

In Vitro Selection of Integration Host Factor Binding Sites

STEVEN D. GOODMAN,^{1*} NERISSA J. VELTEN,¹ QIAN GAO,¹
SCOTT ROBINSON,² AND ANCA M. SEGALL²

*Department of Basic Sciences, University of Southern California School of Dentistry, Los Angeles, California,¹
and Department of Biology, San Diego State University, San Diego, California 92182-4614²*

Received 24 September 1998/Accepted 9 March 1999

Integration host factor (IHF) is a bacterial protein that binds and severely bends a specific DNA target. IHF binding sites are approximately 30 to 35 bp long and are apparently divided into two domains. While the 3' domain is conserved, the 5' domain is degenerate but is typically AT rich. As a result of physical constraints that IHF must impose on DNA in order to bind, it is believed that this 5' domain must possess structural characteristics conducive for both binding and bending with little regard for specific contacts between the protein and the DNA. We have examined the sequence requirements of the 5' binding domain of the IHF binding target. Using a SELEX procedure, we randomized and selected variants of a natural IHF site. We then analyzed these variants to determine how the 5' binding domain affects the structure, affinity, and function of an IHF-DNA complex in a native system. Despite finding individual sequences that varied over 100-fold in affinity for IHF, we found no apparent correlation between affinity and function.

In solution, B-form DNA is isotropically flexible only when it vastly exceeds its persistence length (150 bp; reviewed in reference 8). Nevertheless, DNA binding proteins often impose an anomalous structure onto their DNA targets despite typically contacting fewer than 30 bp. These deformations can be dramatic, however, such that the generated bend itself has a functional consequence. Since these DNA bending proteins generally require no coupled energy source for binding, the deformations are probably derived from the favorable thermodynamics of protein-DNA interactions. Proteins accomplish this feat by using various mechanisms, for example, charge neutralization of phosphates and/or destabilization of DNA base-stacking interactions (25, 30).

What role does the DNA sequence play in its own distortion? While specific protein-DNA interactions are most often mediated via specific hydrogen bonding between DNA bases and amino acid residues, there is an ever-growing class of proteins that select their target site based on indirect readout, i.e., a structure or subset of structures that are specifically recognized by a protein (34).

The bacterial protein integration host factor (IHF) epitomizes such proteins. It is a small heterodimeric protein consisting of homologous subunits, α and β , that binds and bends DNA specifically. Although known DNA targets are selected over bulk DNA by at least 1,000-fold (40), there is an inherent degeneracy of sequence in the target selection. IHF was originally discovered as a protein required for efficient integration of the bacteriophage lambda into the *Escherichia coli* chromosome (36). Subsequently, IHF has been shown to participate in virtually every type of nucleoprotein system (e.g., transcription, replication, and recombination). While its role is always that of an accessory factor, its involvement can be anything from a nominal 2-fold influence, as in promoter activation (23), to as great as a 10,000-fold effect, as in lambda integrative recombination.

It is the complex of IHF with DNA that makes it unique

among bending proteins; the recently solved cocrystal structure of IHF bound to one of its natural binding sites shows that the DNA is bent by 180° into a virtual U-turn (24, 25). IHF typically binds to a 30- to 35-bp sequence which can be divided into at least two domains. The 3' region is significantly conserved, and sites share the consensus WATCAANNNTTR (where W is A or T, R is purine, and N is any base). The lone cytosine base is conserved in every known natural site. Unlike the 3' region, the 5' region seems almost random; natural sites are typically AT rich, but no obvious patterns of sequence emerge as conserved.

How does the IHF binding site accommodate specific binding and bending? Of particular interest is the 5' binding domain, since it appears variant in each sequence. Recent evidence suggests that this domain is most successful at binding when an appropriately positioned run of adenines, or an A-tract, is present (11). A-tracts typically consist of three to six consecutive adenines that, in the proper sequence context, create an intrinsically rigid structure with a narrow minor groove; adjacent sequences typically possess an anisotropic bend. What facets of A-tracts, if any, attract IHF for preferred binding?

We have attempted to answer these questions by performing a systematic evolution of ligands by exponential enrichment (SELEX) analysis. Selective pressure in vitro for high-affinity binding was applied to a population of IHF binding sites where the wild-type 3' domain was held constant and the 5' domain was randomized. Individual sequences were then compared for binding affinity, gross structure of nucleoprotein complexes, and the ability to function in bacteriophage lambda site-specific recombination. We found that the 5' region can vary the affinity for IHF at least 100-fold. Although comprehensive rules for these sequence determinants could not be deduced, this region contributes significantly to the structure but little to the function of these nucleoprotein complexes. Finally, we found that while the base composition of this region is skewed in native sites, there appears to be no gross base composition advantage for either affinity or function.

MATERIALS AND METHODS

Bacterial strains and plasmids. *E. coli* DH5 was the host for all the plasmids used in this work. pHN868, pHN872, and pHN873 were derived from pBR322

* Corresponding author. Mailing address: Department of Basic Sciences, University of Southern California School of Dentistry, 925 West 34th St., Los Angeles, CA 90089. Phone: (213) 740-3867. Fax: (213) 740-7560. E-mail: sgoodman@hsc.usc.edu.

and have been described previously (4). pHN868 possesses a 1.0-kb *HindIII-BamHI attR*-containing insert. pHN872 and pHN873 contain two inserts, a 1.2-kb *PstI* segment from pUC4K, which confers kanamycin resistance, and a 1.0-kb *PstI-BamHI attL*-containing segment. pHN872 contains the wild-type *attL*, and pHN873 contains a mutant *attL* (OH') possessing four missense mutations (A37C, A38C, T43G, and T44G) at the H' site.

SELEX. Oligonucleotides were purchased from the University of Southern California Microchemical Core Facility. Our randomized H' pool of *attL* was constructed from a starting template, oSG42L, with the sequence 5' GCC TGC TTT TTT ATA CTA AGT TGG CAN NNN NNN NNN NNN NNN NNN NNC AAT TTG TTG CAA CGA ACA GGT CAC TA 3'. This sequence corresponds to the top-strand bases -11 to +62 of *attL*. We found that we could amplify an intact *attL* (119 bp) by using this single-stranded template and two additional oligonucleotides. oSG47 (5' GGA ATT CAA ATA ATG ATT TTA TTT TGA CTG ATA GTG ACC TGT TCG TTG CAA CAA ATT G 3') possesses 27 bases of complementary DNA sequence to oSG42L in addition to sequence that encompasses the entire 3' end of the *attL* locus (+88) and five additional bases which create an *EcoRI* restriction site. The third oligonucleotide, oSG46 (5' GGA ATT CCG TTG AAG CCT GCT TTT TTA TAC TAA GTT GGC 3'), corresponds to bases -20 to +13 of *attL* and also possesses an *EcoRI* linker. Exponential PCR amplification of a population of *attL*s was then initiated under the following conditions: strands were melted at 95°C for 3 min, and *Taq* polymerase (Promega) was added. Subsequent temperature cycling was performed by heating to 94°C for 40 s, annealing at 47°C for 30 s, and extending at 72°C for 40 s. This first amplification to create duplex starting material was performed for only five cycles.

Once the *attL* population was prepared, it was used as a substrate in an electrophoretic mobility shift assay (EMSA). Twenty-one replicates of 300 ng (80 pmol total) of SELEX-derived *attL* were each incubated with 60 nM IHF in 50 mM Tris-Cl-50 mM KCl-50 µg of bovine serum albumin (BSA) per ml-3.75 µg of salmon sperm DNA per ml-10% glycerol-1 mM EDTA at pH 7.8 for 20 min at room temperature in a final volume of 20 µl. Each reaction mixture was loaded onto an 8% nondenaturing polyacrylamide gel (29:1 acrylamide-to-bisacrylamide ratio) and electrophoresed (6 to 10 V/cm) for 2 to 3 h. The gels were then stained with ethidium bromide solution (0.5 µg/ml) for 1 h and visualized under long-wave UV light (365 nm). All DNA that was chased into the vicinity (± 0.5 cm) of the wild-type shifted product was excised and purified from the gel fragment by the crush-and-soak method (26). DNA concentrations were estimated by measurement of absorbance at 260 nm.

The amplification portion of the SELEX strategy was then used. Gel-purified IHF-shifted DNA was used as a template for PCR with oligonucleotides oSG46 and oSG49 (5'GGA ATT CAA ATA ATG ATT TTA TTT TGA CTG ATA GTG ACC TGT TCG 3', a shortened version of oSG47 and one with a melting temperature more similar to that of oSG46) at a final concentration of 400 nM. Shifted DNA from the first round was added to 3 µg/ml. The cycling conditions for amplification were identical to that of the original SELEX substrate reaction, except that nine cycles were performed. The small number of cycles was crucial since we found that under our conditions the polymerizing reaction started to fail significantly beyond nine cycles. This results in increasing formation of a heteroduplex, which does not seem to be a substrate for IHF binding (data not shown). In addition, we found that the highest yield was attained by keeping the reaction volume to 100 µl but performing 20 reactions in parallel. PCR products were pooled, and unincorporated oligonucleotides were removed by using the QIAquick PCR purification kit (Qiagen Inc.). All subsequent rounds of SELEX EMSA were performed the same way.

Cloning *attL*. PCR products were cloned into the *EcoRI* site of pBR322 by using the ligation express kit (Clontech) under conditions described by the manufacturer and transformed into *E. coli* DH5. Plasmid DNA from transformants was analyzed by restriction and DNA sequencing. Once a unique *attL*-containing plasmid was isolated, the *attL* portion was amplified by PCR as described above with oligonucleotides oSG46 and oSG49. QIAquick-purified *attL*s were eluted in 10 mM Tris-1 mM EDTA (TE), pH 8.0.

Isotopic labeling of DNA. Linear DNA was isotopically labeled with ³²P in either of two ways. For PCR amplicons, oligonucleotides were first labeled isotopically with [γ -³²P]ATP by using the RTS T4 kinase-labeling system (Life Technologies Inc.). These labeled oligonucleotides were then used in subsequent PCRs (see below). For restriction fragments with 5' protruding ends, a fill-in reaction was used with Sequenase version 2.0 and an α -³²P-labeled deoxynucleoside triphosphate. For either labeling method, free nucleotides were removed with the QIAquick nucleotide removal kit (Qiagen Inc.) or on Sephadex G-50 spin columns (26).

Quantitative EMSA. Substrates were synthesized by PCR with oSG46 and oSG49, using wild-type *attL*-containing plasmid (pHN872) as a template. Quantitative EMSAs with our wild-type *attL* were performed similarly to the experiments described by Yang and Nash (40). Labeled *attL* PCR amplicon (119 bp) and IHF to 10 nM were added simultaneously to 50 mM Tris-HCl (pH 7.8)-60 mM KCl-50 µg of BSA per ml-10% glycerol to a final volume of 20 µl. The reaction was allowed to reach equilibrium by incubation at 25°C for 40 min. For competition experiments, EMSA mixtures were similarly assembled except that IHF was limiting at 50 pM. The DNA substrate consisted of 10 fmol of ³²P-labeled *attL* with a variable amount of unlabeled competitor DNA.

In all cases, the reaction mixtures were loaded onto an 8% polyacrylamide gel

immediately after incubation. IHF nucleoprotein complexes were separated from free DNA by electrophoresis (10 V/cm) in 0.5× Tris-borate-EDTA (TBE) (26). Dried gels were then exposed to a PhosphorImager screen (Molecular Dynamics) and quantitated with ImageQuant software. The fraction of DNA shifted into complex was estimated by dividing the amount of complexed DNA by the total amount. For each competition experiment, a range of competitor concentrations were used. Curve fitting with Cricket software was used to estimate the 50% inhibitory concentration (IC₅₀).

Quantitative in vitro recombination. All recombination reactions were performed in 25 to 50 mM Tris-Cl (pH 7.8)-60 to 70 mM KCl-250 µg of BSA-5 mM spermidine-0.5 mM EDTA-10% glycerol in a final volume of 15 µl. Purified Int, IHF, and Xis were gifts from Howard Nash. In all the reactions, Int and IHF were added to a final concentration of 70 to 140 nM and 50 nM respectively, unless specified. In excisive recombination reactions, Xis was added to 30 to 60 nM. Plasmid DNA substrates were extracted and purified with the QIAprep spin miniprep kit (Qiagen). Substrates derived from PCR amplicons (typically 100 to 200 bp) were purified of free nucleotides and unincorporated oligonucleotides via the QIAquick PCR purification kit. The concentration of DNA substrates was typically 1 to 2 nM. Salmon sperm DNA was added to 20 µg/ml in reactions with PCR-derived substrates. The reaction mixtures were always incubated at 25°C for the times indicated. In some cases, reactions were terminated by incubation at 60°C for 5 min and MgCl₂ was added to 10 mM along with select restriction enzymes. Digests of the DNA followed by gel electrophoresis were performed to distinguish substrates from products.

Construction of *attPs*. Amplicons (119 bp) of select *attL*s along with the wild-type *attR*-containing plasmid, pHN868, were used as substrates for in vitro excisive recombination. Recombination reactions were performed as described above, except that the incubations were extended to 3 h and the reaction mixture sizes were doubled. Recombination yielded an integrated linear-form DNA molecule the sum of the sizes of the two constituent substrates and featured an *attB* and an *attP* (5.1 kb). Following the incubations, the reaction products were digested with *Bam*HI, which separated the *attB* (0.8 kb)- from the *attP* (4.3 kb)-containing fragments. The ends of the DNA were filled in with T4 DNA polymerase. The restriction fragments were gel purified, self-ligated, and transformed into *E. coli*. Plasmid DNA isolated from transformants was mapped by restriction enzyme analysis to verify that the *attPs* contained a 350-bp *EcoRI* fragment indicative of the recombinant *attP*.

Hydroxyl radical footprinting. Footprinting reactions were performed on PCR-amplified *attL* with oligonucleotides oSG46 and oSG49. To examine the top-strand protections, oSG46 was labeled isotopically with [γ -³²P]ATP, and ³²P-labeled and unlabeled oSG46 and unlabeled oSG49 were used to PCR amplify the designated *attL*s. Hydroxyl radical footprinting was carried out as described by Yang and Nash (38).

RESULTS

Experimental strategy. We created a population of randomized IHF binding sites and used IHF to select for high-affinity DNA targets. We chose the H' site of bacteriophage lambda *attL/attP*, which is required for efficient phage recombination, as the parental target DNA. IHF requires a region comprising 32 bp for maximum binding (+15 to +46 of the *attL/attP* locus of bacteriophage lambda). The H' site was chosen because it has been extensively mutagenized and studied (2, 4, 6, 9-11, 15-18, 21, 24, 27, 28, 35, 38-40) and because its structure bound to IHF was recently solved by X-ray crystallography (24). IHF bound to the H' site has been shown to act as an architectural element; similar deformations of DNA at this site replace IHF for function (6, 27).

To identify high-affinity targets, we used a version of the SELEX protocol (32). Briefly, this entails synthesizing a DNA template in which the ends have a defined sequence and surround a region of random sequence. The selection is based on binding IHF and partitions high-affinity members from low-affinity members. High-affinity members are amplified by PCR, and the selection-amplification cycle is repeated until the selection has enriched the population to a suitable percentage of high-affinity members. Although we had desired to randomize an entire 32-base site, this turned out to be impractical for sites in excess of 22 bases due to the large quantity of DNA and protein required to ensure a completely random population. Instead, we focused on the 21 bp comprising the 5' region of H', +15 to +35 of *attL-attP*, next to the conserved cytosine at +36. The ends of the amplifying oligonucleotides were de-

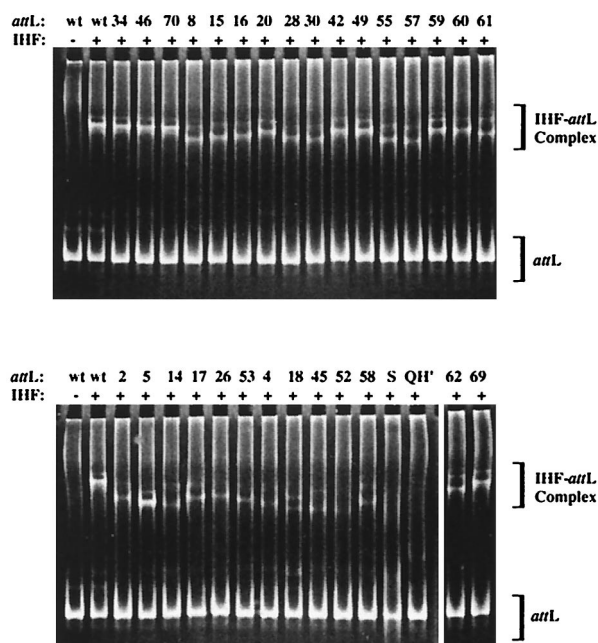


FIG. 2. EMSA of *attL* with IHF. Each *attL* (200 nM) was incubated in the presence of IHF (60 nM) for 40 min at 25°C. Each reaction mixture was then resolved by polyacrylamide gel electrophoresis. The gels were stained in ethidium bromide and visualized under 302-nm light. Complexed DNA has a retarded migration pattern relative to free DNA. Numbers atop each lane indicate the *attL*. Lanes marked wt contain wild-type *attL*. Lane S contains the original *attL* population from which selection was performed. Lane QH' contains *attL* with four base mutations in the 3' binding domain (see the text).

PCR amplicons of each of the 37 *attL*s were then analyzed in an EMSA with IHF under the same conditions as used for the original selection. Only 28 of the *attL*s were capable, to different degrees, of forming discrete complexes. We refer to these *attL*s as class I. The remaining nine that failed to shift in the presence of IHF (data not shown) have consequently been designated class II. As shown in Fig. 2, class I members yielded complexes that migrated to different degrees. Only three *attL*s, 1-34, 1-46, and 1-70, had gel shifts indistinguishable from that of wild-type *attL*. No *attL* had a greater shift than the wild type.

A cursory inspection of the class I *attL*s showed that specific portions of each sequence were conserved (Fig. 3 and 4). Based on the original design of the IHF site population, 21 bases 5' to the conserved cytosine (+36) were randomized. Others have identified the 3 bases 5' from this cytosine as being highly conserved in natural sites (7). In fact, Craig and Nash (2) originally defined the consensus as WATCAANNNTTR (where W is A or T, R is purine, and N is any base). We found that all class I *attL*s except one (*attL* 1-5) possessed the sequence WWT 5' to the +36 cytosine. In addition, 17 of these 28 members possessed an adenine directly 5' to this sequence, yielding a conserved sequence of AWWT 5' to the +36 cytosine (bases +32 to +35). These three bases were much less highly conserved among class II members, with only 1-36 yielding an exact match. Thus, selection for affinity and the screening for discrete complexes during EMSA mostly coincided with the canonical 3' consensus.

Despite these matches to published consensus sequences, a plethora of other sequence motifs were observed in our populations, demonstrating that high affinity and gross structure could be accommodated by multiple sequences. In fact, for the 21 randomized bases, none of the 37 *attL*s possessed more than

12 identical bases and two contained as few as 3 identical bases compared to the wild-type *attL* sequence.

Within the wild-type H' sequence, the most conspicuous sequence motif is an A tract between bases +19 and +24 that others have speculated enhances DNA binding via an accommodating sequence-directed architecture (11, 28). Several groups have demonstrated that point mutations within the A tract cause a significant and sometimes severe binding defect (9, 11, 16, 18, 28). To the best of our knowledge, no mutation in this region has ever enhanced DNA affinity over the wild type. Only three class I *attL*s (1-42, 1-15, and 1-14) possessed an A tract at these coordinates. In each case, the length of the A tract was 4 adenines or less. Whether these adenine residues are important for binding in these variants is not known, but the lack of A tracts in the other high-affinity class I members precludes this motif from being essential.

As shown in Fig. 4, additional conservations among 27 class I members occurred at bases +14, +16, +18, +19, +22, and +23. By oligonucleotide design, base +14 should have been exclusively adenine. The fact that it varied within our population indicates that our starting oligonucleotides were not homogeneous and/or there were additional PCR artifacts. Indeed, several sequenced isolates that were not further analyzed possessed either additions or deletions within the *attL* fragment (data not shown). Since we did not examine the proportion of adenines at +14 in the unselected starting material, variability at +14 may not be relevant. Among the most highly conserved bases were those at +18 and +19, where at least half of the bases were guanine, and base +22, which was mostly adenine (16 of 27 bases). These conservations were absent in the class II members. With the exception of the adenine at base +22, these conservations were also absent in a published alignment of 27 natural sequences (7).

Quantitative binding analysis of *attL*-IHF interactions. To further characterize and classify our SELEX-derived *attL* populations, we determined the apparent equilibrium dissociation constants (K_d) by a competition assay for all members of each class as well as the original randomized SELEX starting material. To start, a binding isotherm was generated by using an EMSA with a constant amount of IHF and increasing concentrations of our wild-type *attL* substrate (119 bp). Based on the amounts of free and bound DNA, we were able to estimate the concentration of free *attL* which yields half the maximal concentration of bound complex, the K_d , for the wild-type substrate at 0.7 nM (40). We measured the K_d of all variant sites by a competition method: isotopically labeled wild-type *attL* (0.5 nM) was incubated with limiting amounts of purified IHF (50 pM) in the presence or absence of different amounts of a specific competitor *attL*. Dissociation constants were based on the measured IC_{50} (the concentration of the competitor that yielded 50% inhibition of the wild-type shifted complex in the absence of competitor [see Materials and Methods]). With the wild-type *attL* as the competitor, the average IC_{50} in three trials was 0.8 nM, very similar to the value we measured by using the more direct approach of a binding isotherm. The IC_{50} s determined for each SELEX-derived *attL* are shown in Fig. 3.

Each class had distinctive affinity characteristics. Consistent with our preliminary designations, class I members had a higher average affinity than class II members did. No SELEX-derived site had a higher affinity than the wild-type sequence, although site 1-69 had a binding affinity within a factor of 3 of the wild type. Although the average affinity of each class was distinct, there was overlap between individual members of class I and class II. The worst binder, 1-47, had only a threefold-lower affinity than did the worst class I binder. It is noteworthy,

TABLE 1. Quantitative recombination

<i>attL</i>	% Excisive recombination ^a	% Integrative recombination ^b
Wild type	100	100
QH'	34	7
SELEX substrate	4	ND ^d
Class I ^c		
1-69	28	ND
1-42	58	100
1-5	<2	ND
1-70	25	33
1-46	50	35
1-49	30	ND
1-34	44	ND
1-15	31	ND
1-62	34	ND
1-8	40	ND
1-61	46	67
1-14	32	ND
1-20	57	ND
1-16	18	67
1-55	45	39
1-28	42	31
1-60	52	ND
1-57	30	ND
1-30	30	ND
1-18	23	ND
1-58	26	ND
1-26	50	ND
1-2	28	15
1-59	42	80
1-53	56	35
1-17	13	75
1-4	39	ND
1-52	26	12
Class II ^c		
1-36	50	110
1-50	13	8
1-29	15	ND
1-65	15	9
1-40	16	ND
1-45	43	ND
1-48	28	ND
1-54	23	ND
1-47	42	85

^a Recombination reaction mixtures were incubated at 25°C for 20 min. All recombination efficiencies are the average of two trials and were normalized against the wild-type *attL*. 100% is equivalent to 19% product formation.

^b Recombination reaction mixtures were incubated at 25°C for 60 min. All recombination efficiencies are the average of two trials and were normalized against the wild-type *attP*. 100% is equivalent to 19% product formation.

^c Columns are listed in decreasing order of IHF binding affinity within each class.

^d ND, not determined.

still bound well, consistent with our estimate of high-affinity binding. We propose that the noncanonical binding pattern for 1-5 is due to the presence of a new IHF site out of register with the original wild-type H' site (see below). This easily explains the complete failure of this site in excisive recombination.

DISCUSSION

The IHF protein is highly conserved among gram-negative eubacteria (19, 22). This conservation extends beyond apparent amino acid sequence homology to a specific recognition sequence and subsequent structural deformation of the bound

DNA target. Despite DNA site specificity, DNA targets possess significant amounts of sequence degeneracy between sites. This degeneracy is most apparent in the 5' region of the DNA target. In this work, we studied variants of the lone natural site from the bacteriophage lambda recombination locus *attL*. Using a minimal *attL* locus (119 bp) as a template, we generated DNA targets for IHF de novo via the SELEX strategy. By this approach, we randomized the 5' region of the IHF site while keeping the 3' region and the remainder of the *attL* locus unaltered. Specific members of the population were initially selected for affinity with IHF by EMSA. Selected *attL*s were subsequently assessed individually in EMSA for affinity and gross structure and as substrates in recombination assays. Although the affinity of IHF for individual members of our selected sequences varied over 100-fold, we could find no relationship to recombination function.

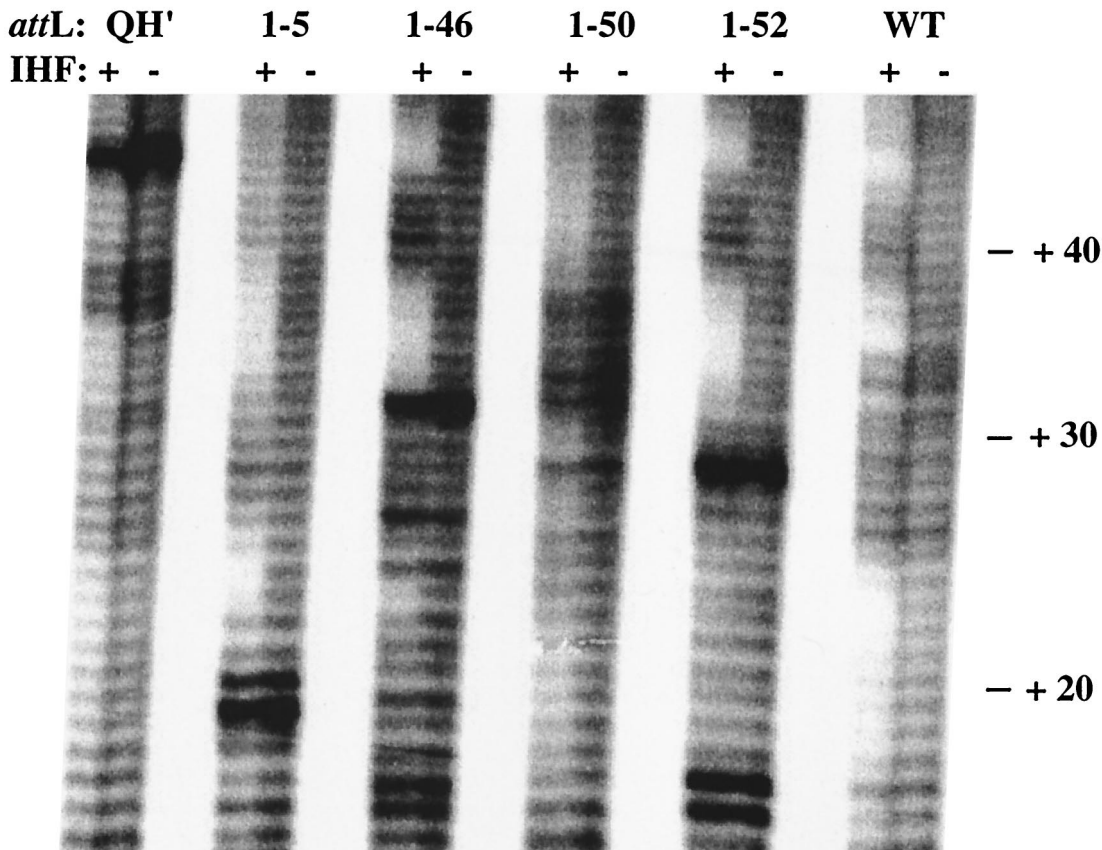
Our SELEX selection yielded IHF binding species that, as a group in complex with IHF, produced a smear during EMSA. We found that the smear consisted of complexes with the original population of SELEX-derived *attL*s and was, at least in part, the result of a population of distinctly shifted IHF-*attL* complexes with different migrations. No members, complexed with IHF, produced shifts even more retarded than the wild-type *attL*-IHF complex. In addition, none of the 37 members had an altered migration on polyacrylamide gel electrophoresis in the absence of IHF, indicating that the altered migration was indeed due to the nucleoprotein complex. We conclude that the 5' region of the binding site plays a critical role in IHF binding and in nucleoprotein complex structure.

The affinity of each SELEX-derived *attL* substrate was measured by competition with the wild-type *attL* for IHF. This obviates the necessity for the mutant sites to form an electrophoretically stable complex. We found that the ability to compete for binding was not absolutely related to the ability to form a stable complex with IHF during gel electrophoresis. No member of our population bound with greater affinity than the wild-type site, although the strongest binders had IC_{50} s within a factor of 3. Interestingly, when the wild-type 5' binding domain of H' was left intact and the conserved 3' region was mutated at the most conserved residues (QH'), binding was reduced to a level 14-fold lower than the wild type and no shifted bands were observed during EMSA. In addition, the IC_{50} s of the starting material and the lowest-affinity *attL*s were 50- and 100-fold higher, respectively, demonstrating the critical contribution of the 5' region to affinity. These results are consistent with binding affinity being distributed between the 5' and 3' domains.

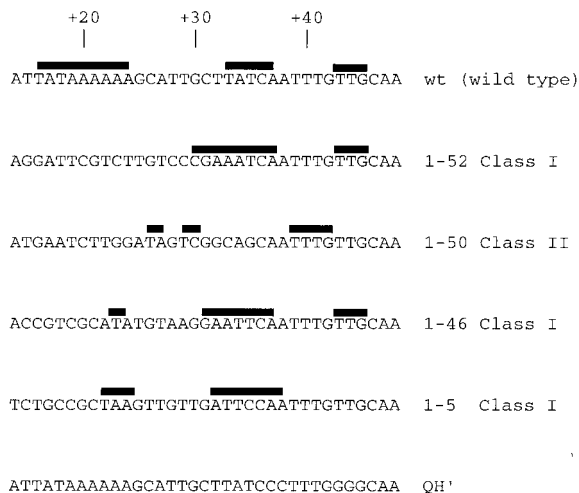
In general, the functionality of the SELEX-derived sites in excisive or integrative lambda site-specific recombination does not correlate with their affinity for IHF. It is very likely that the recombination defect was ameliorated in the presence of Int. We have previously shown that the nucleoprotein intermediates on various *att* sites are codependent on Int and IHF to facilitate the formation of the appropriate structure (27); Int cooperates even with the nonspecific bending protein HU to make a functional nucleoprotein complex. Since the nucleoprotein complex between Int, IHF, and *attL* or *attP* is the functional substrate for recombination, it seems more likely that the variant sequences are applying their effect on recombination via these nucleoprotein complexes rather than any intermediate complex between either *att* site and IHF. We are beginning to investigate the role that Int may play in mitigating the defects of our H' variants in recombination.

Of the 40 *attL*s tested, only *attL* 1-5 was inactive as a substrate for excisive recombination. IHF binds to this site with high affinity, suggesting that complex formation per se does not

A



B



C

```

                    WATCAANNNTTR
5' TCTGCCGCTAAGTTGTTGATTCCAATTTGTTGCAA 3'
3' AGACGGCGATTCAACAACCTAAGGTTAAACAACGTT 5'
      RRTNNNNACTAW
    
```

FIG. 5. Hydroxyl radical footprinting. Preformed complexes of IHF (1 μM) and top strand-labeled *attL* DNA (0.5 nM) were exposed to hydroxyl radicals generated by the Fenton reaction. The conditions were adjusted to create about one cleavage per DNA molecule. The reaction mixtures were separated by denaturing gel electrophoresis. (A) Autoradiography of six *attL* top strands in the absence and presence of IHF. Base positions are indicated in the right-hand margin. (B) Summary of protection patterns. Each solid bar represents bases that are moderately to strongly protected from hydroxyl-radical attack. (C) Alignment of the pertinent region of the H' locus of *attL* 1-5 with the IHF 3' consensus (2) for the top and bottom strands.

cause the recombination defect. Based on our footprinting analysis, we propose that the true failure of this site lies in the presence of an alternative IHF site out of register with the wild-type H', resulting in a nucleoprotein complex with the wrong geometry for recombination. To identify potential alternative and overlapping IHF sites, we used the MacTargsearch program, which was developed to identify possible IHF sites based on similarity to a set of 27 known IHF sites (7). It is particularly useful in identifying possible sites based on the conserved 3' region. It is much less effective in identifying sites based on the unconserved 5' region, a primary reason for this investigation. Of all the SELEX-derived sites, only *attL* 1-5 was distinguished as possessing a second possible site. One of these, which has a weaker similarity to bona fide IHF sites, is in register with the H' site, while the second, with a much stronger similarity to IHF sites, is out of register with the H' site. The misalignment of the IHF site with respect to the loci of Int-mediated catalysis is known to interfere strongly with lambda recombination (21).

While many reports, including the solution of the cocrystal structure of IHF bound to this very site (H'), have described the interactions of IHF with its DNA site, the specifics of this nucleoprotein association are still not fully defined. There appear to be no unique contacts between the protein and the DNA bases, although specific components may aid IHF in recognizing both regions of its binding site. Conventional protein-DNA binding specificity is thought to be driven by unique hydrogen bond donor and acceptors; in contrast, IHF appears to rely on indirect readout (34). Structural configurations associated with both the protein and target DNAs (e.g., ionic interactions, stacking interactions, etc.) limit the degrees of freedom of interactions which restrict and impose a high-affinity binding surface even when the protein-DNA base contacts are degenerate. The *E. coli* bacterial DNA binding proteins Fis and H-NS, which also modulate nucleoprotein functions, also share these characteristics. Fis binds and bends DNA site specifically and, like IHF, has a highly degenerate consensus sequence, making site identification difficult (1, 12). The DNA binding protein H-NS is even more problematic, since DNA sites are bound with high affinity but without any sequence conservation. H-NS has been shown to bind preferentially to existing DNA deformations (29, 37). Thus, the phenomenon of indirect readout coupled to DNA deformation is not unique to IHF.

Evidence of contacts or at least intimate proximity between IHF and the 5' region at various IHF sites has been demonstrated at virtually every base pair in the 5' domain from +35 to +19 (Fig. 1) (reviewed in references 19 and 25). Beyond +19, the number of demonstrable interactions falls off dramatically, although some investigators have shown biochemical evidence of interaction as distal as base +14 (31). In our selected sequences, we have found that strong conservation between positions +35 and +32 is associated with stability of the IHF-DNA complex during EMSA. This patch of four bases coincides with the 5' boundary of the 3' region observed in the great majority of natural IHF sites. By selecting IHF sites *de novo*, we demonstrate that this 4-base stretch is essential for forming at least one type of complex. Cocrystal analysis did not find unique hydrogen bonding of amino acid side chains to the canonical H' sequence in this region.

While it is beyond the scope of this paper to interpret the contribution of specific motifs, there appear to be patterns of conservation. Data from other studies demonstrate the importance and the proximity of IHF to the region from +19 to +29 (reviewed in reference 19). Indeed, the six consecutive adenines of H' are found from +19 to +24 and one of two kinks induced by IHF is found between bases +28 and +29. The A-tract figures prominently in the cocrystal structure of IHF bound to H' and demonstrates that, at least in this complex, IHF forms a clamp around a particularly narrow minor groove. One possibility is that all high-affinity members of class I have or are predisposed to form narrow minor grooves in the vicinity of +19 to +24 and/or sequences that readily kink about bases +28 and +29. While there is no obvious conservation among our class I *attL*s at coordinates +28 and +29, there is conservation at bases +18, +19, and +22. Conservation of these three bases persists over a 20-fold range of affinities and suggests that they specify important binding features that at least stabilize IHF-DNA interactions.

Evidence of a noncanonical structure associated with these sequences (+19 to +29) can be seen in our preliminary footprinting analysis. With the exception of the wild-type *attL* sequence, all of the *attL*s tested here displayed strong cleavage enhancements upon exposure to hydroxyl radicals in the absence of IHF. These enhancements result from increasing sol-

TABLE 2. Base composition of the 5' domain of selected IHF binding sites

H' DNA sequence	% A+T (no. in region/total no.) ^a in:	
	+35 to +14	+31 to +14
Class I	57 (340/594)	49 (236/486)
Class II	58 (114/198)	57 (93/162)
McClure consensus ^b	76 (454/594)	72 (352/486)
Wild-type H'	82 (18/22)	78 (14/18)

^a The percentage is the fraction of the sum of A · T base pairs in the designated region of each group of H' sequences divided by the total number of bases and multiplied by 100.

^b The McClure consensus (7) was derived from 27 natural sequences aligned with the native H' site.

vent accessibility of the hydroxyl radical target within the ribose ring through sequence-directed distortion. It is not clear whether these distortions play a role in enhancing IHF binding, although it is interesting that in *attL* 1-52 the enhancements coincide with the IHF-induced kink between +28 and +29 in the wild-type H' site. A more detailed biochemical analysis, focused on this region of class I *attL*s of similar IHF binding affinity, would clarify the nature and extent of these contacts.

In addition to base conservation (or lack thereof), we observed an intriguing phenomenon concerning the nucleotide content of the IHF-selected sequences. We compared the nucleotide content at bases +14 to +35 for both class I and class II with the corresponding region of the 27 natural IHF sites cited by Goodrich et al. (7), including the wild-type H' site. Remarkably, the base content differed dramatically between the selected sites and the native sequences (Table 2). These 22 bp in the native sequences are extremely AT rich, while they are only slightly enriched for A · T base pairs in the selected sequences regardless of class. Since natural sites and class I selected sequences conserve A · T base pairs from +35 to +32, this 4-base stretch weighs significantly in the overall base composition of the region. If we examine only bases +31 to +14, the result becomes even more striking: while 72% of the base pairs in this region are A · T among the native sequences, there is virtually no bias in base composition among our SELEX-derived sequences. We interpret this to mean that IHF sites, while typically possessing AT-rich 5' regions, are not dependent on this base composition. This implies that, on average, many motifs of various base compositions can accommodate high-affinity binding. The biological role of IHF may be a more significant determinant in target site selection and conservation than a specific binding motif that is independent of base composition. It is also possible that there are alternative binding modes of IHF, particularly within the 5' domain, as indicated in our preliminary footprint analysis.

We are currently attempting to identify critical determinants in the 3' region of the IHF site by using a SELEX approach. Despite its greater conservation, the 3' region is still mystifying, since IHF contacts appear degenerate in the cocrystal structure. Finally, to understand the essential determinants of IHF function, a similar selection will be used with recombination as the partition selector in order to ask how efficient recombination is associated with DNA affinity.

ACKNOWLEDGMENTS

We are particularly indebted to Richard Deonier for his many helpful suggestions and for comments on the manuscript. We also thank Bruce Teter and Howard Nash for comments on the manuscript.

This work was supported by Public Health Service grants GM55392

(to S.D.G.) and GM52847 (to A.M.S.) and by a grant from Pioneer Hi-Bred International Inc. (to S.D.G.).

REFERENCES

1. **Betermier, M., D. J. Galas, and M. Chandler.** 1994. Interactions of Fis protein with DNA: bending and specificity. *Biochimie* **76**:958–967.
2. **Craig, N. L., and H. A. Nash.** 1984. *E. coli* integration host factor binds to specific sites in DNA. *Cell* **39**:707–716.
3. **Cui, Y., Q. Wang, G. D. Stormo, and J. M. Calvo.** 1995. A consensus sequence for binding of LRP to DNA. *J. Bacteriol.* **177**:4872–4880.
4. **Gardner, J. F., and H. A. Nash.** 1986. Role of *Escherichia coli* IHF protein in lambda site-specific recombination: a mutational analysis of binding sites. *J. Mol. Biol.* **191**:181–189.
5. **Goodman, S. D., and H. A. Nash.** 1989. Functional replacement of a protein-induced bend in a DNA recombination site. *Nature* **341**:251–254.
6. **Goodman, S. D., S. C. Nicholson, and H. A. Nash.** 1992. Deformation of DNA during site-specific recombination of bacteriophage λ : replacement of IHF protein by HU protein or sequence-directed bends. *Proc. Natl. Acad. Sci. USA* **89**:11910–11914.
7. **Goodrich, J. A., M. L. Schwartz, and W. R. McClure.** 1990. Searching for and predicting the activity of sites for DNA binding proteins: compilation and analysis of the binding sites for *Escherichia coli* integration host factor (IHF). *Nucleic Acids Res.* **18**:4993–5000.
8. **Hagerman, P. J.** 1988. Flexibility of DNA. *Annu. Rev. Biophys. Biophys. Chem.* **17**:265–286.
9. **Hales, L. M., R. I. Gumpport, and J. F. Gardner.** 1994. Determining the DNA sequence elements required for binding integration host factor to two different target sites. *J. Bacteriol.* **176**:2999–3006.
10. **Hales, L. M., R. I. Gumpport, and J. F. Gardner.** 1994. Mutants of *Escherichia coli* integration host factor: DNA-binding and recombination properties. *Biochimie* **76**:1030–1040.
11. **Hales, L. M., R. I. Gumpport, and J. F. Gardner.** 1996. Examining the contribution of a dA+dT element to the conformation of *Escherichia coli* integration host factor-DNA complexes. *Nucleic Acids Res.* **24**:1780–1786.
12. **Hengen, P. N., S. L. Bartram, L. E. Stewart, and T. D. Schneider.** 1997. Information analysis of Fis binding sites. *Nucleic Acids Res.* **25**:4994–5002.
13. **Irvine, D., C. Tuerk, and L. Gold.** 1991. SELEXION, systematic evolution of ligands by exponential enrichment with integrated optimization by non-linear analysis. *J. Mol. Biol.* **222**:739–761.
14. **Kim, S., and A. Landy.** 1992. Lambda Int protein bridges between higher order complexes at two distant chromosomal loci *attL* and *attR*. *Science* **256**:198–203.
15. **Kim, S., L. Moitoso de Vargas, S. E. Nunes-Duby, and A. Landy.** 1990. Mapping of a higher order protein-DNA complex: two kinds of long-range interactions in λ *attL*. *Cell* **63**:773–781.
16. **Lee, E. C., M. P. MacWilliams, R. I. Gumpport, and J. F. Gardner.** 1991. Genetic analysis of *Escherichia coli* integration host factor interactions with its bacteriophage λ H' recognition site. *J. Bacteriol.* **173**:609–617.
17. **Lee, E. C., L. M. Hales, R. I. Gumpport, and J. F. Gardner.** 1992. The isolation and characterization of mutants of the integration host factor (IHF) of *Escherichia coli* with altered, expanded DNA-binding specificities. *EMBO J.* **11**:305–313.
18. **MacWilliams, M., R. I. Gumpport, and J. F. Gardner.** 1997. Mutational analysis of protein binding sites involved in formation of the bacteriophage λ *attL* complex. *J. Bacteriol.* **179**:1059–1067.
19. **Nash, H. A.** 1996. The HU and IHF proteins: accessory factors for complex protein-DNA assemblies, p. 149–179. *In* E. C. C. Lin and A. S. Lynch (ed.), *Regulation of gene expression in Escherichia coli*. R. G. Landes Co., Austin, Tex.
20. **Numrych, T. E., R. I. Gumpport, and J. F. Gardner.** 1990. A comparison of the effects of single-base and triple-base changes in the integrase arm-type binding sites on the site-specific recombination of bacteriophage lambda. *Nucleic Acids Res.* **18**:3953–3959.
21. **Nunes-Duby, S. E., L. I. Smith-Mungo, and A. Landy.** 1995. Single base-pair precision and structural rigidity in a small IHF-induced DNA loop. *J. Mol. Biol.* **253**:228–242.
22. **Oberto, J., K. Drlica, and J. Rouviere-Yaniv.** 1994. Histones, HMG, HU, IHF: *me me combat*. *Biochimie* **76**:901–908.
23. **Peacock, S., H. Weissbach, and H. A. Nash.** 1984. *In vitro* regulation of phage λ cII gene expression by *Escherichia coli* integration host factor. *Proc. Natl. Acad. Sci. USA* **81**:6009–6013.
24. **Rice, P. A., S. Yang, K. Mizuuchi, and H. A. Nash.** 1996. Crystal structure of an IHF-DNA complex: a protein-induced DNA U-turn. *Cell* **87**:1295–1306.
25. **Rice, P. A.** 1997. Making DNA do a U-turn: IHF and related proteins. *Curr. Opin. Struct. Biol.* **7**:86–93.
26. **Sambrook, J., E. F. Fritsch, and T. Maniatis.** 1989. *Molecular cloning: a laboratory manual*, 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
27. **Segall, A. M., S. D. Goodman, and H. A. Nash.** 1994. Architectural elements in nucleoprotein complexes: interchangeability of specific and non-specific DNA binding proteins. *EMBO J.* **13**:4536–4548.
28. **Shindo, H., F. Kanke, M. Miyake, U. Matsumoto, and M. Shimizu.** 1995. The binding specificity and affinity of *E. coli* integration host factor (IHF) are influenced by the flexibility of flanking regions of its recognition sites. *Biol. Pharm. Bull.* **18**:1328–1334.
29. **Spurio, R., M. Falconi, A. Brandi, C. L. Pon, and C. O. Gualerzi.** 1997. The oligomeric structure of nucleoid protein H-NS is necessary for recognition of intrinsically curved DNA and for DNA bending. *EMBO J.* **16**:1795–1805.
30. **Strauss, J. K., and L. J. Maher.** 1994. DNA bending by asymmetric phosphate neutralization. *Science* **266**:1829–1833.
31. **Sun, D., L. H. Hurley, and R. M. Harshey.** 1996. Structural distortions induced by integration host factor (IHF) at the H' site of phage λ probed by (+)-CC-1065, pluramycin, and KMnO₄ and by DNA cyclization studies. *Biochemistry* **35**:10815–10827.
32. **Tuerk, C., and L. Gold.** 1990. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249**:505–510.
33. **Vant-Hull, B., A. Payano-Baez, R. H. Davis, and L. Gold.** 1998. The mathematics of SELEX against complex targets. *J. Mol. Biol.* **278**:579–597.
34. **von Hippel, P. H.** 1994. Protein-DNA recognition: new perspectives and underlying themes. *Science* **263**:769–770.
35. **Werner, M. H., G. M. Clore, A. M. Gronenborn, and H. A. Nash.** 1994. Symmetry and asymmetry in the function of *Escherichia coli* integration host factor: implications for target identification by DNA-binding proteins. *Curr. Biol.* **4**:477–487.
36. **Williams, J. G. K., D. L. Wulff, and H. A. Nash.** 1977. A mutant of *Escherichia coli* deficient in a host function required for phage lambda integration and excision, p. 357–361. *In* A. Bukhari, J. Shapiro, and S. Adhya (ed.), *DNA insertion elements, plasmids and episomes*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
37. **Yamada, H., S. Muramatsu, and T. Mizuno.** 1990. An *Escherichia coli* protein that preferentially binds to sharply curved DNA. *J. Biochem.* **108**:420–425.
38. **Yang, C.-C., and H. A. Nash.** 1989. The interaction of *E. coli* IHF protein with its specific binding sites. *Cell* **57**:869–880.
39. **Yang, S.-W., and H. A. Nash.** 1994. Specific photocrosslinking of DNA-protein complexes: identification of contacts between integration host factor and its DNA target. *Proc. Natl. Acad. Sci. USA* **91**:12183–12187.
40. **Yang, S.-W., and H. A. Nash.** 1995. Comparison of protein binding to DNA *in vivo* and *in vitro*: defining an effective intracellular target. *EMBO J.* **14**:6292–6300.