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)

James A.Goodrich, Michael L.Schwartz+ and William R.McClure
Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Received May 10, 1990; Revised and Accepted July 24, 1990

ABSTRACT

An analysis of the sequence information contained in a compilation of published binding sites for *E. coli* integration host factor (IHF) was performed. The sequences of twenty-seven IHF sites were aligned; the base occurrences at each position, the information content, and an extended consensus sequence were obtained for the IHF site. The base occurrences at each position of the IHF site were used with a program written for the Apple Macintosh computers in order to determine the similarity scores for published IHF sites. A linear correlation was found to exist between the logarithm of IHF binding and functional data (relative free energies) and similarity scores for two groups of IHF sites. The MacTargsearch program and its potential usefulness in searching for other sites and predicting their relative activities is discussed.

INTRODUCTION

Recent increases in the number of known DNA binding proteins and their associated binding sites have led to numerous findings of potential binding sites based upon sequence similarity to consensus sequences for binding proteins. These potential sites are often defined by the fact that they have greater than a minimum number of matches with the consensus sequence of interest. A more complete means by which to search for potential binding sites uses a matrix containing base occurrences at each position of the site. The base occurrences arise from the compilation and alignment of all known binding sites for the protein of interest. Methods utilizing this scheme have proven useful in the past (1,2); however, the programs used to perform the analyses have either been specific for the binding site studied or difficult to use. In this study we present a program that can easily be used to search for and score potential sites for any binding protein for which a reasonable number of binding sites have been defined. The potential of this program is demonstrated in an analysis of the binding sites for the *Escherichia coli* integration host factor.

E. coli IHF is a heterodimeric protein with a molecular weight of about 20,000 Daltons. The subunits of the protein are expressed from the himA (3,4) and hip/himD genes (5). Hydroxyl radical and DNase I footprints of IHF bound to DNA extend over forty base pairs. As judged by gel retardation assays, IHF is known to bend DNA at a number of different sites (6–11). Recently, methylation interference and hydroxyl radical protection experiments have indicated that IHF contacts bases in the minor groove as the DNA is bent around the protein (12). As reviewed by Friedman (13), IHF has been implicated as playing a role, often regulatory, in a number of cellular processes, including site-specific recombination, phage packaging, transposition, plasmid replication, and transcription.

A number of consensus sequences for IHF binding have been published (11,14–18). Many researchers have used one or another of these consensus sequences to search for potential IHF sites. We have compiled and aligned the sequences of twenty-seven IHF binding sites. An extended consensus sequence was determined. Scores that represent the similarity of each IHF site to the matrix of base occurrences generated by the compilation were determined. Using data for the binding or function of two groups of IHF site mutants, we found a correlation between the logarithm of the site activity and the similarity scores of the sites.

THE PROGRAM

A program named MacTargsearch was written and compiled in Microsoft BASIC (b) for the Apple Macintosh. The program requires a Macintosh with at least 1 megabyte of memory and an Apple ImageWriter. The program is an adaptation of that written and used by Mulligan et al. (1,19) to search for and score *E. coli* promoters. MacTargsearch allows a user to search a sequence of bases for sites that have similarity to a target file that the user creates.

Target files contain the information about a site of interest, including the weighting matrix used to evaluate sites. The information is derived from a compilation of known sites such as that performed for *E. coli* promoters (20) and cyclic AMP receptor protein (2). Sites may have a single region of up to 48 bases or two regions of up to 24 bases each that are separated

+ Present address: Department of Pathology and Laboratory Medicine, Division of Neuropathology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA

by spacers of variable length. The user must enter the size of both regions, the base scores associated with the four bases at each position in the site, the number and sizes of the spacers (if any), and the score associated with each spacer. The base scores comprising the weighting matrix in the IHF target file used in this study were the occurrences of each of the four bases at every position in the 27 IHF sites compiled. These base scores are shown in Figure 2.

During a search, similarity scores (S.S.) are determined for all possible sites in both the forward and reverse directions in the sequence file being searched. Similarity scores are calculated according to the equation used by Mulligan et al. (1) to calculate homology scores for *E. coli* promoters.

$$\text{S.S.} = \left(\frac{\text{sum of base scores for site + spacer score} - \text{baseline score}}{\text{maximum score} - \text{baseline score}}\right) 100 \qquad 1.$$

We have used 'similarity' instead of 'homology' to describe this score, because the result does not bear directly on the presumed homologous evolutionary descent of the various sites evaluated. The baseline score is a number used to correct for the random occurrence of bases in a sequence, and is equal to 25% of the sum of all of the base scores and spacer scores in the target file. The maximum score is the sum of the maximum spacer score and the maximum base scores at each position of both regions of the target file.

Starting with the first base in the sequence file, the program sums the base scores for all positions in the first region as defined by the target file. If spacer regions are designated in the target file, the program then jumps the appropriate number of bases for the first spacer and sums the base scores for all positions in the second region. The spacer score is added to the base scores sum and the similarity score is calculated for the possible site. Similarity scores are calculated for all of the possible sites in a sequence file. The base sequence in the sequence file is treated as one strand of double stranded DNA; therefore, all bases in the sequence file are considered to be both the first base of a possible site in the forward direction and the complement of the last base of a possible site in the reverse direction. During the search, all sites found to have a score greater than a set minimum are sent to the printer.

MacTargsearch is rather flexible in that a base-scores-matrix for any site of up to 48 bases can be entered as a target file. It is possible to examine larger sites with multiple searches. This program is useful for searching for sites as well as for predicting the relative strengths of a group of known sites for performing a function that can be correlated with sequence similarity. Use of the program is not complicated. Requests for a copy of the MacTargsearch program may be sent to the authors along with a formatted 3.5 inch disk. An 'About' file included in the program folder explains how to use the program for any DNA site of interest.

RESULTS

Sequence Compilation

Sequences containing twenty-seven IHF sites have been analyzed. Alignment of the sequences was performed based upon the identical IHF site consensus sequences of Leong et al. (15) and Gamas et al. (18), and the footprint protection patterns caused by IHF binding to the sites. All of the 27 IHF sites compiled have exhibited binding in vitro and have either been footprinted or established by mutation. The IHF site does not appear to be symmetrical and one orientation for each site conformed to the consensus better than the other orientation.

Figure 1A shows the percent occurrence of the base found most often at each position when 125 bases around the twenty-seven IHF sites were aligned and analyzed. The base positions on the horizontal axis are numbered relative to the first highly conserved base pair in the consensus sequence, which is labelled +1. Bases 5′ of the most highly conserved base have negative base positions, while those 3′ of the most highly conserved base have positive base positions. Upper case letters above the bars in the histogram are used to represent strongly conserved bases, *i.e.* those bases with a percent occurrence greater than 63% (>4 standard deviations). Lower case letters are used to represent weakly conserved bases, *i.e.* those bases with a percent occurrence between 44% and 63% (2 to 4 standard deviations). The pattern of base occurrences is also shown in panel B as a histogram in which the information content at each position was calculated according to (21,22):

$$I_{seq} = \sum_{b=A}^{T} f_b \log_2 \frac{f_b}{p_b} \qquad 2.$$

where f_b = the frequency of base b at a given position among compiled sites and p_b = the frequency of base b in the *E. coli* genome (0.25 was used for all four bases). The representations in panel A and panel B are qualitatively very similar. Both show that the most conserved sequences (or highest information content) occur in the region from +1 to +13 and that additional sequence conservation extends in both directions.

The histograms in Figure 1 allowed us to define a region that we considered as an IHF site for subsequent work. The region between the two vertical lines in the figure is the 48 base pairs that are considered to represent an IHF site. The boundaries are −22 and +26 in the sequence. Some similarity between sites can be seen downstream of +26. We decided not to include this region in the IHF site because of the lack of evidence from DNase I footprinting, hydroxyl radical protection data, and mutational analysis that this region is part of the IHF binding site.

Figure 2 contains a compilation of the 48 bases that represent one strand of each of the twenty-seven IHF sites considered in this study. The hierarchy of base occurrences at each position is shown toward the bottom of the figure. Based on this hierarchy an extended consensus site sequence was derived. In the consensus sequence, upper case letters represent strongly conserved bases and lower case letters represent weakly conserved bases. The bottom line of letters represents bases that occur with high frequency that are not the most frequently occurring base at a given position. These bases occur with a frequency that is greater than 4 standard deviations from the expected mean occurrence of the remaining 3 bases (not including the most frequently occurring base at that position).

Similarity Scores of Compiled Sites

Similarity scores were determined for all of the IHF sites compiled in Figure 2. The IHF target file consisted of a single region of 48 bases with the base scores listed near the bottom of Figure 2. The maximum site score and baseline score contained in the IHF target file were 682 and 323, respectively. The IHF sites included in Figure 2 are listed in order of decreasing similarity score. The similarity scores of the 27 sites compiled

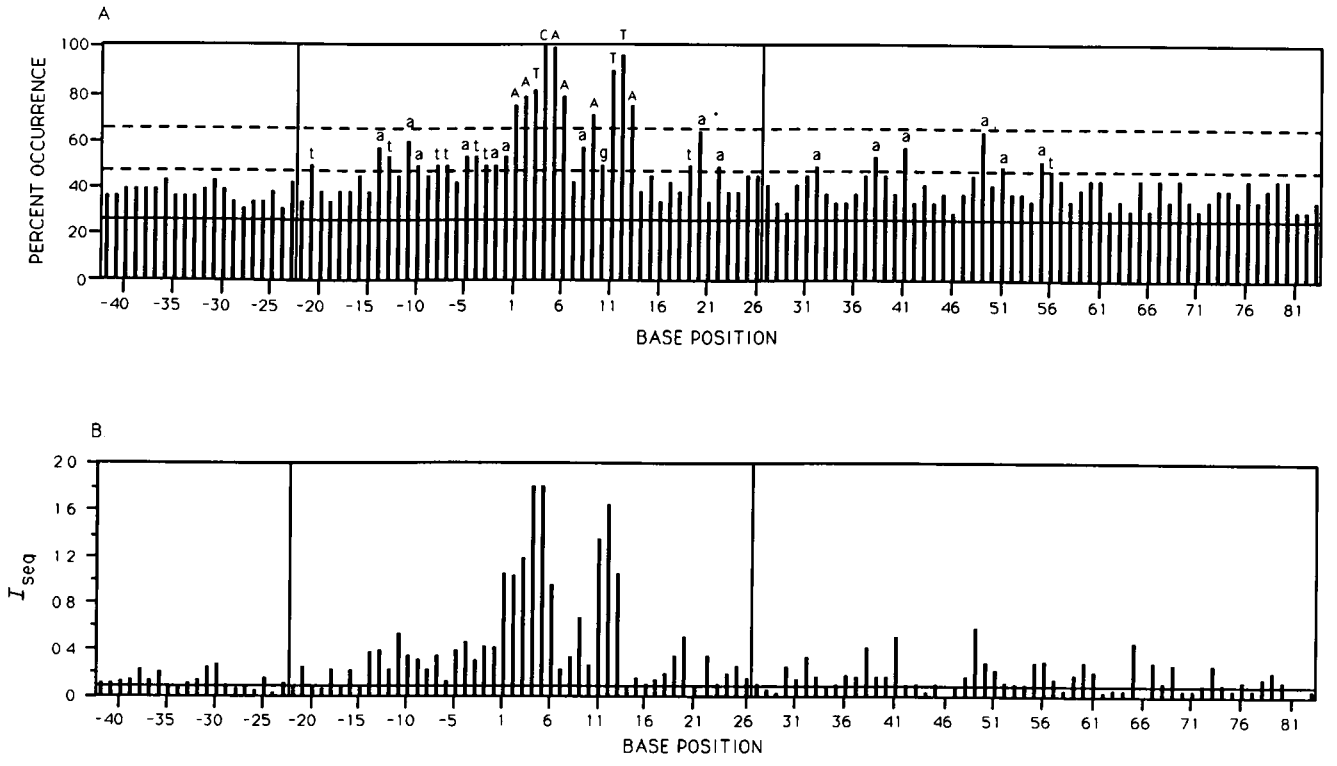
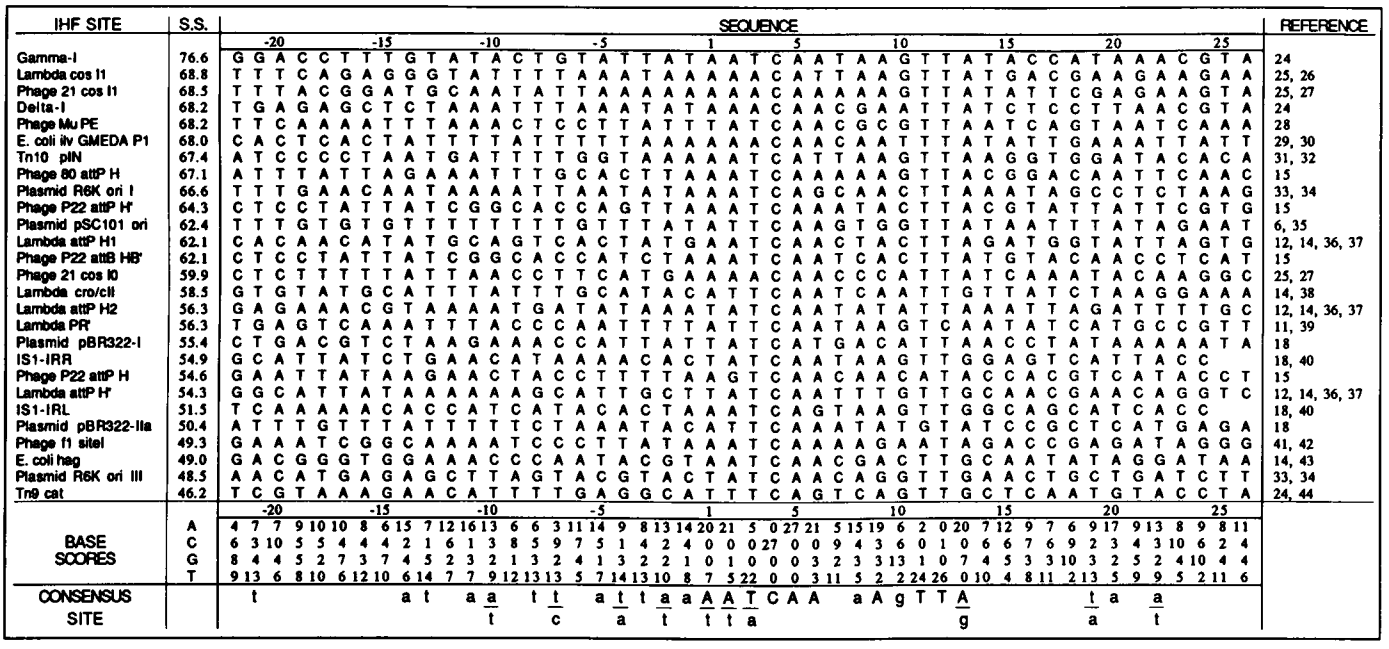


Figure 1. (A) Percent occurrence of the base found most frequently at each position in a 125 base region among 27 aligned IHF sites. References pertaining to the IHF sites used are given in Figure 2. The alignment of sites and the numbering of base positions are discussed in the text. The percent occurrence at each position was calculated as the number of occurrences of the most highly conserved base at a given position divided by the number of bases compiled at the position (this quotient was then multiplied by 100). The two horizontal dashed lines across the histogram indicate the percent occurrence at 2 and 4 standard deviations (Poisson) above the theoretical mean of 25% (shown as a solid horizontal line). The standard deviation was calculated as the square root of the mean number of occurrences; one standard deviation was equal to 9.6 percent occurrence. The two vertical lines represent the boundaries of the region that is considered as an IHF site. The letters above columns in the histogram are the bases that are most conserved at the position of the column. (B) The information content (I_{seq} in bits) at each position in a 125 base region among 27 aligned IHF sites. I_{seq} is calculated for each position using equation 2 in the text. At positions for which $f_b = 0$, an estimated occurrence of 0.5 was used in the calculation (23). This causes the maximum bits of information at a given position to be decreased from 2 to 1.79 for a group of 27 sites. The horizontal line represents the corrected base line taking into account small-sample-error (21). This correction would result in 0.09 bits being subtracted from each I_{seq} in the histogram. The IHF site (the 48 base region between the vertical lines) contains a total of 21.4 bits before the small sample correction, and 17.0 bits after correction. The 125 base region contains a total of 20.5 bits of information after the small sample correction is made.

range from 46 to 77. The quantitative evaluation of these sites suggests a rational basis for defining a consensus sequence for an IHF site. Typically, a DNA binding site is described with a consensus sequence to provide a concise indication of its location and identity in a region of interest. The difficulty in doing so for IHF binding sites is that conserved sequence information is distributed over (at least) 48 base pairs. The inclusion of all of these large and small contributions is very unwieldy. We suggest that the 10 base sequence (represented as upper case letters in Figure 2) and the similarity score be used to describe an IHF site. The following reasoning has led us to suggest this combined approach for indicating an IHF site. First, the average similarity score for the known IHF sites is 60; the minimum number of base pairs required to obtain a similarity score of 60 is ten. In fact, these are the 10 bases designated with upper case letters in Figure 2. These 10 bases also contain 58% of the information content in the (48bp) IHF site. Second, the overall similarity score is also required to describe the site accurately because expressions such as '8 out of 10 consensus base pairs' distorts the real similarity to the complete IHF site. For example, of the 13 sites in Figure 2 that are above average in similarity score, only five have 10 out of 10 matches to the upper case letters in the consensus sequence; moreover, the similarity scores of these five sites range from 62 to 77. It may turn out that when more IHF sites are compiled and more binding measurements are reported that the overall scores and the minimum number of base pairs required to define an average binding site score will change. In the meantime, we suggest that the combined use of the 10 base sequence (to indicate the site location) and the similarity score (to indicate similarity to the entire site) is the most concise description of an IHF binding site.

The usefulness of search programs such as MacTargsearch arises from their ability to allow a user to distinguish between actual sites and sequences that appear to be sites but are not. Such sites can be called false positives; they arise with a predictable frequency from any search, whether it be a search performed by eye or by a computer program. The frequency with which MacTargsearch finds false positives was assessed in two ways. First, the genomes of phages T7 and lambda were searched and scores that were greater than or equal to the lowest score for an actual site (46.2) were considered as false positives. Second, similarity scores were determined for sites that had been published as possible binding sites, but were later found to bind IHF only weakly. The similarity scores for these sites were compared to the scores for actual IHF sites.

There is no known contribution of IHF to phage T7 infection.



IHF SITE	S.S.	SEQUENCE	REFERENCE
		-20 -15 -10 -5 1 5 10 15 20 25	
Gamma-I	76.6	G G A C C T T T G T A T A C T G T A T T A T A A T C A A T A A G T T A T A C C A T A A A C G T A	24
Lambda cos I1	68.8	T T T C A G A G G G T A T T T T A A A T A A A A A C A T T A A G T T A T G A C G A A G A A G A A	25, 26
Phage 21 cos I1	68.5	T T T A C G G A T G C A A A T A T T A A A A A A A A A C A A A A A A G T T A T A T T C G A G A A G T A	25, 27
Delta-I	68.2	T G A G A G C T C T A A A T T T A A A T A T A A A A C A A C G A A T T A T C T C C T T A A C G T A	24
Phage Mu PE	68.2	T T C A A A A T T T A A A C T C C T T A T T T A T C A A C G G C G T T A A T C A G T A A T C A A A	28
E. coli ilv GMEDA P1	68.0	C A C T C A C T A T T T T A T T T T A A A A A A C A A C A A T T T A T A T T G A A A T T A T T	29, 30
Tn10 pIN	67.4	A T C C C C T A A T G A T T T T G G T A A A A A A T C A T T A A G T T A A G G T G G A T A C A C A	31, 32
Phage 80 attP H	67.1	A T T T A T T A G A A A T T T T G C A C T T A A A T C A A A A A G T T A C G G A C A A T T C A A C	15
Plasmid R6K ori I	66.6	T T T G A A C A A T A A A A T T T A A T T A A A A T C A G C A A C T T A A A T A G C C C T C T A A G	33, 34
Phage P22 attP H'	64.3	C T C C T A T T A T C G G C A C C A G T T A A A T C A A A T A C T T A C G T A T T A T T C G T G	15
Plasmid pSC101 ori	62.4	T T T G T G T G T T T T T T T T G T T T A T A T T C A A A G T G G T T A T A A T T T A T A G A A T	6, 35
Lambda attP H1	62.1	C A C A A C A T A T G C A G T C A C T A T G A A T C A A C T A C T T A G A T G G T A T T A G T G	12, 14, 36, 37
Phage P22 attB H8'	62.1	C T C C T A T T A T C G G C A C C A T C T A A A T C A A T C A C T T A T G T A C A A C C T C A T	15
Phage 21 cos I0	59.9	C T C T T T T T A T T A A C C T T C A T G A A A A C A A C C C A T T A T C A A A T A C A A G G C	25, 27
Lambda cro/cII	58.5	G T G T A T G C A T T T A T T T G C A T A C A T T C A A T C A A T T G T T A T C T A A G G A A A	14, 38
Lambda attP H2	56.3	G A G A A A C G T A A A A T T G A T A T A A A T A T T C A A A T A T A T T T A A A T T A G A T T T T G C	12, 14, 36, 37
Lambda PR'	56.3	T G A G T C A A A T T T A C C C A A T T T T A T T T C A A T A A G T C A A T A T C A T G C C G T T	11, 39
Plasmid pBR322-I	55.4	C T G A C G T C T A A G A A A C C C A T T A T T A T T A T C A T G A C A T T T A A C C C T A T A A A A A A T A	18
IS1-IRR	54.9	G C A T T A T C T G A A C A T A A A A C A C T A T T C A A T T A A G T T G G A A G T C A T T A C C	18, 40
Phage P22 attP H	54.6	G A A T T A T A A G A A C T A C C C T T T T A A A G T C A A C A A C A A T A C C A C G T C A T A C C C T	15
Lambda attP H'	54.3	G G C A T T A T A A A A A A A G C A T T G C T T A T C A A A T T T G T T G C A A A C G A A A C A A G G T C	12, 14, 36, 37
IS1-IRL	51.5	T C A A A A A C A C C A T C A T A C A C T A A A A T C A G T A A G T T G G C A A G C A A T C A C C	18, 40
Plasmid pBR322-IIa	50.4	A T T T G T T T A T T T T T C T A A A T A C A T T C A A A A T A T G T A T C C G C T C A T G A G A A	18
Phage f1 site I	49.3	G A A A T C G G C A A A A A T C C C C T T A T A A A T C A A A A A G A A T A G A A C C G A A G A T A G G G	41, 42
E. coli hag	49.0	G A C G G G T G G A A A A C C C A A T A C G T A A T C A A C G A A C T T G C A A A T A T A G G A T A A	14, 43
Plasmid R6K ori III	48.5	A A C A T G A G A G A G C T T A G T A C G G T A C T T A T C A A C A G G T T G A A C T G C T G A T C T T	33, 34
Tn9 cat	46.2	T C G T A A A G A A C A T T T T G A G G C A T T T C A G T C A G T T G C T C A A T G T A C C T T A	24, 44
		-20 -15 -10 -5 1 5 10 15 20 25	
BASE SCORES A		4 7 7 9 10 10 8 6 15 7 12 16 13 6 6 3 11 14 9 8 13 14 20 21 5 0 27 21 5 15 19 6 2 0 20 7 12 9 7 6 9 17 9 13 8 9 8 11	
BASE SCORES C		6 3 10 5 5 4 4 4 2 1 6 1 3 8 5 9 7 5 1 4 2 4 0 0 0 27 0 0 9 4 3 6 0 1 0 6 6 7 6 9 2 3 4 3 10 6 2 4	
BASE SCORES G		8 4 4 5 2 7 3 7 4 5 2 3 2 1 3 2 4 1 3 2 2 1 0 1 0 0 0 3 2 3 3 13 1 0 7 4 5 3 3 10 3 2 5 2 4 10 4 4	
BASE SCORES T		9 13 6 8 10 6 12 10 6 14 7 7 9 12 13 13 5 7 14 13 10 8 7 5 22 0 0 3 11 5 2 2 24 26 0 10 4 8 11 2 13 5 9 9 5 2 11 6	
CONSENSUS		t a t a a t t t a a A A T C A A a A g T T A t a a	
SITE		t c a t t t a g t a	

Figure 2. Compilation and analysis of the sequences of 27 IHF sites. 48 bases for each of 27 IHF sites are aligned with base positions as described in the text. The bases at positions 25 and 26 of the IS1-IRR and IS1-IRL sites are absent due to the fact that the C residues at position 24 of these sites are the end bases of IS1. The base scores are determined by summing the number of occurrences of each base at each position in the IHF site. The base scores are used during the formation of the IHF target file, along with the maximum score (682) and the baseline score (323). The IHF consensus sequence is shown at the bottom of the figure. An explanation of the consensus sequence and the limits used during its determination is given in the text. The entries under the column labelled 'S.S.' are the similarity scores for the IHF sites.

We tested the infectivity of wild type T7 in plaque assays on *E. coli*. We found that plating T7 on strain NK7817 (W3110 str^R lacZΔrl Δ3[hip]::cat Δ82[himA]::Tn10) resulted in a 30% reduction in the number of plaque forming units as compared to that obtained with strain W3110. Plaques formed on the two strains were of equal size and morphology. The similarity scores were obtained for all of the possible sites in the 39936 base pairs of phage T7 using the IHF target file. The scores range from −35 to 48.7. The mean similarity score is 1.2 with a standard deviation of 10.7. Seven possible sites were found with scores greater than or equal to 46.2. The overlap region between the similarity scores for known sites and those found for the 79872 possible sites in the phage T7 genome spans only 2.5 similarity score units. If any of these seven sites actually bind IHF, they are predicted to have binding constants only about 20% that of an average known IHF site with similarity score equal to 60 (See discussion below of Figure 5 correlation.). Alternatively, if none of these predicted sites actually bind IHF, then the frequency of false positives obtained in the search of the phage T7 genome was 1.75×10^{-4} per base pair. In either case, we expect by chance to find about one very low scoring IHF site in every 5,000 base pairs of DNA sequence, or one average scoring IHF site in 50,000 base pairs of DNA sequence.

As a comparison to the phage T7 results, we searched the genome of phage lambda, a phage known to require IHF for multiple functions in vivo. The mean similarity score for all possible sites in the lambda genome was 0.1 with a standard deviation of 11.5. Twenty possible sites were found to have similarity scores greater than or equal to 46.2. Six of these sites are actual IHF binding sites that are included in Figure 2. Another site near the Pbl promoter has been found to be an IHF site (11,39), but it was not included in Figure 2 because of the lack of published footprints. The 13 remaining sites with scores of 46.2 or greater must be considered false positives at this point because of the lack of information about IHF binding to these regions of phage lambda. These 13 sites in the phage lambda genome represent a frequency of 2.68×10^{-4} false positives per base pair.

Figure 3 contains a list of sequences that have either been proposed as IHF sites and later found to be very weak sites at best or sites that have been found to bind IHF but were not included in Figure 2 for a specific reason. Similarity scores have been determined for all of these sites and they are included in Figure 3. The pBR322-III, Lambda *cos* I0, Lambda *cos* I0', Lambda *cos* I2, Lambda *cos* I3, Lambda *cos* I4, Phage 21 *cos* I2, Delta II, and IS1-int sites were all proposed to be IHF sites and later found to bind IHF very weakly, if at all (refer to references in Figure 3 for each of these sites). Most of these sites conform rather well to at least one of the previously published IHF consensus sequences (11,14−18); however, the similarity scores of the sites in this group range from 12.5 to 37.0, well below the lowest score for the group of binding sites included in Figure 2. The Delta II sequence, with a similarity score of 34.0, is interesting in that it matches 4 of the 6 previously published consensus sequences exactly (15−18); however, it binds IHF very weakly in vitro (24). The analysis of these sites as well as the small number of false positives found in a search of the phage T7 genome demonstrate the usefulness of MacTargsearch and the IHF target file for finding or excluding potential IHF sites.

The remainder of the sites in Figure 3 merit discussion. The two sites named Plasmid pBR322-IIb (18) and Plasmid R6K *ori*

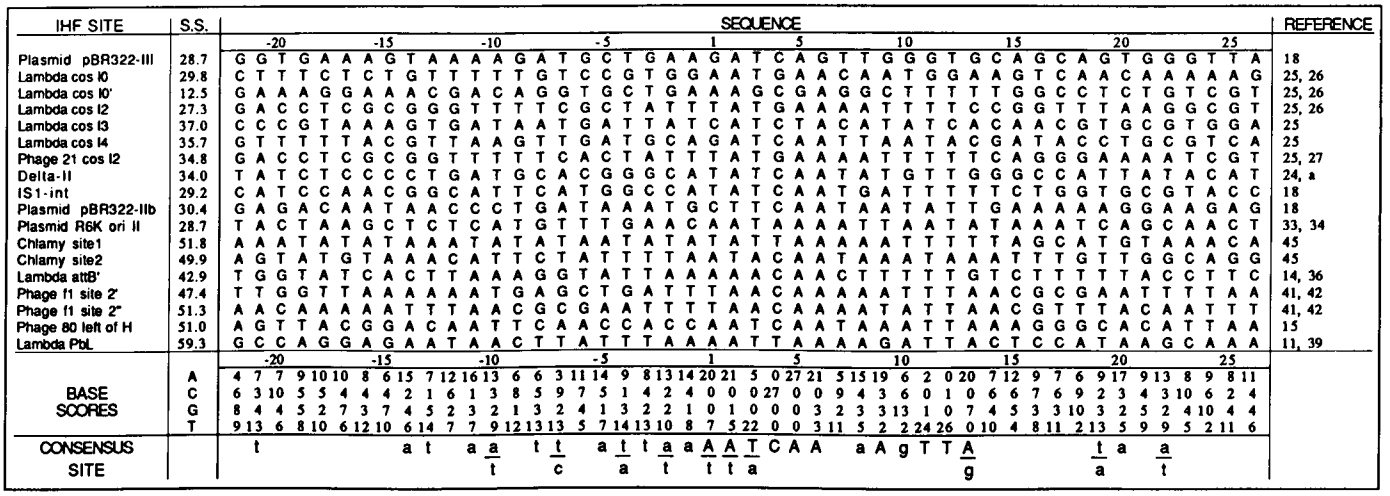


Figure 3. Sequences and similarity scores of proposed IHF sites. The sites contained in this figure are a diverse group of sites from the literature that were not included in the compilations in Figures 1 and 2. All of these sites are discussed in the text. The base scores matrix from the compilation of the 27 sites shown in Figure 2 and the extended IHF consensus sequence are provided for comparison.

[a] N. Grindley, personal communication

II (33) were proposed as IHF binding sites in regions found to give large footprints in the presence of IHF. In each of these cases IHF was proposed to bind to two sites giving a single large footprint. Only the sites with the higher similarity scores were included in Figure 2 (pBR322-IIa and R6K *ori* III) because of the lack of evidence that two molecules of IHF are binding to produce either of these two footprints. Chlamy site1 and site2 are two sites on chloroplast DNA from *Chlamydomonas* that lie within regions that give in vitro footprints in the presence of IHF (45). The Lambda *attB'* sequence, with a similarity score of 42.9, was alluded to as a possible IHF binding site by Craig and Nash (14). The sites named Phage f1 site 2' and site 2'' were identified by DNase I footprinting as weak sites, although their similarity scores are similar to that of f1 site 1 (41). Footprinting over the site Phage 80 left of H was found only at high IHF concentrations, and it was postulated that binding of IHF to this site may be inhibited by binding to the Phage 80 *attP* H site (15). Finally, the Lambda Pbl site has been found to regulate transcription from the lambda Pbl promoter in vitro (39). Kur et al. have stated that IHF binding can be footprinted at the Pbl site (11). During our search of the lambda genome, we found this site to have a high similarity score of 59.3. We chose not to include this site in Figure 2 because of the lack of published footprint or mutation data for IHF binding to the site.

Several putative IHF sites (uncharacterized sites that have been identified by similarity to a previously published consensus sequence) have been reported in the literature (46–55). These sites are located in bacteriophage genomes, plasmids, insertion elements, and regions of the *E. coli* genome. Analysis of these putative sites with MacTargsearch produced a wide range of similarity scores, one third of which were below the lowest similarity score obtained for a known IHF site (results not shown). Experimental analysis of these, and other, putative IHF binding sites will allow the weighting matrix contained in the IHF target file to be updated.

Correlation Between IHF Site Activity and Similarity Score

A number of studies looking at the binding of IHF to DNA have considered multiple sites and have reported the relative binding affinities of each of these sites for IHF. These semiquantitative comparisons can be used to examine the correlation between binding affinities and similarity scores for some sites. The pBR322-I and pBR322-II sites have a similar affinity for IHF, and both have a greater affinity for IHF than does the pBR322-III site (18); their similarity scores are consistent with this hierarchy. The Phage P22 *attP* H′ and *attB* HB′ sites were found to bind IHF more tightly than the Phage P22 *attP* H site (15); their similarity scores agree with this analysis. Plasmid R6K *ori* I binds IHF better than pR6K *ori* III, as predicted by the similarity scores for these two sites (33). The binding hierarchy for the Gamma-delta sites in decreasing order is Gamma-I, Delta-I, and Delta-II (24). Their similarity scores also decrease in this order.

The similarity scores of two sites are not consistent with the binding hierarchies. The Tn9 cat site binds IHF with an affinity comparable to the Delta-I site (24), but its similarity score is much lower. The binding order for the Lambda *attP* sites is H2 > H′ > H1 (14), while the similarity scores are 56.3, 54.3, and 62.1 respectively. We are unable to explain the lack of correlation between binding and similarity scores for the Tn9 cat and Lambda *attP* H1 sites. It is possible that for some sites, in vitro binding to DNA fragments will not directly correlate with the sequence information that has evolved to provide binding and function for IHF in vivo. It also seems likely that DNA conformation at an IHF site may contribute to the binding constant of IHF, given that the protein bends DNA (See below).

A number of mutations have been created and characterized at IHF sites. With the exception of a series of mutations in the Tn10 pIN site (31), the mutations are predominantly block substitutions that decrease IHF binding. Figure 4 contains the sequences of the wild type and mutated IHF sites, along with the effects of the mutations on IHF binding or function, and the similarity scores of the sites. The shaded areas of the mutant sequences represent the bases that have been substituted into the natural IHF sites. For all but three of the mutations, changes in similarity score in going from the wild type to the mutant sequence correlate with the effect of the mutation on IHF binding or function. Two of the discrepancies result from point mutations at a single position in the Tn10 pIN site (24A and 24T). The

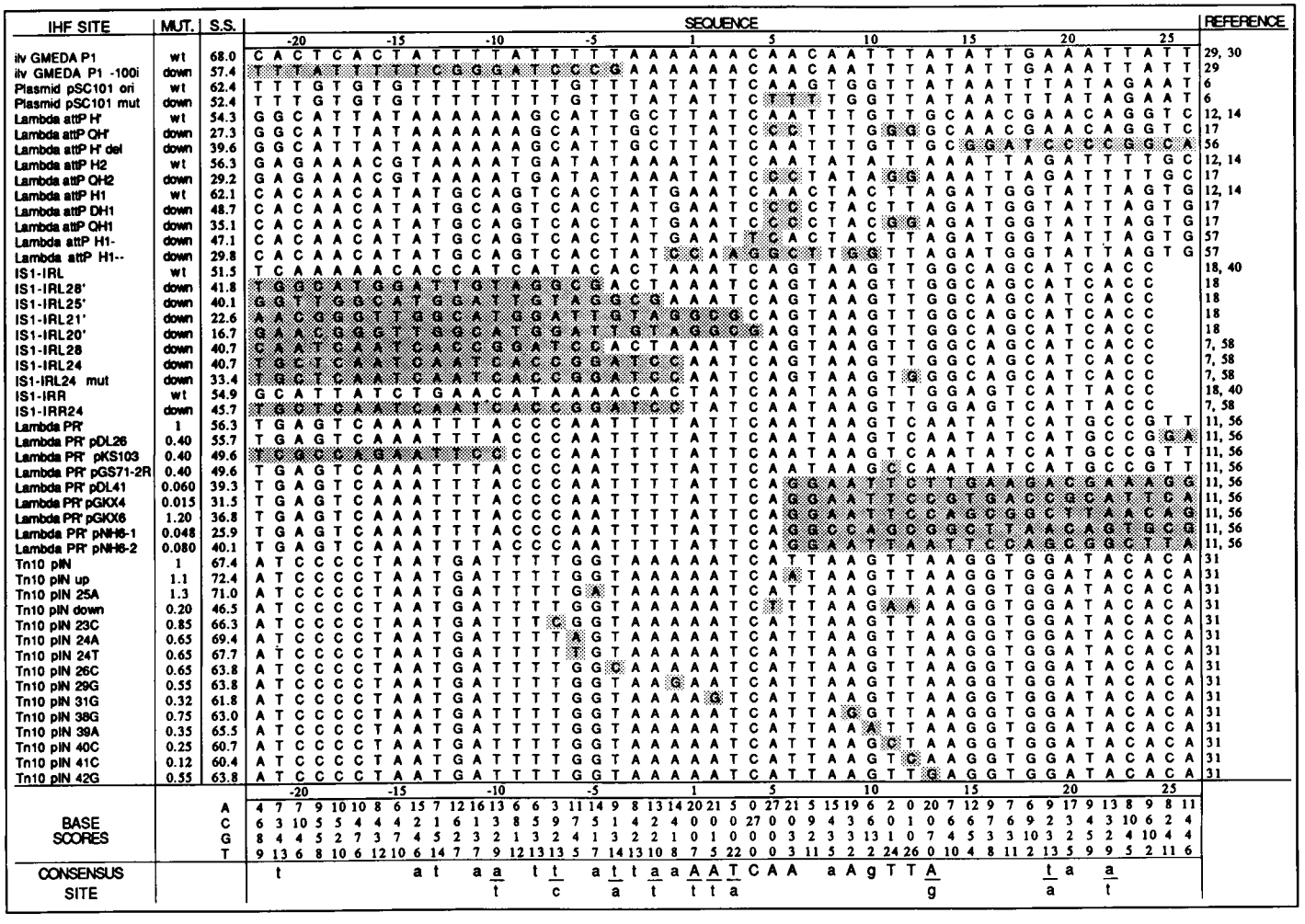


Figure 4. Sequences of mutant IHF sites. Shaded bases are those that have been substituted into the wild type IHF sites. The wild type IHF sites are also provided for comparison to the mutant sites. The effects of the mutations on in vitro IHF binding or in vivo function (for the Tn10 pIN sites) relative to the wild type sites are given in the column labelled MUT. The numbers in this column represent the measured IHF site activities relative to the wild type site for two groups of mutant IHF sites. Relative in vitro binding affinities were measured for a group of lambda PR' IHF site mutants (56), and relative in vivo transposition frequencies were measured for a group of Tn10 mini-transposons containing either the wild type or a mutant Tn10 pIN IHF site (31). The quantitative measurements for the wild type and mutant PR' and Tn10 pIN IHF sites are used to derive the points plotted in Figure 5, as described in the text. The base scores matrix from the compilation of the 27 sites shown in Figure 2 and the extended IHF consensus sequence are provided for comparison.

similarity scores for these two sites are not significantly different from the similarity score for the wild type Tn10 pIN site. The third discrepancy occurs for a block substitution at the Lambda PR' site. Lambda PR' pGKX6 is a deletion/substitution mutant that Kur et al. have created; they found it to be slightly up for IHF binding in vitro (11,56). The similarity score of the PR' IHF site changes from 56.3 to 36.8 as a result of the mutation. We are unable to explain this strong discrepancy. It will be interesting to see if other sites like PR' pGKX6 can be created, and more important, to determine whether point mutations are able to decrease the binding of IHF to PR' pGKX6.

Quantitative Correlations

Quantitative data from a series of binding site mutants can potentially be analyzed for their correlation with the similarity scores produced by MacTargsearch. We have performed such an analysis, assuming that the sequence information contained in a binding site can be correlated with the free energy associated with the binding of a protein to that site. Kur et al. (56) have quantitated the in vitro IHF binding affinity of 8 mutant PR' IHF sites relative to the wild type PR' IHF site. In Figure 5, we have plotted the logarithm of the in vitro relative binding affinities of IHF to the PR' wild type and mutant IHF sites versus the similarity scores of the PR' sites.

The best linear fit through the data as determined by least squares is shown as a solid line in Figure 5. The slope of the solid line in Figure 5 is 0.051 with a standard deviation of 0.009. The PR' pGKX6 mutant that acts as an up mutant but gives a very low similarity score was omitted from the least squares analysis (triangle). The correlation coefficient (0.91) and the standard deviation of the slope both indicate that the logarithm of the relative binding affinities correlate well with the similarity scores for the PR' IHF sites.

A second set of quantitative data has been obtained by Huisman et al. (31). A functional assay was used to quantitate the frequency of in vivo transposition by mini-transposons containing either the wild type or one of the 14 mutant Tn10 pIN IHF sites (31). One set of transposition data obtained by Huisman et al. (31) used constructs that contained one wild type inside end and one outside end with either the wild type or a mutant IHF site. Although

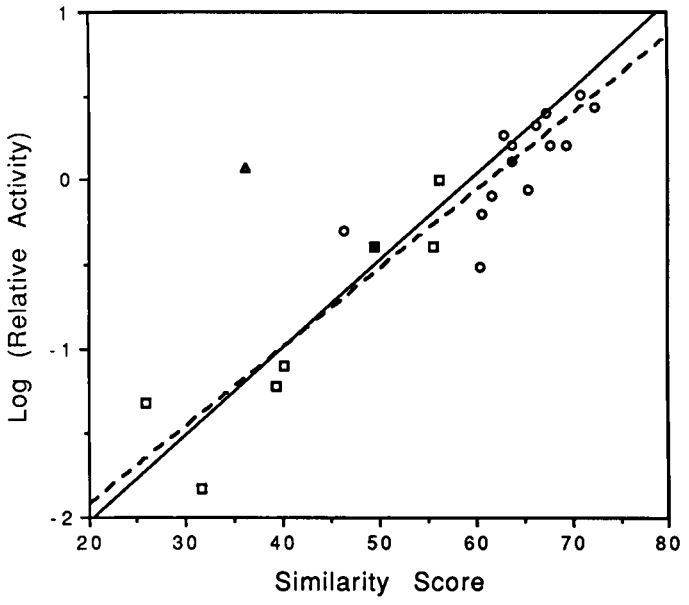


Figure 5. Correlation between relative activities and similarity scores for two groups of IHF sites. The relative binding affinities (relative to the wild type PR′ IHF site) of IHF to the PR′ IHF site and 8 mutants were obtained from Kur et al. (56) and are provided in Figure 4. The relative affinities are plotted logarithmically versus the similarity scores as square symbols (the single closed square represents data for two mutants). The solid line represents the best linear least squares fit through the PR′ IHF site data (slope = 0.051 ±0.009, Y intercept = −3.0, correlation coefficient = 0.91). The triangle depicts the data for the PR′ pGKX6 mutant. The data for this mutant is provided for reference, but was not included in the least squares analysis as discussed in the text. Relative in vivo transposition frequencies of mini transposons containing the wild type and 13 mutant Tn10 pIN IHF sites were published by Huisman et al. (31) and are provided in Figure 4. The relative transposition frequencies (relative to the wild type Tn10 pIN IHF site transposition frequency) were multiplied by a factor of 2.5, in order to obtain corrected Tn10 IHF site data that could be correlated with the line generated from the PR′ IHF sites. The corrected Tn10 IHF site transposition frequencies were plotted logarithmically versus the similarity scores as circles (the single closed circle represents data for two mutants). The dashed line represents the best linear least squares fit through the PR′ and Tn10 IHF site data (slope = 0.045 ±0.004, Y intercept = −2.8, correlation coefficient = 0.94). The two sets of data were also plotted using the procedures of Berg and von Hippel (2) and as suggested by Stormo (22). The results were comparable in both cases to that shown above, *i.e.* correlation coefficients were 0.92 and the slope values were uncertain to ± 10% (data not shown). Both of these alternative correlations required the omission of the PR′ pGKX6 mutant datum.

this is a functional assay, not an in vitro binding assay, we reasoned that measurements of transposition frequency might be proportional to the binding affinities for the wild type and mutant Tn10 pIN IHF sites. If this were so, then the Tn10 pIN IHF site data should fall on the line generated from the PR′ IHF site data in Figure 5. In order to perform such an analysis the transposition frequency data had to be in a form that allowed it to be plotted relative to the logarithm of the binding affinity of the wild type PR′ IHF site. Information about the IHF binding affinity of the wild type Tn10 pIN IHF site relative to that of the wild type PR′ IHF site is not available. We decided to enter the similarity score of the wild type Tn10 pIN IHF site into the equation of the line through the PR′ IHF site data in order to determine the factor by which to multiply the relative transposition frequency data for the Tn10 pIN IHF sites. The relative transposition frequencies (relative to the wild type Tn10 pIN IHF site transposition frequency) were multiplied by a factor of 2.5, and the logarithm of these corrected data were plotted versus the similarity scores in Figure 5.

The dashed line in Figure 5 represents the least squares fit through all of the Tn10 pIN and PR′ points (except the point representing the data for the PR′ pGKX6 mutant). The slope of this line is 0.045 with a standard deviation of 0.004. In practical terms, this slope corresponds to a free energy change of −0.61 kcal/mol for every increase of 10 similarity score units, or a change of 22 similarity score units for every unit change in the logarithm of the IHF binding constant. The addition of the Tn10 data to the PR′ data decreases the slope of the line by 10% while reducing the error associated with the slope by half. The correlation coefficient increases to 0.94 with the addition of the Tn10 data. It is pleasing to see that a strong correlation exists between two sets of data obtained by different methods and the similarity scores determined using MacTargsearch and the IHF target file. Realizing that the consistency present in this analysis could be coincidental, it would be interesting to obtain in vitro binding data for the set of Tn10 pIN IHF site mutants, as well as for other IHF binding sites, and to compare that data to the lines in Figure 5. In addition, the predictive value (searching) and refinement of the evaluation protocol demonstrated here will be much improved when more systematic IHF binding results are available.

DNA Conformation and IHF Function

The analysis of IHF sites presented above is based upon the assumption that the relative activity of a DNA site is a summation of the individual contributions of each base in the site to the overall function of the site. It is possible, if not probable, that DNA conformation and flexibility may play a significant role in the function of many protein binding sites. For example, IHF may bind more tightly to sites that are inherently oriented to bend around the IHF molecule, as opposed to those oriented to bend away from the IHF molecule. Studies addressing the contributions of sequence to DNA conformation and flexibility argue that groups of base pairs, not individual base pairs, are very important. We have analyzed the sequences of the 27 IHF sites for dinucleotide periodicities and occurrences. Dinucleotide periodicities were analyzed through the use of an autocorrelation function similar to that used by Trifonov and Sussman to analyze chromatin DNA nucleotide sequences (59). We have not found any strongly conserved dinucleotide periodicities among the IHF sites. We further analyzed the occurrence of all 16 dinucleotides in the 125 base region represented in the base-scores-matrix used to derive Figure 1. In the 48 base region we have considered to represent an IHF site, the dinucleotide occurrences were as expected based upon single base frequencies. The only preferences found were for AA and TT dinucleotides between positions −13 and −10 and for AT dinucleotides between positions +13 and +19. Higher than expected occurrences of AA and TT dinucleotides were found in the regions upstream and downstream of the 48 base IHF site. The significance of specific dinucleotides at a given site needs to be addressed by mutation before strong arguments can be made for dinucleotide preferences at positions in and around an IHF site.

We have attempted to estimate the contribution of DNA conformation and flexibility to the binding of IHF. The free energy associated with bending DNA around an arc of a circle can be calculated using the formulation of Gray and Hearst (60),

$$\Delta G = \frac{(kTP)}{2L}\left(\frac{(2\pi)\Delta\theta}{360}\right)^2 \text{(kcal/mol)} \qquad 3.$$

where P = the persistence length in bp, L = bp in the bend, and $\Delta\theta$ = the bend angle in degrees. We have chosen to estimate the maximal contribution of bending to IHF binding by using equation 3 and an experimentally determined estimate of the contribution of bending to the free energy associated with the formation of nucleosomes. Shrader and Crothers have determined the maximal free energy associated with DNA flexibility in the formation of nucleosomes in vitro to be 3 kcal/mol (61). This result was obtained by comparing the formation of nucleosomes on their best binding sequences to nucleosome formation on bulk nucleosomal DNA. Nucleosomes bend DNA 630° over a length of 140 base pairs. Assuming that the persistence length of DNA is 150 base pairs, that IHF bends DNA 140° (62) over a length of 40 base pairs, and that both IHF and nucleosome induced bends approximate an arc of a circle, we calculate the bending free energy associated with IHF binding as ±0.52 kcal/mole. This free energy is equivalent to a ±0.4 change in the logarithm of the IHF binding constant, or ±9 similarity score units based on Figure 5.

Recent studies have addressed the role of DNA conformation on the function of the lambda *attP* IHF sites (63,64,65). Goodman and Nash found that the lambda *attP* H2 IHF site can be functionally replaced by a stretch of DNA containing a static bend (63). This finding brings up the possibility that some of the IHF sites may not have evolved towards an optimum DNA conformation for binding because this could obviate the regulatory contribution of IHF. In light of this possibility, we may not be able to derive sequence information pertaining to DNA conformation from a simple analysis of the aligned sequences of IHF sites. The target file used to score the IHF sites did not contain any weighting factors for the contributions of DNA flexibility and conformation to IHF binding. As additional studies provide information concerning the contributions of DNA flexibility and conformation to the binding of IHF, these factors can be incorporated into the weighting matrix and the similarity score equation.

ACKNOWLEDGEMENTS

The work was supported by a grant (GM30375) from the National Institutes of Health; J.A.G. was supported by a National Institutes of Health predoctoral training grant (GM08067). We thank W. Syzbalski for communicating results on IHF binding prior to publication, and N. Grindley for providing an unpublished sequence of the ends of the $\gamma\delta$ transposon. We thank D. Morisato and N. Kleckner for providing *E. coli* strain NK7817. We thank Barbara Bachmann of the *E. coli* Genetic Stock Center for providing strain W3110 (CGSC # 4474).

REFERENCES

1. Mulligan, M.E., Hawley, D.K., Entriken, R., and McClure, W.R. (1984) *Nucleic Acids Res.* 12, 789–800.
2. Berg, O.G. and von Hippel, P.H. (1988) *J. Mol. Biol.* 200, 709–723.
3. Miller, H.I. (1984) *Cold Spring Harbor Symp. Quant. Biol.* 49, 691–698.
4. Mechulam, Y., Fayat, G., and Blanquet, S. (1985) *J. Bacteriol.* 163, 787–791.
5. Flamm, E.L. and Weisberg, R.A. (1985) *J. Mol. Biol.* 183, 117–128.
6. Stenzel, T.T., Patel, P., and Bastia, D. (1987) *Cell* 49, 709–717.
7. Prentki, P., Chandler, M., and Galas, D.J. (1987) *EMBO J.* 6, 2479–2487.
8. Robertson, C.A. and Nash, H.A. (1988) *J. Biol. Chem.* 263, 3554–3557.
9. Kosturko, L.D., Daub, E., and Murialdo, H. (1989) *Nucleic Acids Res.* 17, 317–334.
10. Higgins, P.N., Collier, D.A., Kilpatrick, M.W., and Krause, H.M. (1989) *J. Biol. Chem.* 264, 3035–3042.
11. Kur, J., Hasan, N., and Szybalski, W. (1989) *Gene* 81, 1–15.
12. Yang, C. and Nash, H.A. (1989) *Cell* 57, 869–880.
13. Friedman, D.I. (1988) *Cell* 55, 545–554.
14. Craig, N.L. and Nash, H.A. (1984) *Cell* 39, 707–716.
15. Loeng, J.M., Nunes-Duby, S., Lesser, C.F., Youderian, P., Susskind, M.M., and Landy, A. (1985) *J. Biol. Chem.* 260, 4468–4477.
16. Gamas, P., Galas, D., and Chandler, M. (1985) *Nature* 317, 458–460.
17. Gardner, J.F. and Nash, H.A. (1986) *J. Mol. Biol.* 191, 181–189.
18. Gamas, P., Chandler, M.G., Prentki, P., and Galas, D.J. (1987) *J. Mol. Biol.* 195, 261–272.
19. Mulligan, M. and McClure, W.R. (1986) *Nucleic Acids Res.* 14, 109–126.
20. Hawley, D.K. and McClure, W.R. (1983) *Nucleic Acids Res.* 11, 2237–2255.
21. Schneider, T.D., Stormo, G.D., Gold, L., and Ehrenfeucht A. (1986) *J. Mol. Biol.* 188, 415–431.
22. Stormo, G.D. (1988) *Ann. Rev. Biophys. Biophys. Chem.* 17, 241–263.
23. Stormo, G.D. and Hartzell III, G.W. (1989) *Proc. Natl. Acad. Sci. USA* 86, 1183–1187.
24. Wiater, L.A. and Grindley, N.D.F. (1988) *EMBO J.* 7, 1907–1911.
25. Xin, W. and Feiss, M. (1988) *Nucleic Acids Res.* 16, 2015–2030.
26. Bear, S.E., Court, D.L., and Friedman, D.I. (1984) *J. Virol.* 52, 966–972.
27. Feiss, M., Frackman, S., and Sippy, J. (1985) *J. Mol. Biol.* 183, 239–240.
28. Krause, M.H. and Higgins, P.N. (1986) *J. Biol. Chem.* 261, 3744–3752.
29. Pereira, R.F., Ortuno, M.J., and Lawther, R.P. (1988) *Nucleic Acids Res.* 16, 5973–5989.
30. Lawther, R.P., Wek, R.C., Lopes, J.M., Pereira, R., Taillon, B.E., and Hatfiled, G.W. (1987) *Nucleic Acids Res.* 15, 2137–2155.
31. Huisman, O., Errada, P.R., Signon, L., and Kleckner, N. (1989) *EMBO J.* 8, 2101–2109.
32. Simons, R.W., Hoopes, B.C., McClure, W.R., and Kleckner, N. (1983) *Cell* 34, 673–682.
33. Filutowicz, M. and Appelt, K. (1988) *Nucleic Acids Res.* 16, 3829–3843.
34. Stalker, D.M., Kolter, R., and Helinski, D.R. (1979) *Proc. Natl. Acad. Sci. USA* 76, 1150–1154.
35. Vocke, C. and Bastia, D. (1983) *Proc. Natl. Acad. Sci. USA* 80, 6557–6561.
36. Landy, A. and Ross, W. (1977) *Science* 197, 1147–1160.
37. Hsu, P.L., Ross, W., and Landy, A. (1980) *Nature* 285, 85–91.
38. Griffo, G., Oppenheim, A.B., and Gottesman, M.E. (1989) *J. Mol. Biol.* 209, 55–64.
39. Kur, J., Hasan, N., and Szybalski, W. (1989) *Virology* 168, 236–244.
40. Ohtsubo, H. and Ohtsubo, E. (1978) *Proc. Natl. Acad. Sci. USA* 75, 615–619.
41. Greenstein, D., Zinder, N.D., and Horiuchi, K. (1988) *Proc. Natl. Acad. Sci. USA* 85, 6262–6266.
42. Beck, E. and Zink, B. (1981) *Gene* 16, 35–58.
43. Szekely, E. and Simon, M. (1983) *J. Bacteriol.* 155, 74–81.
44. Alton, N.K. and Vapnek, D. (1979) *Nature* 282, 864–869.
45. Thompson, R.J. and Mosig, G. (1988) *Nucleic Acids Res.* 16, 3313–3326.
46. Kennedy, M., Chandler, M., and Lane, D. (1988) *Biochem. Biophys. Acta* 950, 75–80.
47. Dempsey, W.B. (1987) *J. Bacteriol.* 169, 4391–4392.
48. McIntire, S.A. and Dempsey, W.B. (1987) *J. Bacteriol.* 169, 3829–3832.
49. Funnell, B.E. (1988) *Proc. Natl. Acad. Sci. USA* 85, 6657–6661.
50. Abeles, A.L., Friedman, S.A., and Austin, S.J. (1985) *J. Mol. Biol.* 185, 261–272.
51. Brody, H. and Hill, C.W. (1988) *J. Bacteriol.* 170, 2040–2044.
52. Tsui, P., Helu, V., and Freundlich, M. (1988) *J. Bacteriol.* 170, 4950–4953.
53. Dorman, C.J. and Higgins, C.F. (1987) *J. Bacteriol.* 169, 3840–3843.
54. Eisenstein, B.I., Sweet, D.S., Vaughn, V., and Friedman, D.I. (1987) *Proc. Natl. Acad. Sci. USA* 84, 6506–6510.
55. Mechulam, Y., Blanquet, S., and Fayat, G. (1987) *J. Mol. Biol.* 197, 453–470.
56. Kur, J., Hasan, N., and Szybalski, W. (1990) *Mol. Gen. Genet.* 221, 411–420.
57. Thompson, J.F., Waechter-Brulla, D., Gumport, R.I., Gardner, J.F., Moitoso de Vargas, L., and Landy, A. (1986) *J. Bacteriol.* 168, 1343–1351.
58. Prentki, P. and Kirsch, H.M. (1984) *Gene* 29, 303–313.
59. Trifonov, E.N. and Sussman, J.L. (1980) *Proc. Natl. Acad. Sci. USA* 77, 3816–3820.
60. Gray, H.B. and Hearst, J.E. (1968) *J. Mol. Biol.* 35, 111–129.
61. Shrader, T.E. and Crothers D.M. (1989) *Proc. Natl. Acad. Sci. USA* 86, 7418–7422.
62. Thompson, J.F. and Landy A. (1988) *Nucleic Acids Res.* 16, 9687–9705.
63. Goodman, S.D. and Nash, H.A. (1989) *Nature* 341, 251–254.
64. Snyder, U.K., Thompson, J.F., and Landy A. (1989) *Nature* 341, 255-257.
65. Goodman, S., Yang, C., Nash, H., Sarai, A., and Jernigan, R. (1990) In Sarma R.H. and Sarma M.H. (ed.), *Structure and Methods. Volume 2: DNA Protein Complexes and Proteins.* Adenine Press, Schenectady, NY, pp. 51–62.