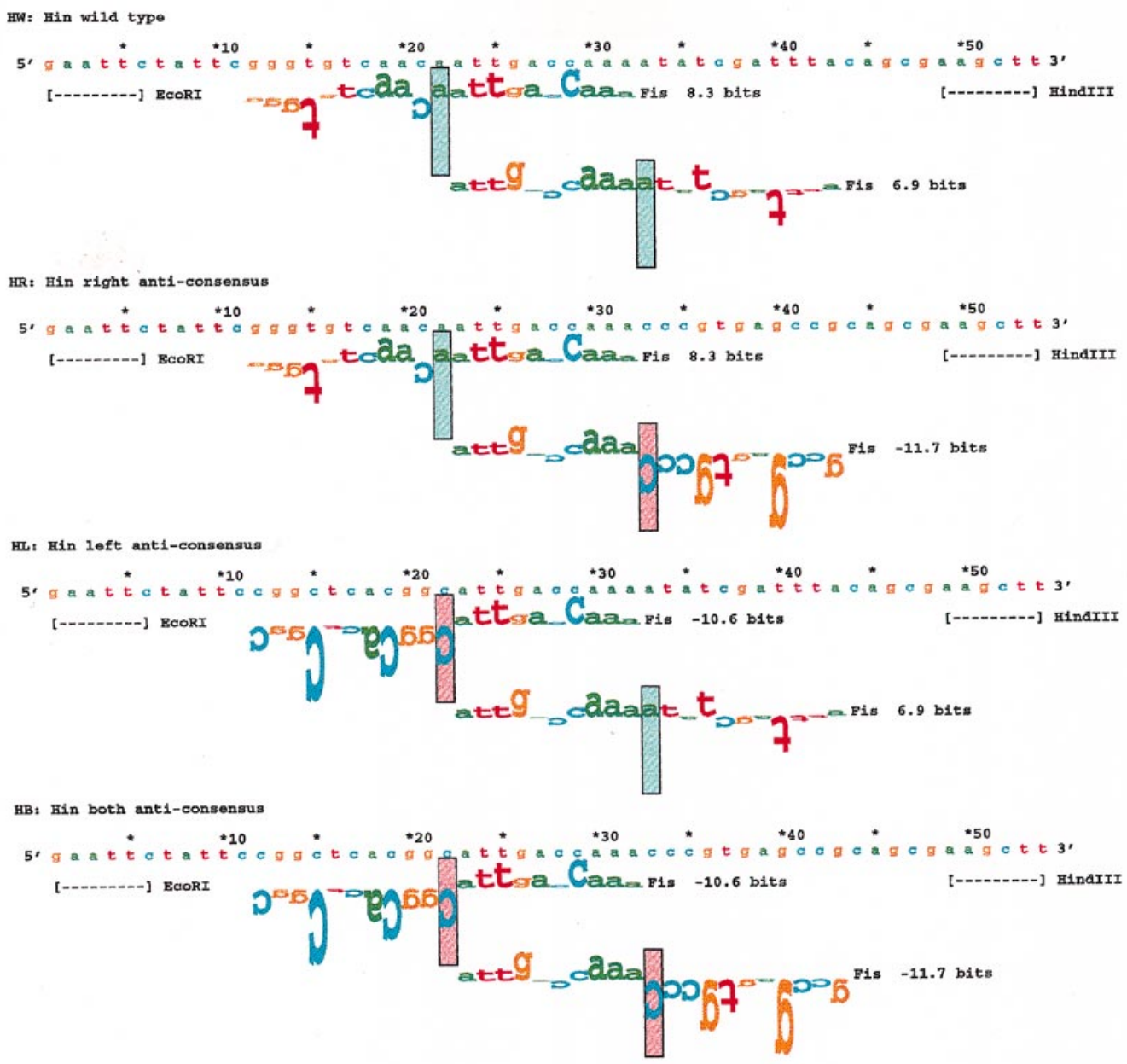

Figure 6. Oligonucleotide design of mutant *S.typhimurium hin* Fis binding sites. In the first construction, at the top of the figure, the wild-type sequence containing the proximal Fis site from the *S.typhimurium hin* region (HW, *hin* wild-type) is given (coordinates 117–158 of GenBank accession V01370), flanked by *Eco*RI and *Hin*dIII restriction sites. The known proximal site (R_i = 8.3 bits) is indicated next to the predicted medial site (R_i = 6.9 bits). Both sites are shown by walkers (34). The vertical bars extend from –5 to +2 bits. Normal orientation of letters indicates positive contribution to the sequence conservation of the site, while inverted letters indicate negative contribution. The height of each letter is given by the information weight matrix according to equation **1**. The total sequence conservation is the sum of the letter heights. Sites with conservation more than zero bits (green bars) are expected to be bound by Fis, those less than zero bits (pink bars) are expected not to be bound by Fis. In the second construction, the right anti-consensus Fis site sequence (HR, *hin* right) was used to destroy the medial site, leaving the proximal site intact. In the third construction, the left anti-consensus (HL, *hin* left) was used to destroy the proximal site while leaving the medial site intact. In the fourth construction, both (HB, *hin* both) sites were destroyed. See Materials and Methods section for design details.

Three of the sites are the footprinted proximal sites. In addition, the medial site for *gin* is supported by footprint data, but it was not identified as a Fis site (9). To test whether the proposed medial site exists at *hin* we performed gel shift experiments on DNAs in which we presumably had knocked out neither, one, or both of the sites. The DNA design is shown in Figure 6 using sequence walkers, a graphical representation of the individual information content at specific binding sites (34). Characters representing the sequence are either oriented normally and placed above a line indicating favorable contact, or upside-down and placed below the line indicating unfavorable contact. Functional sites therefore have most letters pointing upwards, while those we have destroyed have many letters pointing downwards. The walkers also show that we did not inadvertently create any other Fis sites.

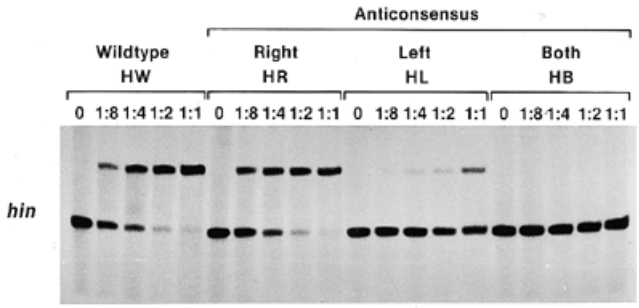

Figure 7. Mobility shift experiments for *hin*. Gel shifts of DNA containing the *hin* proximal and medial Fis binding sites. Each lane contains increasing concentrations of Fis protein added, beginning with no Fis protein, Fis diluted 1 to 8, etc. The 1:1 ratio is 1000 nM Fis. Letter designations refer to the sequences given in Figure 6.

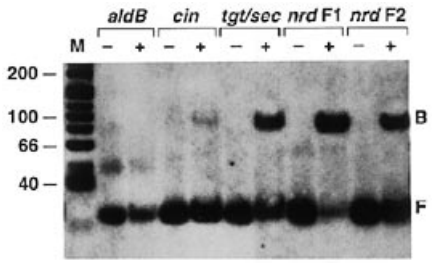

Figure 8. Mobility shift experiments for predicted Fis sites. Gel shifts of hairpin structures containing the *aldB*II at 238 (–5.3 bits), *cin* external site at 174 (2.8 bits), *tgt*/*sec* at –73 (10.8 bits), *nrd*F1 site at –283 (14.6 bits) and *nrd*F2 site at –272 (5.0 bits). Each lane contains 20 µl of 1 nM DNA with either no Fis (–) or 1000 nM Fis added (+). The marker lane (M) contains 10 ng of biotinylated ϕX174 *Hin*fI digested DNA standards (Life Technologies, Inc.), with sizes indicated in bp. The original X-ray film and photograph were intentionally overexposed to reveal the weaker *cin* shift product. B, DNA bound to Fis; F, free DNA.

The results of shifting the *hin* sequences of Figure 6 are shown in Figure 7.

Under our experimental conditions *hin* does have a second site as predicted, since the knockout of the stronger proximal site still allowed the DNA to shift (Fig. 7, HL). However, more Fis protein was required to shift an equivalent amount of DNA than for the wild-type proximal site, indicating that Fis binds weakly to the medial site. This is consistent with the weaker sequence conservation of the medial site (R_i = 6.9 bits) compared to the proximal site (R_i = 8.3 bits).

To further investigate the predictive ability of our individual information model, we synthesized five oligonucleotides representing various interesting sites:

(i) We were curious as to whether the predicted *cin* external site (7 bases from the proximal site, R_i = 2.8 bits), could bind Fis, even though it is so close to the proximal site (Fig. 5).

(ii) Four lines of evidence indicate that a site predicted to be at –73 of the *tgt*/*sec* promoter (R_i = 10.8 bits, Fig. 1, # 23) should bind Fis (see fig. 2 of ref. 34). We decided to test this prediction directly.

(iii) Footprinting data covers two overlapping sites spaced 11 bases apart within the *nrd* promoter at –283 (*nrd*F1, R_i = 14.6 bits) and –272 (*nrd*F2, R_i = 5.0 bits, Fig. 1, #19) relative to the start of transcription (Fig. 3), but only one site had been identified (44). We decided to test both.

(iv) The site previously identified as Fis site II at 238 in the *aldB* promoter (53) has a negative R_i value (–5.3 bits) and therefore should not bind Fis.

Figure 8 shows that all four sites having positive R_i values are able to bind Fis as predicted (33). Although the gel is not quantitative, the band intensities correlate well with information content:

(i) The *cin* external site was bound weakly, suggesting that it might be involved in site-specific inversion.

(ii) The site within the *tgt*/*sec* promoter region upstream from the start site of the *queA* gene (34) had been previously footprinted (45), however, that footprint extended up to and included a DNase I hypersensitive region at –79. In addition, a secondary shift product was observed when that DNA was used in a gel shift experiment. There are actually two adjacent Fis sites in that region since we were able to shift the one at –73.

(iii) The two sites within the *nrd* promoter were also previously footprinted (44). We show here that when separated they are individually able to bind Fis. These two sites are likely to be responsible for the single protected region seen on the published footprint from –268 to –285.

(iv) Although it had been observed as a protected region on a footprint (53), the site II at coordinate 238 of the *aldB* promoter has an information content of –5.3 bits and, as expected, it did not shift. This justifies excluding it from the list of known sites (see Materials and Methods). We propose that the DNase I protection observed could be an artifact due to secondary structures formed by a DNA–Fis complex when Fis binds to other surrounding sites. A requirement for binding multiple Fis molecules could explain why a high concentration of Fis is required for protection (53). Alternatively, it could represent a special binding mode of Fis.

In summary, we have shown that information theory can be used to predict Fis binding sites, and we have confirmed some of those sites experimentally. Furthermore, information theory can also predict when a sequence is unlikely to be a binding site. The information theory models can be applied to any nucleic acid binding interaction, so they provide a general tool for researchers to identify and characterize binding sites.

ACKNOWLEDGEMENTS

We thank Reid Johnson for generously supplying Fis protein, Denise Rubens for technical assistance and for sequencing, the Frederick Biomedical Supercomputing Center for access to computer resources, Peter Rogan, R. M. Stephens, Keith Robison and Dhruba Chattoraj for comments on the manuscript.

REFERENCES

1 Craig,N.L. and Kleckner,N. (1987) In Neidhardt,F.C., Ingraham,J.L., Low,K.B., Magasanik,B., Schaechter,M. and Umbarger,H.E. (eds) *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology*. Vol 2. American Society for Microbiology, Washington, DC. pp. 1054–1070.
2 Johnson,R.C. and Simon,M.I. (1987) *Trends Genet.*, **3**, 262–267.
3 Finkel,S.E. and Johnson,R.C. (1992) *Mol. Microbiol.*, **6**, 3257–3265.
4 Finkel,S.E. and Johnson,R.C. (1992) *Mol. Microbiol.*, **6**, 1023.
5 Kostrewa,D., Granzin,J., Koch,C., Choe,H.-W., Raghunathan,S., Wolf,W., Labahn,J., Kahmann,R. and Saenger,W. (1991) *Nature*, **349**, 178–180.
6 Yuan,H.S., Finkel,S.E., Feng,J.-A., Kaczor-Grzeskowiak,M., Johnson,R.C. and Dickerson,R.E. (1991) *Proc. Natl. Acad. Sci. USA*, **88**, 9558–9562.
7 Thompson,J.F. and Landy,A. (1988) *Nucleic Acids Res.*, **16**, 9687–9705.
8 Osuna,R., Finkel,S.E. and Johnson,R. (1991) *EMBO J.*, **10**, 1593–1603.
9 Koch,C., Ninnemann,O., Fuss,H. and Kahmann,R. (1991) *Nucleic Acids Res.*, **19**, 5915–5922.

10 Kostrewa,D., Granzin,J., Stock,D., Choe,H.-W., Labahn,J. and Saenger,W. (1992) *J. Mol. Biol.*, **226**, 209–226.
11 Pan,C.Q., Feng,J.-A., Finkel,S.E., Landgraf,R., Sigman,D. and Johnson,R.C. (1994) *Proc. Natl. Acad. Sci. USA*, **91**, 1721–1725.
12 Bruist,M.F., Glasgow,A.C., Johnson,R.C. and Simon,M.I. (1987) *Genes Dev.*, **1**, 762–772.
13 Hübner,P. and Arber,W. (1989) *EMBO J.*, **8**, 577–585.
14 Ross,W., Thompson,J.F., Newlands,J.T. and Gourse,R.L. (1990) *EMBO J.*, **9**, 3733–3742.
15 Numrych,T., Gumport,R.I. and Gardner,J.F. (1991) *J. Bacteriol.*, **173**, 5954–5963.
16 Gille,H., Egan,J.B., Roth,A. and Messer,W. (1991) *Nucleic Acids Res.*, **19**, 4167–4172.
17 Messer,W., Egan,B., Gille,H., Holz,A., Schaefer,C. and Woelker,B. (1991) *Res. Microbiol.*, **142**, 119–125.
18 Ball,C.A., Osuna,R., Ferguson,K.C. and Johnson,R.C. (1992) *J. Bacteriol.*, **174**, 8043–8056.
19 Ninnemann,O., Koch,C. and Kahmann,R. (1992) *EMBO J.*, **11**, 1075–1083.
20 Condon,C., Philips,J., Fu,Z.-Y., Squires,C. and Squires,C.L. (1992) *EMBO J.*, **11**, 4175–4185.
21 Lazarus,L.R. and Travers,A.A. (1993) *EMBO J.*, **12**, 2483–2494.
22 Dorgai,L., Oberto,J. and Weisberg,R.A. (1993) *J. Bacteriol.*, **175**, 693–700.
23 Nilsson,L. and Emilsson,V. (1994) *J. Biol. Chem.*, **269**, 9460–9465.
24 Osuna,R., Lienau,D., Hughes,K.T. and Johnson,R.C. (1995) *J. Bacteriol.*, **177**, 2021–2032.
25 Pan,C.Q., Johnson,R.C. and Sigman, D.S. (1996) *Biochemistry*, **35**, 4326–4333.
26 Wu,F., Wu,J., Ehley,J. and Filutowicz,M. (1996) *J. Bacteriol.*, **178**, 4965–4974.
27 Schneider,T.D. (1995) Information Theory Primer, http://www-lecb.ncifcrf.gov/~toms/paper/primer/
28 Pierce,J.R. (1980) *An Introduction to Information Theory: Symbols, Signals and Noise*. 2nd edition. Dover Publications, Inc., New York.
29 Schneider,T.D. and Stephens,R.M. (1990) *Nucleic Acids Res.*, **18**, 6097–6100.
30 Papp,P.P., Chattoraj,D.K. and Schneider,T.D. (1993) *J. Mol. Biol.*, **233**, 219–230.
31 Schneider,T.D. (1996) *Methods Enzymol.*, **274**, 445–455. http://www-lecb.ncifcrf.gov/~toms/paper/oxyr
32 Schneider,T.D., Stormo,G.D., Gold,L. and Ehrenfeucht,A. (1986) *J. Mol. Biol.*, **188**, 415–431.
33 Schneider,T.D. (1997) *J. Theor. Biol.*, **189**, in press.
34 Schneider,T.D. (1997) *Nucleic Acids Res.*, **25**, 4408–4415.
35 Papp,P.P. and Iyer,V.N. (1995) *J. Mol. Biol.*, **246**, 595–608.
36 Thompson,J.F., deVargas,L.M., Koch,C., Kahmann,R. and Landy,A. (1987) *Cell*, **50**, 901–908.
37 Bokal,A.J.,IV, Ross,W. and Gourse,R.L. (1995) *J. Mol. Biol.*, **245**, 197–207.
38 Seeman,N.C., Rosenberg,J.M. and Rich,A. (1976) *Proc. Natl. Acad. Sci. USA*, **73**, 804–808.
39 Siebenlist,U. and Gilbert,W. (1980) *Proc. Natl. Acad. Sci. USA*, **77**, 122–126.
40 Sandmann,C., Cordes,F. and Saenger,W. (1996) *Proteins Struct. Funct. Genet.*, **25**, 486–500.
41 Weinreich,M.D. and Reznikoff,W.S. (1992) *J. Bacteriol.*, **174**, 4530–4537.
42 Filutowicz,M., Ross,W., Wild,J. and Gourse,R.L. (1992) *J. Bacteriol.*, **174**, 398–407.
43 Stephens,R.M. and Schneider,T.D. (1992) *J. Mol. Biol.*, **228**, 1124–1136.
44 Augustin,L.B., Jacobson,B.A. and Fuchs,J.A. (1994) *J. Bacteriol.*, **176**, 378–387.
45 Slany,R.K. and Kersten,H. (1992) *Nucleic Acids Res.*, **20**, 4193–4198.
46 Schneider,T.D. (1991) *J. Theor. Biol.*, **148**, 83–123. http://www-lecb.ncifcrf.gov/~toms/paper/ccmm/
47 Schneider,T.D. (1994) *Nanotechnology*, **5**, 1–18. http://www-lecb.ncifcrf.gov/~toms/paper/nano2/
48 Muskhelishvili,G., Travers,A.A., Heumann,H. and Kahmann,R. (1995) *EMBO J.*, **14**, 1446–1452.
49 Haffter,P. and Bickle,T.A. (1987) *J. Mol. Biol.*, **198**, 579–587.
50 Koch,C., Vandekerckhove,J. and Kahmann,R. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 4237–4241.
51 Peck,L.J. and Wang,J.C. (1981) *Nature*, **292**, 375–378.
52 Rhodes,D. and Klug,A. (1981) *Nature*, **292**, 378–380.
53 Xu,J. and Johnson,R.C. (1995) *J. Bacteriol.*, **177**, 3166–3175.
54 Schneider,T.D. and Mastronarde,D. (1996) *Discrete Appl. Math.*, **71**, 259–268. http://www-lecb.ncifcrf.gov/~toms/paper/malign
55 Green,J., Anjum,M.F. and Guest,J.R. (1996) *Mol. Microbiol.*, **19**, 1043–1055.
56 Falconi,M., Brandi,A., Teana,A.L., Gualerzi,C.O. and Pon,C.L. (1996) *Mol. Microbiol.*, **19**, 965–975.
57 Schneider,T.D., Stormo,G.D., Haemer,J.S. and Gold,L. (1982) *Nucleic Acids Res.*, **10**, 3013–3024.
58 Schneider,T.D., Stormo,G.D., Yarus,M.A. and Gold,L. (1984) *Nucleic Acids Res.*, **12**, 129–140.
59 Schneider,T.D. and Stormo,G.D. (1989) *Nucleic Acids Res.*, **17**, 659–674.
60 Sambrook,J., Fritsch,E.F. and Maniatis,T. (1989) *Molecular Cloning, A Laboratory Manual*. 2nd edition. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
61 Hengen,P.N. and Iyer,V.N. (1992) *BioTechniques*, **13**, 57–62.
62 Studier,F.W. and Moffatt,B.A. (1986) *J. Mol. Biol.*, **189**, 113–130.
63 Hunkapiller,T., Kaiser,R.J., Koop,B.F. and Hood,L. (1991) *Science*, **254**, 59–67.
64 Fried,M. and Crothers,D.M. (1981) *Nucleic Acids Res.*, **9**, 6505–6525.
65 Garner,M.M. and Revzin,A. (1981) *Nucleic Acids Res.*, **9**, 3047–3060.
66 Johnson,R.C., Bruist,M.F. and Simon,M.I. (1986) *Cell*, **46**, 531–539.
67 Birnboim,H.C. and Doly,J. (1979) *Nucleic Acids Res.*, **7**, 1513–1523.
68 Hengen,P.N. (1995) In Griffin,A.M. and Griffin,H.G. (eds) *Molecular Biology: Current Innovations and Future Trends*. volume Part 1. Horizon Scientific Press, Wymondham, UK. pp. 39–50.
69 Bronstein,I., Voyta,J.C., Lazzari,K.G., Murphy,O., Edwards,B. and Kricka,L.J. (1990) *BioTechniques*, **8**, 310–314.
70 Hirao,I., Kawai,G., Yoshizawa,S., Nishimura,Y., Ishido,Y., Watanabe,K. and Miura,K. (1994) *Nucleic Acids Res.*, **22**, 576–582.
71 Roth,A., Urmoneit,B. and Messer,W. (1994) *Biochimie*, **76**, 917–923.
72 Newlands,J.T., Josaitis,C.A., Ross,W. and Gourse,R.L. (1992) *Nucleic Acids Res.*, **20**, 719–726.
73 Gosink,K.K., Ross,W., Leirmo,S., Osuna,R., Finkel,S.E., Johnson,R.C. and Gourse,R.L. (1993) *J. Bacteriol.*, **175**, 1580–1589.
74 Bosch,L., Nilsson,L., Vijgenboom,E. and Verbeek,H. (1990) *Biochim. Biophys. Acta*, **1050**, 293–301.
75 Nilsson,L., Vanet,A., Vijgenboom,E. and Bosch,L. (1990) *EMBO J.*, **9**, 727–734.
76 Xu,J. and Johnson,R.C. (1995) *J. Bacteriol.*, **177**, 5222–5231.
77 Glasgow,A.C., Bruist,M.F. and Simon,M.I. (1989) *J. Biol. Chem.*, **264**, 10072–10082.
78 Kahmann,R., Rudt,F., Koch,C. and Mertens,G. (1985) *Cell*, **41**, 771–780.
79 Kahmann,R., Mertens,G., Klippel,A., Bräuer,B., Rudt,F. and Koch,C. (1987) In McMacken,R. and Kelly,T.J. (eds) *DNA Replication and Recombination*. Alan R. Liss, Inc, New York. pp. 681–690.
80 Bétermier,M., Lefrère,V., Koch,C., Alazard,R. and Chandler,M. (1989) *Mol. Microbiol.*, **3**, 459–468.
81 Bétermier,M., Galas,D.J. and Chandler,M. (1994) *Biochimie*, **76**, 958–967.
82 Arnott,S. and Hukins,D.W.L. (1972) *Biochem. Biophys. Res. Commun.*, **47**, 1504–1509.
83 Karplus,M. and Porter,R.N. (1970) *Atoms & Molecules*. Benjamin/Cummings Publishing Co., Menlo Park, CA. pp. 204–207.
84 Carlson,J., Fuchs,J.A. and Messing,J. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 4294–4297.