J. Mol. Biol. (1994) 242, 330–338

Reaching Out

Locating and Lengthening the Interdomain Linker in AraC Protein

Rebecca J. Eustance, Silvia A. Bustos and Robert F. Schleif

Biology Department, Johns Hopkins University
3400 N. Charles Streets, Baltimore, MD, U.S.A.

A genetic method was developed to determine, in proteins, areas which are tolerant of insertions and deletions. Attractive candidates for these areas are linker regions. Such a region was found to include positions 171 to 178 in the *Escherichia coli* regulatory protein AraC. Independent biochemical methods identified amino acid residues 11 to 170 as the minimal dimerization domain of AraC, and amino acid residues 178 to 286 out of the 291 residue protein as the minimal DNA-binding domain. Hence, by both the genetic and biochemical approaches, the interdomain linking region was determined to include amino acid residues 171 to 177. The properties of altered proteins were examined using templates with AraC half-sites more widely separated than in the wild-type case. Both AraC protein containing an insertion in the interdomain linker region and a protein consisting of the minimal functional dimerization and DNA-binding domains separated by a 39 amino acid residue linker were able to bind to and function on such a DNA site. *In vitro*, the proteins with longer linkers bound substantially more stably than wild-type AraC to the DNA containing half-sites for AraC separated by an extra two helical turns of DNA. *In vivo* on an *ara* promoter with the more widely separated AraC half-sites, the proteins could activate transcription much better than wild-type AraC.

Keywords: domain; gene regulation; protein-DNA interaction; protein-protein interaction; transcriptional activation

1. Introduction

In addition to repressing its own synthesis, AraC protein activates transcription of the genes necessary for the catabolism of L-arabinose in *Escherichia coli* (Englesberg, 1961; Englesberg *et al.*, 1965, 1969; Greenblatt & Schleif, 1971; Casadaban, 1976; Ogden *et al.*, 1980). Like many eukaryotic transcription factors and some prokaryotic regulatory proteins (Little & Hill, 1985; Sauer *et al.*, 1988; Corton & Johnston, 1989; Harrison, 1991; Marmorstein *et al.*, 1992) the dimeric regulatory protein AraC contains separable functional domains for dimerization and DNA binding (Bustos & Schleif, 1993). These domains are independent folding units that retain their function within a chimeric protein. On one hand, the protein's ability to bind tightly to DNA with variations in half-site separation or even inversion of a half-site implies a loose connection between the domains (Carra & Schleif, 1993). On the other hand, information concerning the presence of arabinose is communicated from the dimerization to the DNA-binding domain (Lobell & Schleif, 1990; Bustos & Schleif, 1993; Carra & Schleif, 1993). Therefore, at least some use is made of a flexible linker that must connect the domains.

The available evidence on the *ara* system suggests that in the absence of arabinose, the two DNA-binding domains of AraC are connected to a dimerization domain by a flexible linker (Carra & Schleif, 1993) that permits them to bind by virtue of DNA looping resulting in the binding of the DNA binding domain on one subunit to the I_1 half-site at p_{BAD} and the DNA binding domain on the other to the O_2 half-site located 210 bp away (Lobell & Schleif, 1990; Figure 1). When arabinose is added, the connection between the DNA binding domains is postulated to shorten and perhaps stiffen and thereby interfere with binding to distally located half-sites (Carra & Schleif, 1993). Consequently, the domain formerly bound to the O_2 half-site relocates to the weak binding I_2 half-site adjacent to the I_1 half-site (Lobell & Schleif, 1990; Figure 1). Occupancy of this half-site then activates transcription (Reeder & Schleif, 1993).

The postulated constraining of the DNA-binding domains of AraC in the presence of arabinose could come about through two different mechanisms. The

Rebecca J. Eustance and Silvia A. Bustos have contributed equally to this work.

Present address: Silvia A. Bustos, Becton Dickinson Advanced Diagnostics, 54 Loveton Circle, Sparks, MD 21152, U.S.A.

Correspondence: Robert F. Schleif, Biology Department, Johns Hopkins University, 34th and Charles Streets, Baltimore, MA, U.S.A.

0022-2836/94/390330-09 $08.00/0

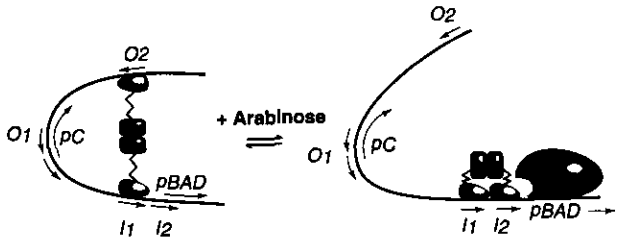


Figure 1. Schematic representation of the *araCBAD* regulatory region. The presence of arabinose breaks the repression loop by inducing a conformational change in AraC that favors the contact of adjacent DNA sites, resulting in activation of p_{BAD}.

dimerization and arabinose-binding domain of AraC could undergo a large conformational change upon arabinose binding. Despite the presence of a flexible linker in AraC, effects of this conformational change could be transmitted to the DNA-binding domains. In this view, the linker plays the relatively passive role of a connector region. At the other extreme, the linker connecting the dimerization and DNA-binding domains could play a more active role in altering the protein's participation in DNA looping events *versus* binding to two adjacent half-sites. In principle, these models could be distinguished by substituting the linker region of AraC by other sequences known to serve a similar function. To begin an investigation of the looping-unlooping phenomenon, we have undertaken first to locate the linker region in AraC protein, and second to explore the consequences of lengthening the linker. All of the following studies are done in the absence of arabinose to preclude the possibility that altering the linker affects AraC protein's response to the sugar.

Lacking the tertiary structure of AraC protein, we developed a genetic method and used a biochemical method to locate and study the flexible linker that connects the dimerization and DNA-binding domains. Since a linker region should be more tolerant of insertions and deletions than a functional domain, the basis of the genetic technique is the facile generation of random insertions and deletions and the identification of those that retain AraC protein activity. In the biochemical method, we progressively deleted the dimerization or DNA-binding domain of AraC in chimeric proteins in order to define the domain boundaries. Both approaches localized the interdomain linker to include amino acid residues 171 to 177.

With the linker region located, it was possible then to insert extra amino acid residues. Although wild-type AraC protein possesses significant flexibility in its ability to bind to DNA in which the AraC half-sites have been inverted or moved further apart, the protein binds weakly when the half-sites

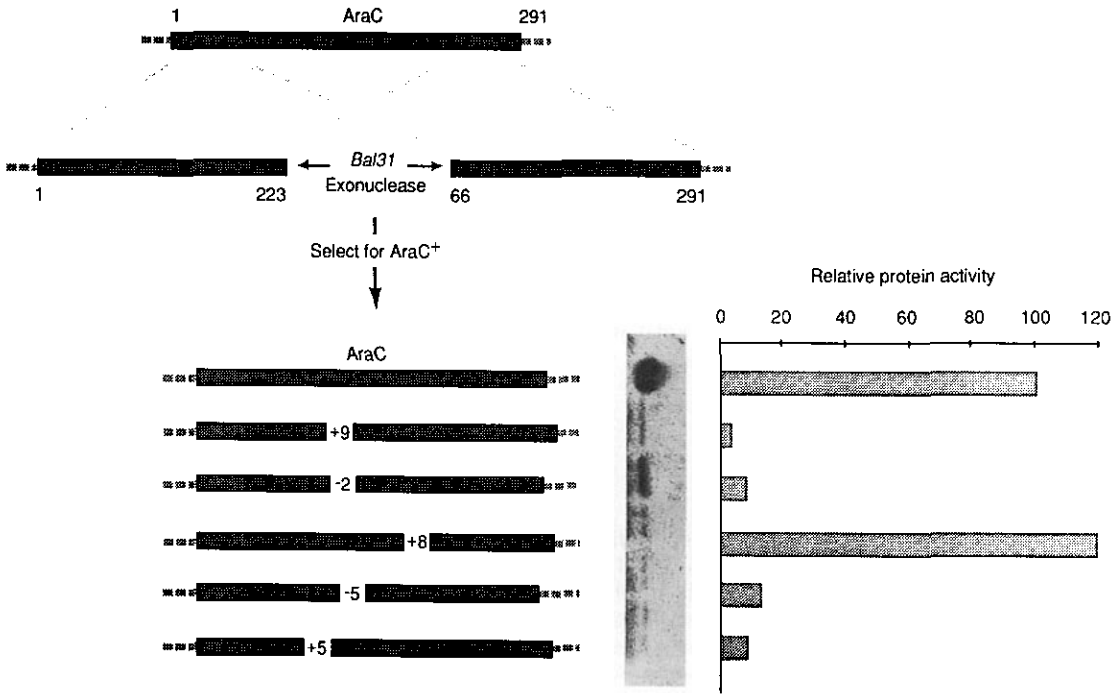


Figure 2. Genetic method for insertion and deletion generation, solubility and relative activity of the products. DNA encoding the indicated sections of AraC protein were joined with the addition of a unique cleavage site between them. After endonuclease digestion, *Bal*31 exonuclease treatment, ligation and transformation, AraC$^+$ transformants were selected on minimal arabinose plates. Insertion or deletion mutants are indicated by + or −, respectively, and the numbers within the protein represent amino acid residues duplicated or deleted. The modified residues are: 112 to 120 (+ 9), 115–116 (− 2), 171 to 178 (+ 8), 122 to 126 (− 5) and 95 to 99 (+ 5). Immunoblots of pure AraC or lysates prepared from cell carrying plasmids encoding these proteins reacted with a polyclonal AraC antibody show that each protein is present (Sambrook *et al.*, 1989). Activation of transcription from I_1-I_1p_{BAD}-*lacZ* (Reeder & Schleif, 1993) is shown with wild-type AraC activity at 100%. Activities are uncorrected for protein levels.

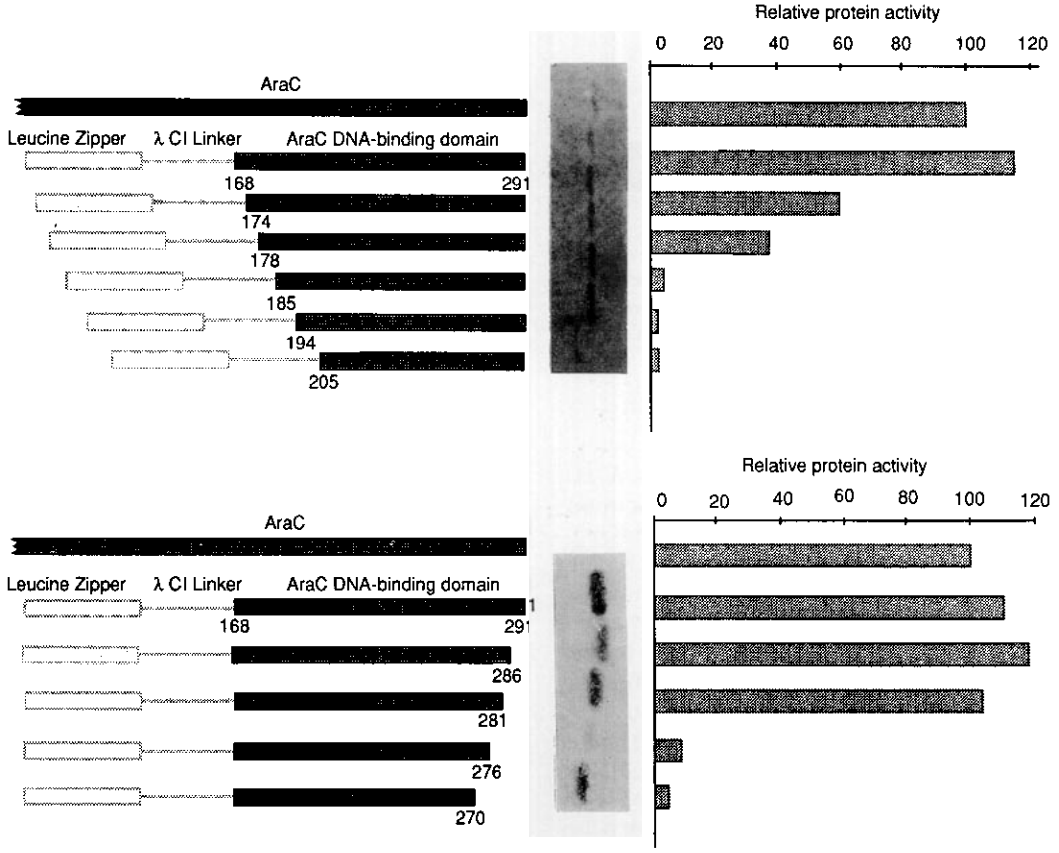


Figure 3. Deletions, solubility and relative activity of truncated chimeric proteins containing the AraC DNA-binding domain. The schematic drawing shows the length of the deleted (leucine-zipper)-(λ CI linker)-(AraC DNA-binding domain) chimeric proteins. Numbers indicate the first or last amino acid residue included in each construct. Unnumbered boundaries were constant among constructs. Immunoblots of bacterial cell lysates were reacted with a monoclonal antibody to AraC (top) or a polyclonal antibody to the C/EBP protein (bottom) which recognizes the leucine-zipper region. Activation of transcription from I_1-I_1p_{BAD}-*lacZ* is shown with respect to wild-type AraC activity. Wild-type AraC is expressed from the same plasmid as the truncated chimeric proteins. Activities are uncorrected for protein levels.

have been separated by 21 bp of DNA (Carra & Schleif, 1993). Therefore we tested the ability of the variant proteins to bind to such a site and to induce transcription *in vivo*. The AraC proteins with longer linker regions both bound to, and activated transcription well from separated half-sites.

2. Materials and Methods

(a) *Plasmids and strains*

In the plasmid used for the genetic technique, the first AraC coding region was made by polymerase chain reaction (PCR†) amplification of pAB1003 (Brunelle & Schleif, 1989) from the *Pst*I site upstream of the promoter at position − 766 to position + 928 which lies within the *araC* gene. This product includes the region coding for amino acid residues 1 to 223. The downstream oligonucleotide in this PCR reaction created a *Hin*dIII site. This DNA fragment was digested with *Pst*I and *Hin*dIII and ligated into a similarly digested pAB1003 vector. Two oligonucleotides were hybridized to create *Hin*dIII, *Bgl*II, *Xho*I, *Sal*I, and *Hin*dIII cleavage sites on a 44 bp fragment. We cloned this fragment into the *Hin*dIII site of the plasmid now coding for amino acid residues 1 to 223 of AraC. A second PCR amplified the region on pAB1003 from + 454 to + 1132. This DNA fragment carries DNA coding for AraC amino acid residues 66 to 291 and includes a translational stop site. The oligonucleotides in this reaction placed a *Sal*I site on either end of the DNA fragment. This DNA was cloned into the *Xho*I and *Sal*I sites in the plasmid. The final construct (depicted schematically in Figure 2) consists of a *lacUV5* promoter followed by a coding region for amino acids 1 to 223 of AraC, a unique *Bgl*II restriction endonuclease site, and a coding region for amino acids 66 to 291 of AraC. All plasmids were transformed into strain RE1 (Δ*araC-leu*1022, *araB*$^+$*A*$^+$*D*$^+$, Δ*lac74*, *galK*$^-$ *str*^{r} *recA938*::*cat*). RE1 was constructed by bacteriophage P1 mediated transduction of the *recA938*::*cat* allele into strain SH321 to make it RecA$^-$ (Miller, 1972; Hahn *et al.*, 1984). We verified the sequence of the construct with double stranded sequencing (Kraft *et al.*, 1988).

To insert the lambda phaga repressor CI linker into the chimeric proteins, two DNA fragments encoding amino acids 93 to 131 of the λ phage CI repressor (Sauer, 1978) were PCR amplified from bacteriophage λ DNA. The products were cloned into a unique *Bam*HI restriction endonuclease site at the domain junction of the leucine-zipper-AraC DNA-binding domain or the AraC dimerization domain-LexA DNA-binding domain chimeras (Bustos & Schleif, 1993). The two linkers were constructed so that both the leucine-zipper-λCI-AraC and the AraC-λCI-LexA

† Abbreviation used: PCR, polymerase chain reaction.

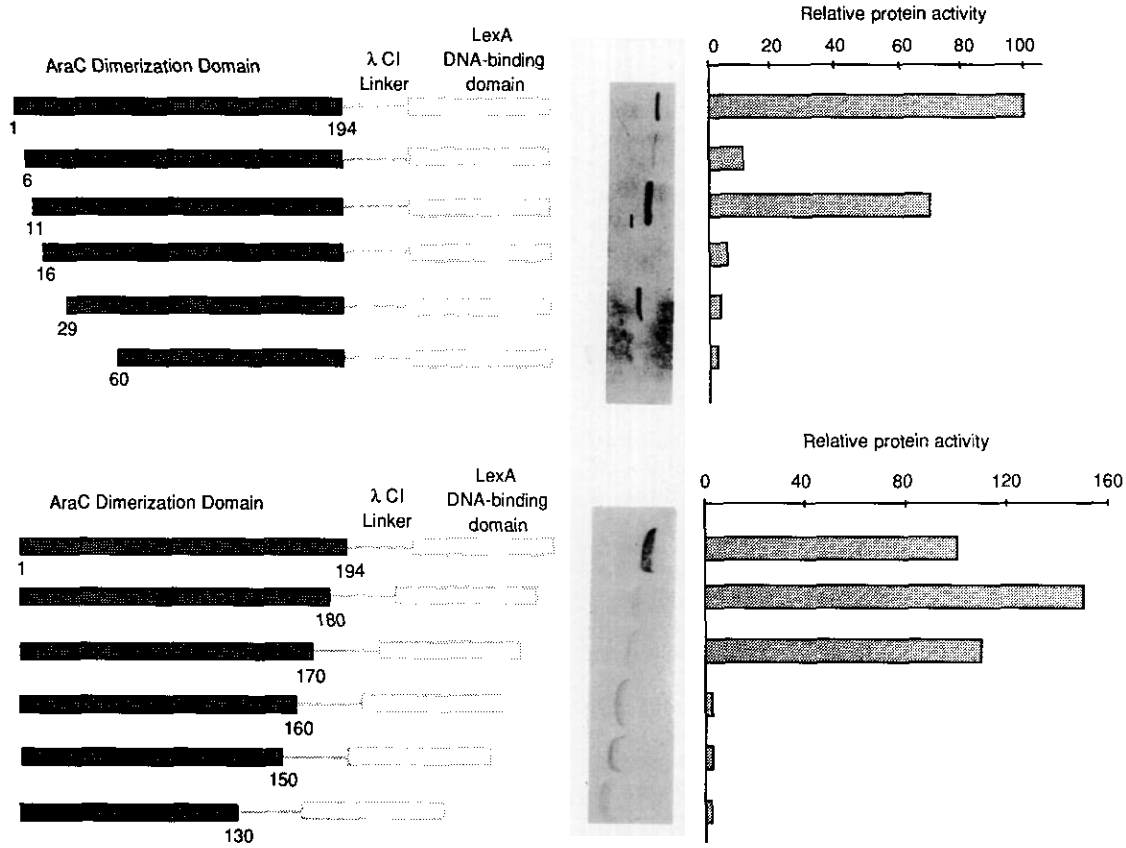


Figure 4. Deletions, solubility and relative activity of truncated chimeric proteins containing the AraC dimerization domain. The schematic drawing represents the (AraC dimerization domain)-(λ CI linker)-(LexA DNA-binding domain) chimeric proteins. Immunoblots of bacterial cell lysates of a *lexA*⁻ strain (Lin & Little, 1989) reacted with a polyclonal antibody to the LexA protein show the presence of each protein. Relative protein activity is shown as repression of the p_{sulA}-*lacZ* promoter with undeleted (AraC dimerization domain)-(λ CI linker)-(LexA DNA-binding domain) as 100%. Activities are uncorrected for protein levels.

constructs retained an intact *Bam*HI site adjacent to the *araC* encoded segment of the chimera but not at the other end. The series of truncations of DNA coding for the AraC domains within the chimeras was made by synthesizing the appropriate PCR products using oligonucleotides that carry a *Bam*HI site and hybridize to different sites within the *araC* gene. The DNA fragments were cloned into a modified pSE380 expression vector (Bustos & Schleif, 1993). All mutants were sequenced using double stranded sequencing (Kraft *et al.*, 1988).

(b) *Exonuclease digests*

The completed plasmid used for the genetic technique (Figure 2) was caesium chloride purified as described (Schleif & Wensink, 1981). The restriction endonuclease *Bgl*II was used to digest 10 μg of the DNA. The ends of the linear DNA were then digested with three units of *Bal*31 exonuclease (Wei *et al.*, 1983) as described by Reeder & Schleif (1993). Samples were removed from the reaction over a period of 2·5 hours and quenched in 25 mM EGTA (pH 8·0), 10 mM Tris-HCl (pH 8·0) at 65°C. Following the digestion, the DNA was made blunt-ended by treatment with the Klenow fragment of DNA polymerase I (USB), ligated closed, and transformed into the strain RE1. Following transformation (Sambrook *et al.*, 1989), the YT medium was removed by centrifugation and cells were plated on plates with A + B salts, 10 μg thiamine/ml, 40 μg leucine/ml, and 0·2% (w/v) arabinose (Schleif & Wensink, 1981). It was necessary to remove the YT medium to ensure the growth of only AraC⁺ strains. All plasmids from colonies that grew within three days were sequenced and no new colonies were detected beyond three days.

(c) *DNA migration retardation assay*

The DNA migration retardation assay was performed as described (Hendrickson & Schleif, 1984). Radiolabeled DNA fragments were synthesized by PCR amplification using 25 ng of the 5'-end-labeled oligonucleotide primer (Carra & Schleif, 1993) with 100 ng of an unlabeled oligonucleotide primer. The template was ES51 (Lobell & Schleif, 1990) with either an I_1-I_2 site or an I_1-*21*-I_1 site (Carra & Schleif, 1993). The DNA fragments were purified as described (Bustos & Schleif, 1993). Assays were performed using pure AraC or crude cell lysates of *E. coli* carrying the appropriate overexpression plasmid (Bustos & Schleif, 1993). The binding reactions contained 0·5 μg of sheared calf thymus DNA for every 0·2 μl crude cell lysate (approximately 4×10^{-2} mg total protein/ml) and labeled DNA in a total volume of 20 μl. The volume of crude lysate in the reactions ranged from 0·2 μl to 6 μl which provides a limiting amount of protein. Binding reactions were incubated at 37°C for five minutes, sufficient for binding equilibrium for all experiments reported here. Cold competitor DNA, when added, consisted of four I_1 half-sites each separated by four base-pairs on an 88 bp linear DNA

fragment. The amounts of the DNA species were obtained using a Molecular Dynamics Phosphorimager® SF.

(d) *β-Galactosidase assay*

Wild-type AraC, AraC mutants, and chimeric proteins were expressed from a modified pSE380 vector (Invitrogen) containing a p15A origin of replication (Bustos & Schleif, 1993) in strain RE1. Cells were grown in minimal salt media with glycerol and Casamino acids as described by Reeder (1993) and the assays were performed as described by Miller (1972). All values were obtained as the average of β-galactosidase levels from two independent assays.

(e) *Immunoblot assays*

Cells carrying the plasmids containing the various *araC* deletion genes were grown in YT medium to about 2×10^8 cells/ml, spun down, resuspended in lysate buffer (Bustos & Schleif, 1993), and approximately 4×10^{-3} mg of total cell protein were loaded per lane on a 10 × 8 cm SDS-containing gel, or approximately 8×10^{-4} mg of total cell protein were loaded per lane on a 18 × 16 cm gel. After electrophoresis the proteins were transferred to nitrocellulose by electroblotting (BioRad) (Sambrook *et al.*, 1989) and incubated with AraC, LexA or C/EBP-specific antibody (Figures 2, 3 and 4). Antigen was detected using alkaline phosphatase-conjugated anti-mouse or anti-rabbit secondary antibodies (Promega) with a Protoblot Western Blot AP system from Promega. The amount of protein used in these assays was in a region where the darkness of the band was proportional to the amount of protein.

3. Results

(a) *Genetic localization of linker regions*

Protease digestion experiments could not identify the linker region in AraC protein. The C-terminal third of the protein is rapidly digested by proteases in the absence of DNA, and the entire protein is resistant to protease cleavage when bound to DNA (Carra, 1993). A genetic method was therefore developed to identify and localize the linker region of AraC.

Figure 2 shows the method we used to generate random duplications and deletions in AraC. A plasmid was constructed containing a promoter and the coding region for the first 3/4 of AraC in one block and the coding region for the last 3/4 of AraC in an adjacent block. Neither block produces a functional protein, but active protein could be obtained, wild-type for example, by the removal of all the redundant coding information and fusion of the remainder of each block. The removal can be accomplished by endonuclease cleavage between the two blocks, digestion with exonuclease *Bal*31, ligation to reform the plasmid circle, and transformation into $AraC^-$ cells where $AraC^+$ candidates are selected. In addition to regenerating the wild-type coding region, internal tandem duplications or internal deletions will result from removal of less than or more than just the redundant coding regions. Further, many different locations and sizes of duplications and deletions will be found due to variations within the sample in the amounts of DNA digested by the exonuclease. Although the method described above can generate functional AraC protein derivatives with deletions scattered over the entire length of the gene, insertions in the form of tandem duplications are confined to the central half of the protein that is duplicated in the original coding blocks.

Control experiments were used to identify the reaction times necessary to obtain digestion endpoints scattered across the entire coding blocks. Variations from molecule to molecule, so as to generate a wide variety of products, were likely to have been significant in these experiments because among the barely functional variants were some retaining only amino acids 1 to 16 fused to amino acids 177 to 291 (data not shown). To ensure the broadest range of possible products, the selection for active fusions was designed to yield even weakly positive $AraC^+$ products. Out of 5000 transformants, aside from the drastic deletions mentioned above, we found five $AraC^+$ transformants with in-frame insertions or deletions in AraC.

As shown in Figure 2, one of the proteins, containing a duplication of amino acids 171 to 178, HPPMDNRV, activated transcription at near wild-type levels, whereas the others were less than 10% as active. To assess the possibility that an insertion or deletion made a particular protein protease sensitive or insoluble, we performed immunoblot assays to detect soluble protein in the cell. This analysis showed that all the variant AraC proteins were present at their expected sizes and soluble at sizeable levels. Thus, the one insertion protein with near wild-type activity identifies a potential linker region in AraC.

(b) *Biochemical identification of the minimal domains*

Independent of the genetic method for locating the linker region connecting the domains in AraC, we also utilized a biochemical method. Previously we have shown that functional chimeric proteins can be constructed in which either the dimerization or DNA-binding domain is provided by AraC protein and the foreign DNA-binding or dimerization domain originates from the LexA protein or the leucine zipper of C/EBP (Bustos & Schleif, 1993). The activities of either of these chimera types are readily assayed and thus it is possible to construct a series of progressively shortened chimeric proteins to locate the minimal dimerization and DNA-binding domains of AraC. One precaution was necessary, however. Because activity of the chimeric proteins could depend upon the linker region provided by AraC, a deleted construct could be inactive either because a linker between the domains was necessary and was deleted, or because amino acids necessary for the function of a domain were removed. Therefore, we included the 39 amino acid linker from the *E. coli* λ phage CI repressor (Sauer, 1978) as a supplemental linker within the chimeras. This linker includes λ phage CI repressor amino acids 93 to 131, MQPSLRSEYEYPVFSHVQAGMFSPE-LRTPTKYDAERWVS. The λ phage CI repressor and the P22 c2 repressor have similar domain

structures and both contain linkers that are not completely flexible and available to the solvent. The P22 c2 repressor, in fact, is only susceptible to tryptic digest at either end of the linker at amino acids 83 and 107. The linker then remains intact and associated with the carboxy terminal oligomerization domain (De Anda *et al.*, 1983). While the λ phage CI linker may not act as a completely flexible linker, it serves our ends as a longer spacer between the AraC protein domains.

Figure 3 shows the results obtained with the (leucine-zipper)-(λ CI linker)-(AraC DNA-binding domain). Shown are the structures of the various deletion proteins, their *in vivo* levels of soluble protein as assayed by immunoblots, and their activities in stimulating transcription from an AraC-inducible promoter (Reeder, 1993; Reeder & Schleif, 1993) fused to *lacZ*. Amino acids 178 to 291 from the AraC DNA-binding domain are sufficient for activation when dimerized by the leucine-zipper domain, but amino acids 185 to 291 are not. The immunoblot assays showed that the levels of the shortened chimeric proteins differed appreciably from protein to protein, presumably as a result of differing *in vivo* stabilities. These levels, however, were higher than those normally found for wild-type AraC protein. Thus, low activity for a protein predominantly results from inactivity rather than insufficiency. Figure 3 thus shows that the N-terminal boundary of the DNA-binding and transcription activation domain lies between amino acids 178 and 185. Similarly, the C-terminal boundary of the domain lies between amino acids 276 and 281. Essentially the same results were obtained when the ability of AraC to bind to the O_I site and repress the *ara* p_C promoter was assayed (data not shown). Since levels of activation paralleled DNA-binding ability, the activation domain of AraC must lie within the stable domain that binds to DNA, and not in a series of amino acids that extends beyond the DNA-binding domain.

Figure 4 shows the analogous results obtained with chimeras of structure: (AraC dimerization domain)-(λ CI linker)-(LexA DNA-binding domain). The function of these proteins was measured with the fusion of the LexA-repressible promoter p_{sulA} to *lacZ* (Lin & Little, 1989). The N-terminal boundary of the dimerization domain lies between amino acids 11 and 29 and the C-terminal boundary between amino acids 160 and 170. No protein is detectable in the immunoblot for the AraC dimerization domain deleted to amino acids 16 to 194, therefore, the minimal N-terminal boundary is given as amino acid residue 29.

A "minimal domains" AraC protein (amino acids 11 to 170)-(λ CI linker)-(amino acids 178 to 286) is also functional; the protein both represses p_C (data not shown) and activates p_{BAD} 50% as well as the wild-type protein. We thus conclude from biochemical measurements that the linker region connecting the dimerization and DNA-binding domains in AraC comprises at a minimum, amino acids 171 to 177, HPPMDNR, and at a maximum, amino acids 161 to 184, RRMEAINESLHPP-MDNRVREACQY. The genetically isolated tandem duplication discussed in the previous section at amino acids 171 to 178 is right in the middle of the biochemically located linker.

(c) *The lengthened linker proteins can bind by reaching further*

To study the binding abilities of AraC and the extended linker proteins, we measured in cell lysates their relative dissociation half-times from I_1-*21*-I_1. The wild-type binding site for AraC consists of two 17 bp half-sites, I_1 and I_2, separated by four base-pairs (Carra & Schleif, 1993). This is denoted by I_1-I_2. When the tight binding half-site, I_1 is substituted for the weaker binding half-site, I_2, and moved an additional 21 bp away, the site is denoted as I_1-*21*-I_1. The relative dissociation rates of the proteins from I_1-*21*-I_1 were obtained by incubating the DNA with AraC until an equilibrium distribution of the protein was achieved. Cold competitor DNA containing four I_1 half-sites each separated by four base-pairs to form optimal binding sites was added, and at various times the amount of AraC still bound to DNA was determined using the DNA migration retardation assay (Hendrickson & Schleif, 1984). The free DNA and DNA complexed with AraC could easily be separated by electrophoresis and quantified.

Figure 5 shows that wild-type AraC quickly dissociates from I_1-*21*-I_1 in the presence of cold competitor DNA. The AraC protein with eight additional amino acids in the linker and AraC protein with a 39 amino acid linker both dissociate much more slowly. Quantification of the band intensities in the DNA migration retardation assay indicates that half of the wild-type protein has dissociated in under 30 seconds, while the altered linker proteins exhibit the same relative dissociation in greater than five minutes.

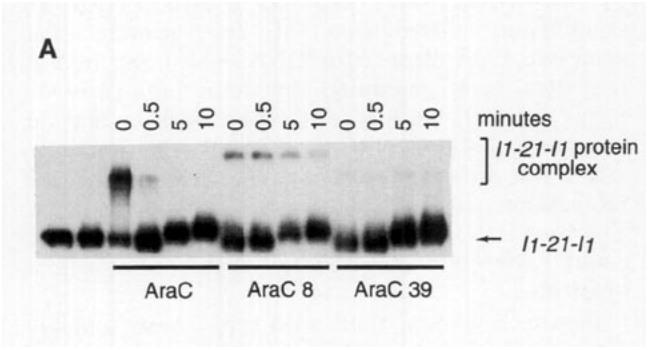
DNA with bound wild-type AraC, AraC with an eight amino acid duplication in the linker, and minimal domain AraC with a 39 amino acid linker all run to different positions in the DNA migration retardation assay when complexed with an I_1-*21*-I_1 site. Why? It seems likely that DNA bending is the root cause and that longer linkers do not require the I_1-*21*-I_1 DNA to bend as much when the protein contacts both half-sites simultaneously. To test this idea we compared the retardation of two DNA fragments, one in which an I_1-*21*-*I1* site was located at the end, and the other where the site was in the middle. The differences in migration rates vanished when the I_1-*21*-I_1 binding site was placed on the end of the DNA fragment as would be expected if the differing migration rates were generated by differing DNA bends at the I_1-*21*-I_1 site (Figure 6; Wu & Crothers, 1984). This shows that the amount of DNA bending in the protein-DNA complex depends

upon the length of the amino acid linker connecting the dimerization and DNA binding domains of AraC.

We note that in typical retardation experiments as DNA bending increases, retardation increases. Our experiments showed the opposite. This results from the fact that the DNA is so sharply bent by the binding of AraC. The retardation of DNA as a function of DNA bend must go through a maximum at some angle intermediate between no bend and a 180° bend. Our situation includes points on the part of the curve where increasing bending results in faster migration.

(d) In vivo *reaching and activation by proteins with longer linkers*

Since the proteins with longer linkers are able to reach to half-sites more widely separated than normal, the next question is whether such proteins can still activate transcription when they bind to such separated half-sites. Figure 7 shows the relative inducing abilities of several AraC derivatives on an I_1-I_1-p_{BAD} promoter fused to *lacZ* and on an I_1-21-I_1-p_{BAD} promoter fused to *lacZ* (Reeder & Schleif, 1993). The proteins with insertions in their linker region were dramatically better than wild-type AraC at inducing I_1-21-I_1-p_{BAD}.



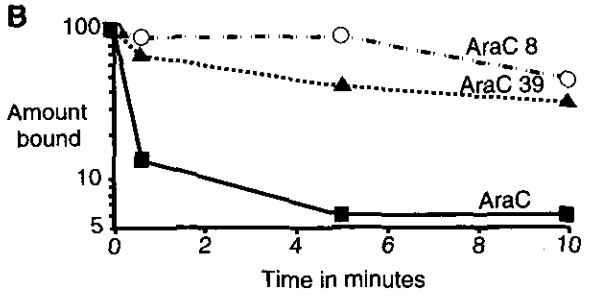


Figure 5. AraC proteins with extended linkers show slower dissociation from I_1-21-I_1 than wild-type AraC. A, DNA migration retardation assays of radiolabeled DNA fragments containing an I_1-21-I_1 operator reacted with 0·2 μl of lysate from *araC*⁻ cells or 0·2 μl, 0·4 μl, or 6 μl of lysates from cells transformed with plasmids encoding wild-type AraC, AraC with eight additional amino acid residues in its linker (AraC 8), or AraC with a 39 amino acid residue linker (AraC 39), respectively. 50 ng of cold competitor DNA 88 bp long was incubated with the reaction for 0·5 min (0·5), 5 min (5), 10 min (10) or was not added (0). The DNA-protein complexes are shown along with the unbound DNA. Lane 1 contains free DNA alone, and lane 2 contains DNA plus lysate from *araC*⁻ cells. B, The DNA species were quantified with a Molecular Dynamics Phophorimager SF. The dissocation of AraC and AraC derivatives from I_1-2$_1$-I_1 is plotted against time with the amount of protein bound to DNA in the absence of competitor given a value of 100.

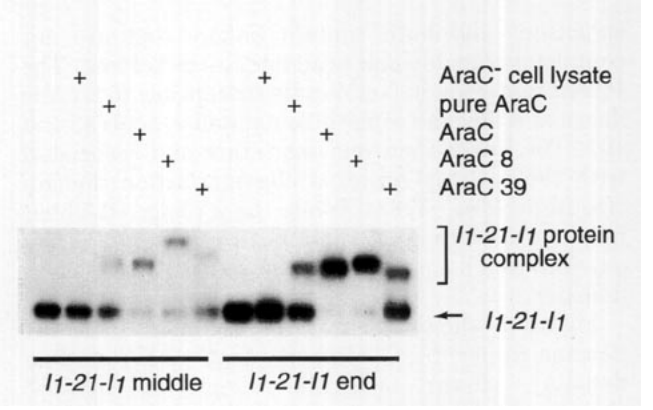


Figure 6. AraC protein bends DNA differently depending on linker length. The DNA migration retardation assay shows radiolabeled DNA fragments of the same size with an I_1-21-I_1 site either at the end of the DNA fragment or in the middle. The DNA fragments are incubated with pure AraC or cell lysates containing no AraC (AraC⁻ cell lysate), wild-type AraC (AraC), AraC with 8 additional amino acid residues in the linker (AraC 8), or AraC with a 39 amino acid residue linker (AraC 39). Lane 1 and lane 7 contain free DNA. DNA-protein complexes and unbound DNA are shown.

4. Discussion

Using both biochemical and genetic techniques we have defined the minimal functional dimerization and DNA binding domains of AraC and located the interdomain linker connecting them. We have shown that increasing the length of the linker region, for example in a minimal domains protein with a 39 amino acid linker, increased the ability of the protein to bind to and activate transcription from promoters in which the half-sites of the AraC DNA-binding site were separated more widely than normal. We have also shown that the AraC interdomain linker likely is very flexible since it tolerates an insertion without disrupting protein function.

Previous reports showed that insertions of one amino acid could be tolerated in Staphylococcal nuclease, throughout much of the protein (Sondek & Shortle, 1990). More sizeable insertions within methionyl-tRNA synthetase inactivated the protein except when inserted in a surface loop (Starzyk *et al.*, 1989). In both cases, however, the insertions were targeted to regions of the proteins selected based on their tertiary structure. One of the advantages of our genetic technique which locates insertion-tolerant

and deletion-tolerant regions is that it can identify surface loops or linker regions in proteins with unknown structure. Additionally, the procedure described here should be useful in creating functional chimeras between two proteins with unknown domain locations. The regions present in functional chimeras would identify the boundaries of the functional domains.

One example of an altered ability of a dimeric protein to bind to half-sites with altered separation is a fusion protein between the DNA binding domain of LexA and the leucine zipper from the jun oncogene protein. This chimera tolerated insertion of two additional base-pairs between its DNA half-sites whereas the wild-type LexA protein could not bind to the site (Oertel-Buchheit *et al.*, 1993). Natural examples appear to be closer to our case of changing the length of a linker arm and finding the resultant protein capable of binding to half-sites with larger than wild-type separation.

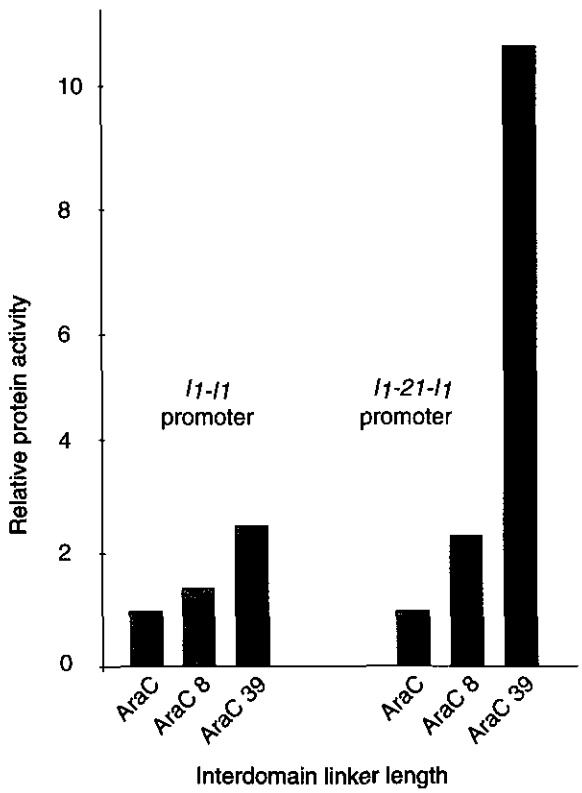


Figure 7. Interdomain linker length correlates with the ability to activate transcription *in vivo* from I_1-*21*-I_1. Relative protein activity is expressed with respect to the ability of wild-type AraC to activate I_1-I_1p_{BAD}-*lacZ* or I_1-*21*-I_1p_{BAD}-*lacZ* (Reeder, 1993). AraC, AraC 8 and AraC 39 represent wild-type AraC, AraC with an 8 amino acid residue duplication, and the minimal domains protein with a 39 amino acid residue linker, respectively. β-Galactosidase levels were assayed in cells co-transformed: with plasmids expressing AraC, AraC 8, or AraC 39 proteins with compatible plasmids carrying I_1-I_1p_{BAD}-*lacZ* or I_1-*21*-I_1p_{BAD}-*lacZ*, respectively. Wild-type AraC is 6-fold less active on I_1-*21*-I_1p_{BAD}-*lacZ* than on I_1-I_1p_{BAD}-*lacZ* with 3500 Miller units and 20,000 Miller units, respectively. Activities are uncorrected for protein levels.

The yeast transcription activators PPR1 and Gal4 have a similar repeated amino acid sequence in their DNA recognition modules and bind to similar DNA half-sites. PPR1, however, has a longer interdomain linker than Gal4 and as a consequence PPR1 is specific for half-sites spaced farther apart on the DNA than Gal4 (Corton & Johnston, 1989; Marmorstein *et al.*, 1992). In addition, the HsdS polypeptide in *E. coli* contains a four amino acid repeat that is present twice in *Eco*R124 and three times in *Eco*R124/3. These type IC restriction modification enzymes recognize the same sequence separated by six nucleotides in the former case, and seven in the latter (Gubler & Bickle, 1991). Presumably the extra repeat specifies a greater separation of the portions of the protein that recognize DNA sequence.

Linker length has also been experimentally altered in LexA protein where two to eight amino acids were deleted from the linker region (Little & Hill, 1985). The resulting protein was still capable of binding its natural DNA site. The yeast Matα2 protein is capable of binding DNA half-sites separated by a variety of spacings and therefore appears to have an extremely flexible linker. If, however, the MCM1 protein is also present, the flexibility of the protein is apparently restricted, and the protein complex is only found bound to the wild-type orientation of half-sites (Sauer *et al.*, 1988).

Because we do not know the length of the linker connecting the dimerization and DNA binding domains in the wild-type AraC protein, we cannot know the true lengths of the linkers in the derivative proteins we have studied. None the less, on the assumption that the linker is a random coil with 4·3 Å per amino acid residue, the additional lengths we have added by the insertion of eight or 31 amino acids is of the same order as the extra 70 Å separating the half-sites in I_1-*21*-I_1. In conclusion, we have located the linker region in AraC protein and have found that an elementary prediction about it, lengthening the linker should permit the DNA-binding domains to reach further apart, was met.

We are grateful to John Little and Steven McKnight for providing the antibodies against LexA and C/EBP, to David Draper, Ed Lattman, David Shortle and Cynthia Wolberger for comments on the manuscript, to the members of the laboratory for critical discussion and comments on the manuscript, and to Jeff Withey for technical assistance. This material is based upon work supported by a National Science Foundation grant BIR-9207432 to S. A. B. and by the National Institutes of Health grant GM18277 to R. S.

References

Brunelle, A. & Schleif, R. F. (1989). Determining residue-base interactions between AraC protein and *araI* DNA. *J. Mol. Biol.* **209**, 607–622.

Bustos, S. A. & Schleif, R. F. (1993). Functional domains of the AraC protein. *Proc. Nat. Acad. Sci., U.S.A.* **90**, 5638–5642.

Carra, J. H. (1993). Ph.D. thesis, Brandeis University, Waltham, MA.

Carra, J. H. & Schleif, R. F. (1993). Variation of half-site organization and DNA looping by AraC protein. *EMBO J.* **12**, 35–44.

Casadaban, M. J. (1976). Regulation of the regulatory gene for the arabinose pathway, araC. *J. Mol. Biol.* **104**, 557–566.

Corton, J. & Johnston, S. (1989). Altering DNA-binding specificity of GAL4 requires sequences adjacent to the zinc finger. *Nature (London)*, **340**, 724–727.

De Anda, J., Poteete, A. R. & Sauer, R. T. (1983). P22 c2 repressor. Domain structure and function. *J. Biol. Chem.* **258**, 10536–10542.

Englesberg, E. (1961). Enzymatic characterization of 17 L-arabinose negative mutants of *Escherichia coli*. *J. Bacteriol.* **81**, 996–1006.

Englesberg, E., Irr, J., Power, J. & Lee, N. (1965). Positive control of enzyme synthesis by gene C in the L-arabinose system. *J. Bacteriol.* **90**, 946–957.

Englesberg, E., Squires, C. & Meronk, F. (1969). The L-arabinose operon in *Escherichia coli* B/r: a genetic demonstration of two functional states of the product of a regulator gene. *Proc. Nat. Acad. Sci., U.S.A.* **62**, 1100–1107.

Greenblatt, J. & Schleif, R. F. (1971). Arabinose C protein: regulation of the arabinose operon *in vitro*. *Nature New Biol.* **233**, 166–170.

Gubler, M. & Bickle, T. A. (1991). Increased protein flexibility leads to promiscuous protein–DNA interactions in type IC restriction-modification systems. *EMBO J.* **10**, 951–957.

Hahn, S., Dunn, T. & Schleif, R. (1984). Upstream repression and CRP stimulation of the *Escherichia coli* L-arabinose operon. *J. Mol. Biol.* **180**, 61–72.

Harrison, S. C. (1991). A structural taxonomy of DNA-binding domains. *Nature (London)*, **353**, 715–719.

Hendrickson, W. & Schleif, R. F. (1984). Regulation of the *Escherichia coli* L-arabinose operon studied by gel electrophoresis DNA binding assay. *J. Mol. Biol.* **178**, 611–628.

Kraft, R., Tardiff, J., Krauter, K. S., & Leinwand, L. A. (1988). Using mini-prep plasmid DNA for sequencing double stranded templates with Sequenase. *Biotechniques*, **6**, 544–546, 549.

Lin, L. & Little, J. W. (1989). Autodigestion and RecA-dependent cleavage of Ind-mutant LexA proteins. *J. Mol. Biol.* **210**, 439–452.

Little, J. W. & Hill, S. A. (1985). Deletions within a hinge region of a specific DNA-binding protein. *Proc. Nat. Acad. Sci., U.S.A.* **82**, 2301–2305.

Lobell, R. B. & Schleif, R. F. (1990). DNA looping and unlooping by AraC protein. *Science*, **250**, 528–532.

Marmorstein, R., Carey, M., Ptashne, M. & Harrison, S. (1992). Solution structure of the DNA-binding domain of Cd2-GAL4 from *S. cerevisiae*. *Nature (London)*, **356**, 408–414.

Miller, J. H. (1972) *Experiments in Molecular Genetics*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Oertel-Buchheit, P., Schmidt-Dorr, T., Granger-Schnarr, M. & Schnarr, M. (1993). Spacing requirements between LexA operator half-sites can be relaxed by fusing the LexA DNA binding domain with some alternative dimerization domains. *J. Mol. Biol.* **229**, 1–7.

Ogden, S., Haggerty, D., Stoner, C., Kolodrubetz, D. & Schleif, R. (1980). The *Escherichia coli* L-arabinose operon: binding sites of the regulatory proteins and a mechanism of positive and negative regulation. *Proc. Nat. Acad. Sci., U.S.A.* **77**, 3346–3350.

Reeder, T. C. (1993). Ph.D. thesis, Brandeis University, Waltham, MA.

Reeder, T. C. & Schleif, R. F. (1993). AraC protein can activate transcription from only one position and when pointed in only one direction. *J. Mol. Biol.* **231**, 205–218.

Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Sauer, R. T. (1978). DNA sequence of the bacteriophage lambda *cI* gene. *Nature (London)*, **276**, 301–302.

Sauer, R. T., Smith, D. L. & Johnson, A. D. (1988). Flexibility of the yeast alpha 2 repressor enables it to occupy the ends of its operator, leaving the center free. *Genes Develop.* **2**, 807–816.

Schleif, R. F. & Wensink, P. (1981). *Practical Methods in Molecular Biology*. Springer-Verlag, New York.

Sondek, J. & Shortle, D. (1990). Accommodation of single amino acid insertions by the native state of staphylococcal nuclease. *Proteins: Struct. Funct. Genet.* **7**, 299–305.

Starzyk, R. M., Burbaum, J. J. & Schimmel, P. (1989). Insertion of new sequences into the catalytic domain of an enzyme. *Biochemistry*, **28**, 8479–8484.

Wei, C. F., Alianell, G. A., Bencen, G. H. & Gray, H. B. (1983). Isolation and comparison of two molecular species of the *BAL 31* nuclease from *Alteromonas espejiana* with distinct kinetic properties. *J. Biol. Chem.* **258**, 13506–13512.

Wu, H. M. & Crothers, D. M. (1984). The locus of sequence-directed and protein-induced DNA bending. *Nature (London)*, **308**, 509–513.

Edited by P. von Hippel

(Received 5 April 1994; accepted 20 June 1994)