Published online 7 July 2007 *Nucleic Acids Research, 2007, Vol. 35, No. 16* ***5275–5283***
doi:10.1093/nar/gkm471

Correlation between binding rate constants and individual information of *E. coli* Fis binding sites

Ryan K. Shultzaberger[1], Lindsey R. Roberts[2], Ilya G. Lyakhov[3], Igor A. Sidorov[1], Andrew G. Stephen[2], Robert J. Fisher[2] and Thomas D. Schneider[1,*]

[1]National Cancer Institute at Frederick, Center for Cancer Research Nanobiology Program, [2]The Protein Chemistry Laboratory, Advanced Technology Program, SAIC - Frederick, NCI - Frederick Bldg. 469, Rm 237 Frederick, MD 21782 and [3]Basic Research Program, SAIC-Frederick, Inc., NCI-Frederick, Frederick, MD 21702, USA

Received March 8, 2007; Revised May 26, 2007; Accepted May 29, 2007

ABSTRACT

Individual protein binding sites on DNA can be measured in bits of information. This information is related to the free energy of binding by the second law of thermodynamics, but binding kinetics appear to be inaccessible from sequence information since the relative contributions of the on- and off-rates to the binding constant, and hence the free energy, are unknown. However, the on-rate could be independent of the sequence since a protein is likely to bind once it is near a site. To test this, we used surface plasmon resonance and electromobility shift assays to determine the kinetics for binding of the Fis protein to a range of naturally occurring binding sites. We observed that the logarithm of the off-rate is indeed proportional to the individual information of the binding sites, as predicted. However, the on-rate is also related to the information, but to a lesser degree. We suggest that the on-rate is mostly determined by DNA bending, which in turn is determined by the sequence information. Finally, we observed a break in the binding curve around zero bits of information. The break is expected from information theory because it represents the coding demarcation between specific and nonspecific binding.

INTRODUCTION

Transcription factors bind to a variety of sequences with different affinities (1). The amount of sequence variability within a set of binding sites is limited by physical requirements for binding, as well as the ability for the site to be distinguished from non-sites in the genome (2). A range of affinities allows for a subtle regulation of transcription. In the case of activators, higher affinity sites will presumably be bound longer than lower affinity sites, and have a greater probability of stabilizing the initiation complex, which in turn has a greater probability of transcribing a gene. Therefore, the affinity of the protein for a site is a direct indicator of the degree that that site will affect the gene expression.

Being able to predict binding affinities for different DNA targets is useful in characterizing genetic regulatory pathways. To do this, we use an information theory-based weight matrix to quantify protein binding to individual sequences (3).

Information theory was developed by Claude Shannon in the late 1940s to describe the movement of information in communications (4). When applied to biological systems it has proven to be useful (2,5–7). Based on the frequency of each base at each position in a set of aligned binding sites, we can determine the strength of an individual site in bits of information. This strength is called the individual information, R_i (rate of individual information transfer, bits per site) for a site (3). Advantages of this approach are discussed in Materials and Methods.

It has been shown that the protein–DNA dissociation constant, K_D, varies with DNA sequences, and can be approximated by different weight matrix approaches (8–12). The information in a binding site should be related to the binding energy (13). Binding energy, in turn, is proportional to the logarithm of the ratio of the association (k_{on}) and dissociation (k_{off}) rate constants of binding. Since the on-rate depends on diffusion of the protein to the DNA binding site, we expected that the on-rate would be independent of the binding sequence. This suggests that the information of binding sites (R_i) should be linearly related to the logarithm of the off-rate. Others have reported differences in binding rate constants as a function of sequence (14–16), but they did not report any relationship between the rate constants and

*To whom correspondence should be addressed. Tel: +1 301 846 5581 (-5532 for messages); Fax: +1 301 846 5598; Email: toms@ncifcrf.gov
Present address: Ryan K. Shultzaberger, Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA

affinity predictions. No one has shown how information theory predictions of individual binding sites are related to binding and dissociation kinetics.

To address this issue, we used surface plasmon resonance (SPR) technology (17–19) and electrophoretic mobility shift assays (EMSA) to measure the binding kinetics for 13 Fis binding sites ranging in predicted site strength, based on our information theory approach. Fis is a pleiotropic homodimeric DNA binding protein involved in site-specific recombination, chromosomal compaction and transcriptional regulation (6,20,21). Because many genomic sites have been experimentally identified, a reliable Fis model could be constructed and verified (6,22), making it a good protein for this analysis.

MATERIALS AND METHODS

Constructing the Fis model

The Fis binding site model was built using the standard Delila programs (23,24) (Figure 1), and was originally presented in (6). Individual information analysis (3,25) of Fis binding sites was computed using a weight matrix from the equation:

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

where $f(b,l)$ is the frequency of each base b at each position l for all positions in an aligned set of binding sites. e(n) is a sample correction value where n is 120, the number of Fis binding sites and their complements that make up our frequency matrix. To determine the strength of a site (R_i), a DNA sequence is compared to the R_{iw} (b,l) weight matrix and the information contribution of each base is summed across the site. There are several advantages to our approach. First, our models are composed of only experimentally verified binding sites, and do not require a training set of unproven 'non-sites' like many neural-networks or HMMs (26–29). Second, our method has no arbitrary parameters, and the theory predicts that all sites with greater than zero bits of information have a negative ΔG of binding (3). Third, the units of measurement, bits, allow direct comparison between different molecular systems. Fourth, the average R_i for all binding sites that define an R_{iw} (b,l) is R_{sequence}, or the total information content [the area under a sequence logo (2)]. The information content is a measure of the sequence conservation and it is determined by the evolution of the sites in the genome (30).

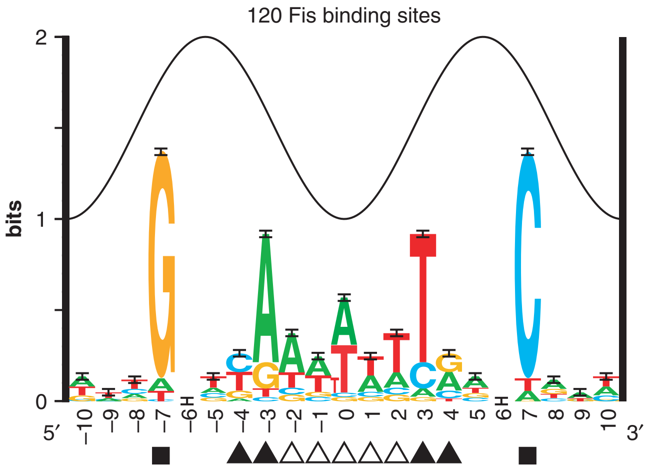

Figure 1. Sequence logo for the Fis protein (6). The heights of letters in each stack are proportional to the frequency of each base at that position. The height of the stack is the information content for that position (23). The total conservation, summed for all positions in the range −7 to +7 is $R_{\text{sequence}} = 7.18 \pm 0.23$ bits per site (2), which is also the average of the individual information of all of the sites (3). The sine wave above the logo represents the 10.6 bp helical twist of B-form DNA (24). The positions presumably bound by the D helices from the major groove at ±7 are marked with squares, the pyrimidine/purine steps that kink the DNA at ±4 and ±3 are marked with filled triangles, and the A/T bases that allow bending into the minor groove are marked with open triangles.

We used a Fis model ranging from −7 to +7 throughout this article. This assumes that positions outside this region do not affect binding and it is consistent with known footprinting data (6). The small amount of information observed in positions −10 to −8 and +8 to +10 (Figure 1) may correspond to overlapping adjacent Fis sites (22).

Individual information analysis was done using the program **scan** and sequence walkers were generated using **lister** (3,31) (Figure 2).

Oligo construction

Thirteen oligos of varying information content were synthesized to measure binding kinetics. Ten of these contain naturally occurring Fis binding sites, where binding has been experimentally verified (32–38). These sites are presented in Table 1 and Figure 2. We chose these oligos to cover a spectrum of strengths from 4.9 to 12.7 bits, as assessed by our information theory approach. The three remaining oligos do not contain characterized binding sites, but have been engineered by us to test binding at additional site strengths.

The first engineered oligo is the Fis consensus of 5′-ATTGGTTAAATTTTAACCAAT-3′ over the range −10 to +10, containing three extra natural bases on each end (Figure 1), which is presumably the highest strength site (14.9 bits), and it does not occur in the *Escherichia coli* genome (named con in Table 1, Figure 2). The second oligo is a slight modification of this consensus, where we mutated the T at position +1 to a G to decrease the strength of the site to 12.8 bits (named mut-con in Table 1, Figure 2). The third oligo is the Fis anti-consensus of 5′-CGGCTGACCCCGGGTCAGCCG-3′, which is made up of the least favorable base at each position (named anti-con in Table 1, Figure 2). The kinetics of binding to this sequence are presumably those of nonspecific interactions of Fis with DNA.

All sequences were inserted into the same hairpin construct: 5′-GCTATCGCG-[Sequence]-ACGATCGCGC-GAA-GCGCGATCGT-[Complement of Sequence]-CG CGA-3′, where there is a 5′ 4 bp overhang of GCTA to allow for future modification, and a 3 bp loop of GAA in the center. This construct has been shown to form tight hairpins (6,39). All oligos were synthesized carrying a

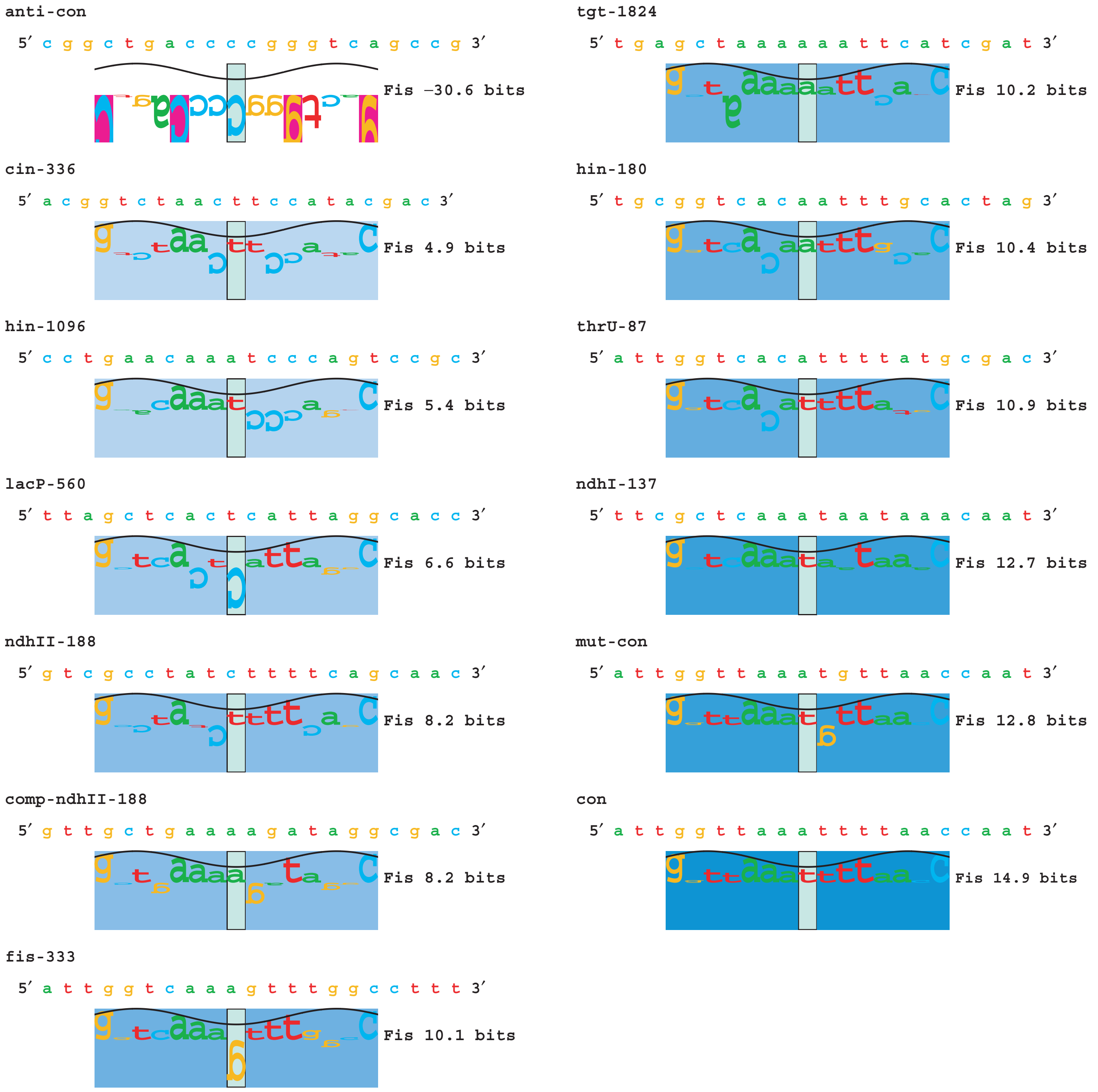
anti-con
5′ c g g c t g a c c c c g g g t c a g c c g 3′
Fis −30.6 bits
cin-336
5′ a c g g t c t a a c t t c c a t a c g a c 3′
Fis 4.9 bits
hin-1096
5′ c c t g a a c a a a t c c c a g t c c g c 3′
Fis 5.4 bits
lacP-560
5′ t t a g c t c a c t c a t t a g g c a c c 3′
Fis 6.6 bits
ndhII-188
5′ g t c g c c t a t c t t t t c a g c a a c 3′
Fis 8.2 bits
comp-ndhII-188
5′ g t t g c t g a a a a g a t a g g c g a c 3′
Fis 8.2 bits
fis-333
5′ a t t g g t c a a a g t t t g g c c t t t 3′
Fis 10.1 bits
tgt-1824
5′ t g a g c t a a a a a a t t c a t c g a t 3′
Fis 10.2 bits
hin-180
5′ t g c g g t c a c a a t t t g c a c t a g 3′
Fis 10.4 bits
thrU-87
5′ a t t g g t c a c a t t t t a t g c g a c 3′
Fis 10.9 bits
ndhI-137
5′ t t c g c t c a a a t a a t a a a c a a t 3′
Fis 12.7 bits
mut-con
5′ a t t g g t t a a a t g t t a a c c a a t 3′
Fis 12.8 bits
con
5′ a t t g g t t a a a t t t t a a c c a a t 3′
Fis 14.9 bits

Figure 2. Sequences used for analysis. Information analysis for individual sequences was computed (3) and displayed using sequence walkers (31). Those positions that are favored according to the $R_{iw}(b,l)$ weight matrix (contribute positive information) are represented by bases above the *x*-axis, whereas those bases that are not favored (contribute negative information) are below the *x*-axis. The height of each base is its information contribution to the site strength. The sum of all base heights is R_i for the sequence, and this is given on the right of the sequence walker. These sequences correspond to those in Table 1. The sequences are sorted by their strength in bits and the saturation of a colored rectangle behind each walker is proportional to that strength. As in Figure 1, the sine wave above the walker represents the 10.6 bp helical twist of B-form DNA (24).

5′-biotin tag (Synthegen, LLC) to allow immobilization of the oligos onto NeutrAvidin (NA)-coated sensor chips (B1 chips, Biacore Inc.). To test whether the orientation of a sequence in the hairpin affects binding, we inverted the ndhII-188 sequence in the hairpin to create comp-ndhII-188.

SPR analysis

NeutrAvidin was purchased from Pierce. EDTA, SDS, NaCl and HEPES (pH 7.4) were purchased from Invitrogen. Potassium glutamate was purchased from Sigma-Aldrich. Tris-HCl (pH 7.5) was purchased from Quality Biological, Inc. Binding experiments were performed on Biacore 2000 and Biacore 3000 instruments (Biacore Inc.). NeutrAvidin was diluted to a final concentration of 25 g/ml in 10 mM sodium acetate, pH 4.5. An immobilization wizard within the Biacore control software was used to immobilize no more than 4000 RU of NA. One RU, or resonance unit, corresponds to a change in the angle of the intensity minimum by 0.0001° as detected by the Biacore. The oligos were diluted to a final concentration of 1 mg/ml in immobilization buffer (10 mM Tris-HCl pH 7.5, 300 mM NaCl, 1 mM EDTA). To prepare double-stranded DNA, the oligos were heated to 95°C for 5 min, snap cooled on ice for 5 min, and incubated at room temperature for 15 min. The sample was then diluted 750-fold in immobilization buffer and injected manually over the surface until between 100 and 150 RUs were captured on the B1 sensor chip.

Purified Fis protein (22) was serially diluted in 1×running buffer (10 mM HEPES pH = 7.4, 350 mM potassium glutamate (40), 3.4 mM EDTA, 0.01% BSA) to concentrations ranging from 100 nM for the high affinity oligos to 1000 nM for the low affinity oligos and injected at 25°C at a flow rate of 100 μl/min for 90 s. All oligos reached a stochastic steady state of Fis binding. Dissociation times were typically 90–360 s depending upon the stability of the complex. Disruption of any complex that remained bound after dissociation was achieved using two 50 μl injections of regeneration solution (0.1% SDS, 3.4 mM EDTA) followed by one EXTRACLEAN command, a running buffer wash to eliminate carry-over into the next experiment. At the beginning of each cycle, the needle was pre-dipped in running buffer before an injection of 100 μl running buffer. Similarly, each cycle was ended by an injection of 100 μl running buffer and an EXTRACLEAN command. Typically, every concentration of protein was injected twice from separate vials. In order to subtract any background noise from each data set, all samples were also run over a sensor chip surface of NA without oligo and injections of running buffer were performed for every experiment ('double referencing') (41). Data were fit to a single exponential decay model using both of the programs Scrubber 1.10 (42) and Biaevaluation 3.1 (Biacore, Inc).

Fis competition electrophoretic mobility shift assay (EMSA)

Using EMSA, we found that nonspecific binding occurred with the long oligos used in the SPR experiments. Therefore we used hairpin oligos containing a Fis site (−7 to +7) with no additional bases, a loop (5′-GCGAAGC-3′) and the complementary sequence of the Fis site for EMSA. (See Supplementary Data Figure 1 for the sequences used.)

Competition EMSA between conF37, a 5′ 6-FAM labeled oligo 5′-GGTTAAATTTAACC-GCGAAGC-GGTTAAAATTTAACC-3′ (Integrated DNA Technologies) containing the consensus Fis binding site, and unlabeled oligos containing naturally occurring and mutated Fis binding sites, was used to determine the K_D of the sites. When a potassium glutamate-containing buffer was used for EMSA, Fis–DNA complexes smear on a gel, therefore we used the following buffer. Binding reactions were carried out in 10 μl of solution, containing 7.7 mM Bis Tris Propane-HCl, 10 mM NaCl, 0.5% glycerol, 10 mM $MgCl_2$, 1 mM DTT, 800 nM Fis, 40 nM labeled conF37 oligo and 1.0, 1.5 or 2.0 μM competitors for 5 min at room temperature, followed by 2.2% agarose gel electrophoresis in 5 mM sodium borate pH = 8.5 (43) for 20 min at 5 V/cm and the gel was scanned by a FMBIO II fluorescent scanner (Hitachi) with 505 nm emission filter (Figure 5). (See Supplementary Data for how the data were analyzed.)

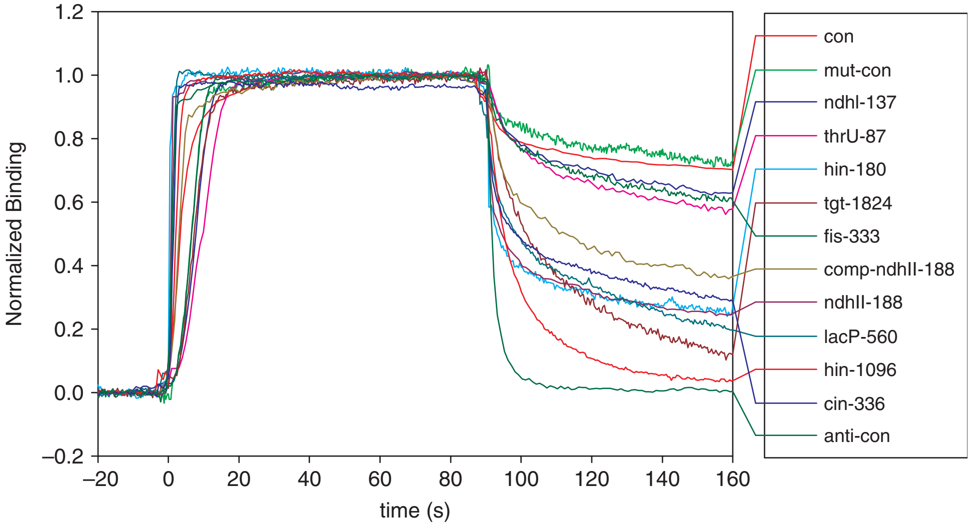

Figure 3. Sensogram of Fis bound to different DNA sequences. All curves were normalized so that saturation of the chip is set to 1. At time zero, Fis was washed onto the SPR chip. At time 90 s, Fis was washed off the chip. The stability measurements reported in Table 1 were determined from the curve after 90 s.

Table 1. Kinetics as determined by SPR

Oligo name	R_i (bits)	Number of experiments	Stability (s^{-1})	Reference
anti-con	−30.6	6	$2.21 \times 10^{-1} \pm 4.08 \times 10^{-3}$	This work
cin-336	4.9	2	$1.24 \times 10^{-1} \pm 2.48 \times 10^{-3}$	(33)
hin-1096	5.4	2	$7.39 \times 10^{-2} \pm 6.07 \times 10^{-4}$	(32)
lacP-560	6.6	4	$1.67 \times 10^{-2} \pm 6.49 \times 10^{-5}$	(34)
ndhII-188	8.2	7	$7.37 \times 10^{-3} \pm 1.10 \times 10^{-4}$	(35)
comp-ndhII-188	8.2	2	$6.24 \times 10^{-3} \pm 4.19 \times 10^{-5}$	(35)
fis-333	10.1	1	$3.45 \times 10^{-3} \pm 3.21 \times 10^{-5}$	(37)
tgt-1824	10.2	1	$2.62 \times 10^{-2} \pm 1.21 \times 10^{-4}$	(36)
hin-180	10.4	2	$7.83 \times 10^{-3} \pm 4.80 \times 10^{-5}$	(32)
thrU-87	10.9	1	$4.06 \times 10^{-3} \pm 6.54 \times 10^{-5}$	(38)
ndhI-137	12.7	1	$2.90 \times 10^{-3} \pm 5.80 \times 10^{-5}$	(35)
mut-con	12.8	6	$8.65 \times 10^{-4} \pm 4.29 \times 10^{-6}$	This work
con	14.9	1	$9.40 \times 10^{-4} \pm 1.39 \times 10^{-5}$	This work

'Oligo' is the name of the synthetic DNA hairpin as defined in this article or the name of the adjacent gene and base coordinate of the site in a GenBank entry (6). The sequences are given in Figure 2 and Supplementary Data Figure 1. R_i is the individual information for that site. 'Number of experiments' is the number of measurements made with each oligo. 'Stability' is the apparent k_{off} that we measure using SPR and analyzed with Scrubber. 'Reference' is the reference describing the binding of Fis to that sequence.

RESULTS AND DISCUSSION

The Fis sequence logo is consistent with models of the Fis/DNA complex (Figure 1) (6,44,22). Sequence conservation at positions ±7 above 1 bit suggests that Fis binds two major grooves on the same face of the DNA (24). However, the distance between the D helices which bind these two major contacts is less than 10.6 bases, one helical twist of B-form DNA, suggesting that the DNA must bend to enable positions ±7 to contact the D helices (24). Indeed, Fis bends DNA (44). The relatively low information content of 7.18 ± 0.23 bits over the range ±7 bases, suggests that Fis is a fairly prolific binder (2, 30). This is consistent with the observed high concentration of Fis in response to nutrient upshifts (as many as 50 000 dimers per cell) (37). Finally, DNA methylation and DNase I hypersensitivity results are consistent with positions of significant sequence conservation (6). The correspondence between the physical and biochemical characterization of Fis binding with the sequence conservation supports the information-theory based Fis binding model.

We chose ten naturally occurring Fis binding sites and three synthetic sites for kinetic analysis. These sites covered a spectrum of strengths and are reported in Figure 2. The terms anti-consensus (anti-con) and consensus (con) refer to the weakest and strongest possible sites based on our model respectively (3).

In order to measure the binding kinetics of these oligos, we used SPR technology. Protein can be flowed over a mat of DNA tethered to a thin gold surface. As the protein associates and dissociates, the change in density on the surface can be monitored, and k_{on} and k_{off} can be determined (17,45). The SPR plots appeared to have one-stage binding, suggesting a simple association–dissociation mechanism (Figure 3).

All data obtained for the Fis dimer (22.4 kDa) on the Biacore machine were transport limited (46). That is, the kinetics of binding that are inferred from these experiments are not only a measurement of binding, but also a measure of the delivery of Fis to the chip surface. However, we were able to measure an apparent k_{off} or 'stability' which is the rate of dissociation of Fis from the surface. Although this is not the true k_{off}, because of the transport limitation, it is proportional since the rate of transport (k_{t}) is constant for all measurements. Additionally, surface effects such as nonspecific interactions of Fis with the chip surface could affect the SPR measure so that it does not entirely represent *in vivo* or in-solution conditions, but as with the rate of transport, such effects should also be constant for all measurements.

The stability kinetics measurement is strongly correlated to the individual information of the sites, with $r^2 = 0.84$ (Figure 4). These values are presented in Table 1. The complexes of Fis with oligos ndhII and comp-ndhII had similar stabilities (7.4×10^{-3} and $6.2 \times 10^{-3} s^{-1}$, respectively) suggesting that orientation within the hairpin had little affect on the stability measurement. The dissociation of the protein from the anti-con oligo is faster than the dissociation from the weakest observed natural binder cin-336, 0.22 s^{-1} versus 0.12 s^{-1}. This is presumably related to the energy difference between the weakest possible specific binding and nonspecific binding for Fis. The stability of the protein with the consensus and mutated consensus is very high, 9.4×10^{-4} and $8.7 \times 10^{-4}\, s^{-1}$, respectively.

The logic of our experiment is based on a series of simple relations:

(1) Information is related to energy by a version of the Second Law of thermodynamics (13). The relationship is generally proportional (TDS in preparation) so we expect that the individual information should relate to the binding energy:

$$R_i \propto -\Delta G \qquad \mathbf{2}$$

This is supported by experiments in a number of systems (8,9,47).

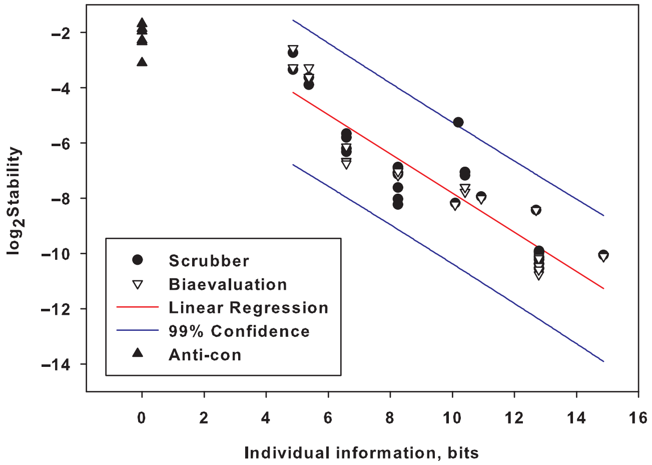

Figure 4. Binding site information is correlated to stability. For each sequence described in Figure 2 and Table 1, we plotted the stability versus the information R_i. Scrubber and Biaevaluation are two implementations of curve fitting by a single exponential decay describing the dissociation. Both were used to evaluate all of the data and slight differences were observed from small deviations in the start and stop points chosen for analysis. We plot each measurement independently. Although the anti-con oligo is presumably nonspecific at −30.6 bits, we plotted it as having 0 bits of information. All points at zero bits are for the anti-con oligo. The regression line (excluding the anti-con) is shown as a red line ($r^2 = -0.84$). 99% confidence limits for the regression are shown with blue lines. The equation for the regression line is $\log_2(\text{Stability}) = -0.70 \times$ Individual information −0.84.

(2) The binding energy is related to the binding constant:

$$\Delta G \propto \log K_\mathrm{D} \qquad \mathbf{3}$$

(3) The binding constant is a function of the on and off rates:

$$K_\mathrm{D} = k_\mathrm{off}/k_\mathrm{on} \qquad \mathbf{4}$$

(4) Once a protein is at a binding site, it will frequently bind irrespective of how strong the binding is, so the on-rate, k_on should be roughly constant and this is observed in various other genetic systems (14–16, 48).

(5) Combining the above

$$R_i \propto -\log k_\mathrm{off} \qquad \mathbf{5}$$

so the more information a binding site has, the larger the number of contacts it can make with the protein (49) and correspondingly the more difficult it becomes for thermal noise to separate the two once they are bound together. The off-rate is strongly dependent on the detailed binding contacts since all of these have to be broken to release the protein.

Although our Biacore experiments gave the relationship of Equation (5) (Figure 4), they did not give us k_on values. To investigate k_on, we performed competitive EMSA experiments to determine K_Ds (Figure 5). The results show a linear relationship between $\log_2(K_\mathrm{D})$ and R_i:

$$\log_2(K_\mathrm{D}) = -0.27 \times R_i - 26.14 \qquad \mathbf{6}$$

with $r^2 = 0.73$ (Supplementary Data). The experiment was repeated and similar results were obtained (data not shown).

Since k_offs and $K_\mathrm{D}s$ were measured by different techniques, the relative ratios between the sites should be correct but they may differ from the absolute values by an unknown multiplicative factor. On the log scale, this is in the additive constant. Using the K_Ds measured by EMSA and the k_offs measured in the Biacore experiments, we calculated k_on according to Equation (4) for each DNA. Unexpectedly, we observed that k_on is related to the information.

By using linear regression of $\log_2(k_\mathrm{on})$ against R_i and $\log_2(k_\mathrm{off})$ against R_i,

$$\log_2(k_\mathrm{on}) = -0.38 \times R_i + 25.35 \qquad \mathbf{7}$$

and

$$\log_2(k_\mathrm{off}) = -0.65 \times R_i - 0.79 \qquad \mathbf{8}$$

we found that 49% of the variance of $\log_2(k_\mathrm{on})$ and 78% of the variance of $\log_2(k_\mathrm{off})$ is explained by the variance of R_i (Supplementary Data). Thus most of the off-rate is explained by the information in the sequence. In addition, a good portion of the on-rate is explained by the sequence, implying that another factor—we suggest sequence bendability—may be involved in the initial binding.

Are the evolved binding targets of Fis the result of the physical properties of DNA? It is possible that the bases that are specifically contacted have been adapted through natural selection to facilitate binding through bending. If this is true, then there should be a correlation between k_on and k_off. Indeed, k_on and k_off increase together with a positive correlation

$$\log_2(k_\mathrm{on}) = 0.69 \times \log_2(k_\mathrm{off}) - 26.51 \qquad \mathbf{9}$$

and 85% of the $\log_2(k_\mathrm{on})$ variance is explained by $\log_2(k_\mathrm{off})$, suggesting that some of the positions are important for both binding and bending (Supplementary Data Figure 2).

This proposal is consistent with our previous observations on the sequence logo of Fis (22). We found that patterns of bases in the Fis sites can be explained in two distinct ways. In Figure 1, the outer bases at ±7, mostly G and C, are consistent with direct binding by Fis into the major groove but these contacts are too close to allow the D helices of Fis to fit into the major groove unless the DNA is also bent. Positions ±4 and ±3 contain pyrimidines and purines (respectively, on the $5' \rightarrow 3'$ strand) which could be contacted directly through the major groove or which could provide a bendable step. Likewise positions −2 to +2 contain A or T which is also consistent with either direct minor groove contacts or with bending into the minor groove. Since the central positions from −4 to +4 do not appear to be contacted in our 3D model (22), binding of Fis may first involve specific contacts followed by bending that perhaps releases those contacts. This implies that the binding rate requires DNA sequence-dependent bending. If so, k_on is controlled by the degree of flexibility of the DNA and that, in turn, is controlled by the DNA sequence. However, if Fis

Name	anti-con	cin-336	hin-1096	lacP-560	ndhII-188	comp-ndhII-188	fis-333	tgt-1824	hin-180	thrU-87	ndhI-137	mut-con	con	No competitor
Ri	**–30.6**	**4.9**	**5.4**	**6.6**	**8.2**	**8.2**	**10.1**	**10.2**	**10.4**	**10.9**	**12.7**	**12.8**	**14.9**	

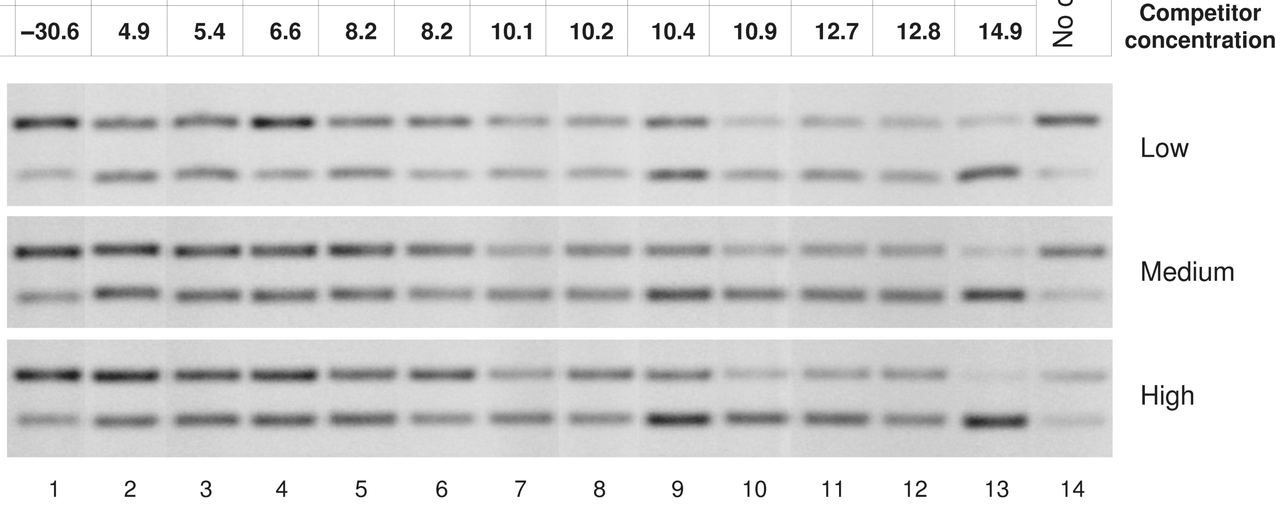

Figure 5. Competition electrophoretic mobility shift assay with three different concentrations of oligos containing different Fis binding sites. (See Supplementary Data Figure 1 for the sequences.) For each concentration, the top band is Fis bound to the consensus 5′ 6-FAM labeled oligo and the bottom band is unbound labeled oligo (see Materials and Methods). The competitor concentrations shown are approximately: 1.0 μM low, 1.5 μM medium, 2.0 μM high; the exact values for each competitor are given in Supplementary Data. Lanes 1 to 13: competitor oligos 1 to 13; Lane 14: no competitor.

makes direct contacts to the central bases while bound (despite our modeling) then DNA sequence should determine the strength of binding, and this is indeed observed. We are led to suggest that both bending and direct contacts are involved in both of the on- and off-stages of Fis binding. Similar experiments relating the information content of binding sites for other proteins that do not bend DNA as strongly as Fis may reveal further insights into the binding process.

The experiments described here suggest that R_i is mostly dependent upon the logarithm of k_{off}. It has previously been shown that the average R_i for all sites is R_{sequence}, the sequence conservation of a set of binding sites (3). Therefore, the results imply that the sequence conservation (the amount of variability among a set of binding sites) for a protein is directly related to the binding kinetics of that protein to its targets. A stronger binding protein that covers the same length of DNA will have a less variable site. Another aspect is that R_{sequence} evolves to match the information needed to find the sites in the genome, $R_{\text{frequency}}$, which is a function of the size of the genome and number of sites (2,30). As a protein evolves to bind a greater number of targets, the average specific binding energy of that protein to its targets would decrease by increased k_{off}.

Our experiment provides preliminary data supporting a distinction between two approaches to understanding the DNA recognition process. In Figure 4, no data points were obtained between the anti-consensus at −30.6 bits and 0 bits, however the lowest positive Fis site, at 4.9 bits

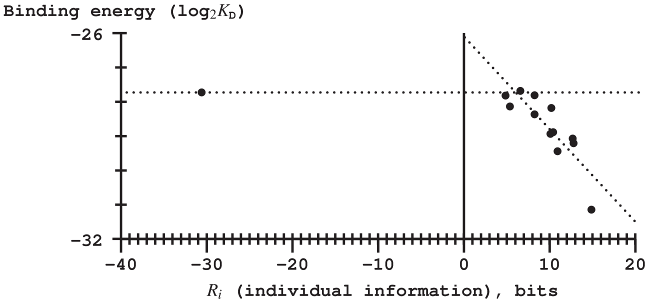

Figure 6. Binding energy is linearly related to binding site information for positive information binding sites but apparently flat for sites with negative information. The curve appears to break near zero bits. The average K_D values were normalized so that the Hin-180 sequence has the published value of Hin-D, 2×10^{-9} M (51). Excluding the anti-consensus at −30.6 bits, the regression line is given in Equation (6) with $r^2 = 0.73$.

has a log$_2$(stability) around −3 and the anti-consensus is around −2 so the curve is linear with a negative slope to near zero bits and then presumably is essentially flat from there to −30.6 bits. As shown in Figure 6, a similar result occurs with a plot of binding energy (log$_2$ K_D) versus information. We suggest that this apparent break at zero bits is a manifestation of the Second Law of Thermodynamics and the channel capacity. That is, the Second Law predicts that sites with positive information should have negative ΔG values and those with negative information should not bind because they have positive

ΔG values (13). Shannon's channel capacity theorem predicts threshold effects in coded systems where there is a sharp boundary between recognized and unrecognized signals (50). The break in the curve therefore provides support for a coding interpretation of the binding interaction between Fis and DNA. This is in contrast with thermodynamic theories of binding, which generate a scale starting at the consensus, and which do not predict a specific boundary (8).

The individual information appears to be well correlated to the kinetics of binding. This not only gives greater confidence in our previous information theory based models, but also shows that it is a reliable approach to characterize genetic systems *in silico*. Furthermore, the relationship between information and energy is subtle (13), and this correlation helps ground the information theory approach into thermodynamics.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Zehua Chen for testing the oligos, Alan Moses, Greg Bowman, Matt Fivash, Jack Kirsch and Danielle Needle for useful discussions. This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under contract N01-CO-12400. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products or organizations imply endorsement by the U.S. Government. This Research was supported [in part] by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research. Funding to pay the Open Access publication charges for this article was provided by NCI.

Conflict of interest statement. None declared.

REFERENCES

1. Record,M.T. Jr, Ha,J.H. and Fisher,M.A. (1991) Analysis of equilibrium and kinetic measurements to determine thermodynamic origins of stability and specificity and mechanism of formation of site-specific complexes between proteins and helical DNA. *Methods Enzymol.*, **208**, 291–343.
2. Schneider,T.D., Stormo,G.D., Gold,L. and Ehrenfeucht,A. (1986) Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, **188**, 415–431. http://www.ccrnp.ncifcrf.gov/~toms/paper/schneider1986/
3. Schneider,T.D. (1997) Information content of individual genetic sequences. *J. Theor. Biol.*, **189**, 427–441. http://www.ccrnp.ncifcrf.gov/~toms/paper/ri/
4. Shannon,C.E., (1948) A mathematical theory of communication. *Bell System Tech. J.*, **27**, 379–423, 623–656. http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html
5. Shultzaberger,R.K., Bucheimer,R.E., Rudd,K.E. and Schneider,T.D. (2001) Anatomy of *Escherichia coli* ribosome binding sites. *J. Mol. Biol.*, **313**, 215–228. http://www.ccrnp.ncifcrf.gov/~toms/paper/flexrbs/
6. Hengen,P.N., Bartram,S.L., Stewart,L.E. and Schneider,T.D. (1997) Information analysis of Fis binding sites. *Nucleic Acids Res.*, **25**, 4994–5002. http://www.ccrnp.ncifcrf.gov/~toms/paper/fisinfo/
7. Rogan,P.K., Faux,B.M. and Schneider,T.D. (1998) Information analysis of human splice site mutations. *Hum. Mutat.*, **12**, 153–171 Erratum in: Hum Mutat 1999;13(1):82. http://www.ccrnp.ncifcrf.gov/~toms/paper/rfs/
8. Berg,O.G. and 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.
9. Barrick,D., Villanueba,K., Childs,J., Kalil,R., Schneider,T.D., Lawrence,C.E., Gold,L. and Stormo,G.D. (1994) Quantitative analysis of ribosome binding sites in *E. coli*. *Nucleic Acids Res.*, **22**, 1287–1295.
10. Roulet,E., Bucher,P., Schneider,R., Wingender,E., Dusserre,Y., Werner,T. and Mermod,N. (2000) Experimental analysis and computer prediction of CTF/NFI transcription factor DNA binding sites. *J. Mol. Biol.*, **297**, 833–848.
11. Liu,X. and Clarke,N.D. (2002) Rationalization of gene regulation by a eukaryotic transcription factor: calculation of regulatory region occupancy from predicted binding affinities. *J. Mol. Biol.*, **323**, 1–8.
12. Udalova,I.A., Mott,R., Field,D. and Kwiatkowski,D. (2002) Quantitative prediction of NF-κB DNA-protein interactions. *Proc. Natl Acad. Sci. USA*, **99**, 8167–8172.
13. Schneider,T.D. (1991) Theory of molecular machines. II. Energy dissipation from molecular machines. *J. Theor. Biol.*, **148**, 125–137. http://www.ccrnp.ncifcrf.gov/~toms/paper/edmm/
14. Kim,J.G., Takeda,Y., Matthews,B.W. and Anderson,W.F. (1987) Kinetic studies on Cro repressor-operator DNA interaction. *J. Mol. Biol.*, **196**, 149–158.
15. Schaufler,L.E. and Klevit,R.E. (2003) Mechanism of DNA binding by the ADR1 zinc finger transcription factor as determined by SPR. *J. Mol. Biol.*, **329**, 931–939.
16. Linnell,J., Mott,R., Field,S., Kwiatkowski,D.P., Ragoussis,J. and Udalova,I.A. (2004) Quantitative high-throughput analysis of transcription factor binding specificities. *Nucleic Acids Res.*, **32**, e44.
17. Fisher,R.J., Fivash,M., Casas-Finet,J., Erickson,J.W., Kondoh,A., Bladen,S.V., Fisher,C., Watson,D.K. and Papas,T. (1994) Real-time DNA binding measurements of the ETS1 recombinant oncoproteins reveal significant kinetic differences between the p42 and p51 isoforms. *Protein Sci.*, **3**, 257–266.
18. Fisher,R.J., Fivash,M.J., Stephen,A.G., Hagan,N.A., Shenoy,S.R., Medaglia,M.V., Smith,L.R., Worthy,K.M., Simpson,J.T., *et al.*, (2006) Complex interactions of HIV-1 nucleocapsid protein with oligonucleotides. *Nucleic Acids Res.*, **34**, 472–484.
19. Rich,R.L. and Myszka,D.G. (2006) Survey of the year 2005 commercial optical biosensor literature. *J. Mol. Recognit.*, **19**, 478–534.
20. Travers,A., Schneider,R. and Muskhelishvili,G. (2001) DNA supercoiling and transcription in *Escherichia coli*: The FIS connection. *Biochimie*, **83**, 213–217.
21. Ussery,D., Larsen,T.S., Wilkes,K.T., Friis,C., Worning,P., Krogh,A. and Brunak,S. (2001) Genome organisation and chromatin structure in *Escherichia coli*. *Biochimie*, **83**, 201–212.
22. Hengen,P.N., Lyakhov,I.G., Stewart,L.E. and Schneider,T.D. (2003) Molecular flip-flops formed by overlapping Fis sites. *Nucleic Acids Res.*, **31**, 6663–6673.
23. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100. http://www.ccrnp.ncifcrf.gov/~toms/paper/logopaper/
24. Schneider,T.D. (1996) Reading of DNA sequence logos: Prediction of major groove binding by information theory. *Methods Enzymol.*, **274**, 445–455. http://www.ccrnp.ncifcrf.gov/~toms/paper/oxyr/
25. Schneider,T.D. and Rogan,P.K. (1999) Computational analysis of nucleic acid information defines binding sites, United States Patent 5867402.
26. Stormo,G.D., Schneider,T.D., Gold,L. and Ehrenfeucht,A. (1982) Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *E. coli*. *Nucleic Acids Res.*, **10**, 2997–3011.
27. Lukashin,A.V., Anshelevich,V.V., Amirikyan,B.R., Gragerov,A.I. and Frank-Kamenetskii,M.D. (1989) Neural network models for promoter recognition. *J. Biomol. Struct. Dyn.*, **6**, 1123–1133.

28. Weller,K. and Recknagel,R.D. (1994) Promoter strength prediction based on occurrence frequencies of consensus patterns. *J. Theor. Biol.*, **171**, 355–359.
29. GuhaThakurta,D. and Stormo,G.D. (2001) Identifying target sites for cooperatively binding factors. *Bioinformatics*, **17**, 608–621.
30. Schneider,T.D. (2000) Evolution of biological information. *Nucleic Acids Res.*, **28**, 2794–2799. http://www.ccrnp.ncifcrf.gov/~toms/paper/ev/
31. Schneider,T.D. (1997) Sequence walkers: a graphical method to display how binding proteins interact with DNA or RNA sequences. *Nucleic Acids Res.*, **25**, 4408–4415. Erratum: NAR 1998, 26(4):1135. http://www.ccrnp.ncifcrf.gov/~toms/paper/walker/
32. Glasgow,A.C., Bruist,M.F. and Simon,M.I. (1989) DNA-binding properties of the Hin recombinase. *J. Biol. Chem.*, **264**, 10072–10082.
33. Finkel,S.E. and Johnson,R.C. (1992) The Fis protein: it's not just for DNA inversion anymore (erratum). *Mol. Microbiol.*, **6**, 1023.
34. Pan,C.Q., Johnson,R.C. and Sigman,D.S. (1996) Identification of new Fis binding sites by DNA scission with Fis-1,10-phenanthroline-copper(I) chimeras. *Biochemistry*, **35**, 4326–4333.
35. Green,J., Anjum,M.F. and Guest,J.R. (1996) The *ndh*-binding protein (Nbp) regulates the *ndh* gene of *Escherichia coli* in response to growth phase and is identical to Fis. *Mol. Microbiol.*, **19**, 1043–1055.
36. Slany,R.K. and Kersten,H. (1992) The promoter of the *tgt*/*sec* operon in *Escherichia coli* is preceded by an upstream activation sequence that contains a high affinity FIS binding site. *Nucleic Acids Res.*, **20**, 4193–4198.
37. Ball,C.A., Osuna,R., Ferguson,K.C. and Johnson,R.C. (1992) Dramatic changes in Fis levels upon nutrient upshift in *Escherichia coli*. *J. Bacteriol.*, **174**, 8043–8056.
38. Bosch,L., Nilsson,L., Vijgenboom,E. and Verbeek,H. (1990) FIS-dependent trans-activation of tRNA and rRNA operons of *Escherichia coli*. *Biochim. Biophys. Acta*, **1050**, 293–301.
39. Lyakhov,I.G., Hengen,P.N., Rubens,D. and Schneider,T.D. (2001) The P1 phage replication protein RepA contacts an otherwise inaccessible thymine N3 proton by DNA distortion or base flipping. *Nucleic Acids Res.*, **29**, 4892–4900. http://www.ccrnp.ncifcrf.gov/~toms/paper/repan3/
40. Merickel,S.K., Sanders,E.R., Vazquez-Ibar,J.L. and Johnson,R.C. (2002) Subunit exchange and the role of dimer flexibility in DNA binding by the Fis protein. *Biochemistry*, **41**, 5788–5798.
41. Myszka,D.G. (1999) Improving biosensor analysis. *J. Mol. Recognit.*, **12**, 279–284.
42. Myszka,D.G. and Morton,T.A. (1998) CLAMP: a biosensor kinetic data analysis program. *Trends Biochem. Sci.*, **23**, 149–150.
43. Brody,J.R. and Kern,S.E. (2004) Sodium boric acid: a Tris-free, cooler conductive medium for DNA electrophoresis. *Biotechniques*, **36**, 214–216.
44. Yuan,H.S., Finkel,S.E., Feng,J.-A., Kaczor-Grzeskowiak,M., Johnson,R.C. and Dickerson,R.E. (1991) The molecular structure of wild-type and a mutant Fis protein: relationship between mutational changes and recombinational enhancer function or DNA binding. *Proc. Natl Acad. Sci. USA*, **88**, 9558–9562.
45. Myszka,D.G., Jonsen,M.D. and Graves,B.J. (1998) Equilibrium analysis of high affinity interactions using BIACORE. *Anal. Biochem.*, **265**, 326–330.
46. Karlsson,R. (1999) Affinity analysis of non-steady-state data obtained under mass transport limited conditions using BIAcore technology. *J. Mol. Recognit.*, **12**, 285–292.
47. Berg,O.G. and vonHippel,P.H. (1988) Selection of DNA binding sites by regulatory proteins. *Trends Biochem. Sci.*, **13**, 207–211.
48. Das,N., Valjavec-Gratian,M., Basuray,A.N., Fekete,R.A., Papp,P.P., Paulsson,J. and Chattoraj,D.K. (2005) Multiple homeostatic mechanisms in the control of P1 plasmid replication. *Proc. Natl Acad. Sci. USA*, **102**, 2856–2861.
49. Mirny,L.A. and Gelfand,M.S. (2002) Structural analysis of conserved base pairs in protein-DNA complexes. *Nucleic Acids Res.*, **30**, 1704–1711.
50. Shannon,C.E. (1949) Communication in the presence of noise. *Proc. IRE*, **37**, 10–21.
51. Pan,C.Q., Finkel,S.E., Cramton,S.E., Feng,J.A., Sigman,D.S. and Johnson,R.C. (1996) Variable structures of Fis-DNA complexes determined by flanking DNA-protein contacts. *J. Mol. Biol.*, **264**, 675–695.