Biochemistry

 Volume 33, Number 36 September 13, 1994

Accelerated Publications

Homeodomain Determinants of Major Groove Recognition†

Joel L. Pomerantz‡,§ and Phillip A. Sharp*,‡

Center for Cancer Research and Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, and Harvard–MIT Division of Health Sciences and Technology, Cambridge, Massachusetts 02139

Received May 31, 1994; Revised Manuscript Received July 12, 1994®

ABSTRACT: The homeodomain is a highly conserved structural module that binds DNA and participates in protein–protein interactions. Most homeodomains contain residues at positions 47 and 51 which mediate recognition of a TAAT core binding sequence in the major groove. The constraints imposed on the identity of these residues by homeodomain structure and DNA docking have been examined in the context of the POU domain of the Oct-1 transcription factor. A bacterial library, in which POU homeodomain residues 47 and 51 have been randomized, was probed on nitrocellulose filters for the binding of DNA fragments containing the consensus octamer sequence. The residues which provide for the highest affinity interaction with the octamer consensus sequence, and the greatest specificity, are the highly conserved wild-type residues valine 47 and asparagine 51. Interestingly, a class of variants containing arginine at position 51 was also detected in the screen and found to have moderate affinity for the consensus sequence but reduced specificity compared to the wild-type protein. A single variant containing arginine at both positions 47 and 51 was detected when the library was probed with fragments containing nucleotide substitutions at positions expected to be contacted by residues 47 and 51. This variant was used to alter the DNA-binding specificity of a transcriptional regulatory complex which depends upon Oct-1 for DNA recognition. These findings suggest that homeodomain structure and DNA docking constrain the versatility of the domain in that only a limited set of amino acid determinants can endow the domain with specific, high-affinity DNA binding.

The homeodomain is a structural module that determines the specificity of action of a wide variety of transcription factors. This specificity is conferred by both its DNA-binding surface as well as by surfaces that are targets for protein–protein interactions with other transcriptional regulators. Structural analyses of homeodomain–DNA complexes have revealed a highly conserved structure and mode of docking DNA (Kissinger et al., 1990; Otting et al., 1990; Wolberger et al., 1991; Klemm et al., 1994). These studies have complemented extensive biochemical and genetic experiments addressing the interaction of the homeodomain with its conserved 5′-TAATNN-3′ binding sequence (Laughon, 1991). The homeodomain is composed of an N-terminal arm which makes contacts with bases (5′-TAATNN-3′) in the minor groove of DNA, and three α helices, the third of which makes base contacts in the major groove (5′-TAATNN-3′). For several homeodomains, specificity for nucleotides 3′ to the TAAT core is determined by residue 50 in helix 3 (Hanes & Brent, 1989, 1991; Treisman et al., 1989; Percival-Smith et al., 1990). Residues that contact the sugar–phosphate backbone are distributed throughout the domain.

The highly conserved nature of homeodomain–DNA interactions raises the issue of how different homeodomain

† This work was supported by U.S. Public Health Service Grant PO1-CA42063 from the National Institutes of Health, by cooperative agreement CDR-8803014 from the National Science Foundation to P.A.S., and partially by the National Cancer Institute Cancer Center Support (core) Grant P30-CA14051. J.L.P. was a Sterling Winthrop Research Fellow in Health Sciences and Technology.

* Corresponding author.
‡ Massachusetts Institute of Technology.
§ Harvard–MIT Division of Health Sciences and Technology.

® Abstract published in *Advance ACS Abstracts,* August 15, 1994.

0006-2960/94/0433-10851$04.50/0

FIGURE 1: The Oct-1 POU domain bound to DNA. Residues 47 and 51 in helix 3 of the POU homeodomain are shown in ball-and-stick representation interacting with adenine 7 (5′-ATGCAAAT-3′) and thymine 8 (5′-ATGCAAAT-3′) of the octamer sequence, respectively. This figure was generated by MOLSCRIPT (Kraulis, 1991) using the coordinates of Klemm et al. (1994).

proteins can exert specific regulatory effects, especially when functional specificity maps to the homeodomain itself (Hayashi & Scott, 1990). The observation that different homeodomains can determine dramatically different biological actions despite little or no difference in DNA-binding specificity has led to the description of mechanisms of specificity other than that provided by monomeric DNA binding. These involve homo- and heterodimerization of homeodomains and the cooperative interaction of the homeodomain with other regulatory proteins (Smith & Johnson, 1992; Pomerantz et al., 1992; Lai et al., 1992; Vershon & Johnson, 1993; Wilson, D., et al., 1993).

The limited variation in the specificity of homeodomains for a TAAT core binding site emphasizes the question of how the highly conserved tertiary structure of the domain constrains its DNA-binding specificity for this core sequence. Most of the amino acid determinants in the homeodomain that form specific contacts with bases in the TAAT core are highly conserved among homeodomains (Laughon, 1991). This raises the possibility that only a limited set of amino acids can occupy these positions and make specific base contacts in the context of homeodomain structure and DNA docking.

The POU homeodomain of the Oct-1 transcription factor participates in prototypical homeodomain–DNA interactions in that specific recognition in the major groove is mediated by residues asparagine (N) 51 and valine (V) 47 (Klemm et al., 1994). Oct-1 is a founding member of the POU domain subfamily of homeodomain transcription factors (Herr et al., 1988) which recognizes the 5′ half of the consensus octamer site (5′-ATGCAAAT-3′) with the POU-specific domain and the 3′ half (5′-ATGCAAAT-3′) with the POU-homeodomain (Figure 1). In the DNA–protein complex, the adenine at position 7 (5′-ATGCAAAT-3′) is contacted by POU homeodomain residue N51. This interaction corresponds to that observed in the engrailed–DNA (Kissinger et al., 1990) and MATα2–DNA (Wolberger et al., 1991) complexes, in which the asparagine accepts a hydrogen bond from the N6 of adenine and donates a hydrogen bond to the N7. Asparagine 51 is one of the most highly conserved residues in all homeodomains, and its interaction with the core adenine is believed to be a signature feature of homeodomain–DNA binding. The thymine at position 8 (5′-ATGCAAAT-3′) interacts with V47 via a van der Waals contact involving the thymine methyl group (Klemm et al., 1994). Homeodomain residue 47 is most often either an isoleucine or a valine, and in the engrailed–DNA cocrystal structure, isoleucine 47 contacts the counterpart thymine (5′-TAAT-3′) via an analagous interaction. Thus, Oct-1 POU homeodomain residues V47 and N51 represent determinants which are used by most homeodomains for base-specific contacts to the TAAT core sequence in the major groove of DNA.

In order to examine how the structural context of homeodomain–DNA binding constrains the nature of amino acids involved in specific interactions with the TAAT core sequence, an Oct-1 POU domain bacterial expression library was generated in which residues 47 and 51 in the POU homeodomain were randomized. This library was prepared so that the DNA binding of a large number of POU domain variants could be screened by the direct binding of radioactive DNA probes on nitrocellulose filters. This library has been used to investigate what residues at positions 47 and 51 are consistent with binding to the octamer consensus site. In addition, the library has been used to detect Oct-1 POU domain variants which have novel DNA-binding specificities, and which can alter the binding specificity of a transcriptional regulatory complex which depends upon Oct-1 for DNA recognition.

MATERIALS AND METHODS

Construction and Screening of the Bacterial PA-Oct-1 POU Domain Expression Library. A protein A–Oct-1 POU domain expression library was constructed by a strategy we refer to as "deletion insertion randomization" (DIR). DIR is useful when restriction sites flanking a region desired to be randomized are not available and cannot be engineered silently. Oligonucleotide-mediated mutagenesis (Sayers et al., 1992) is employed using a population of oligonucleotides which has been synthesized to randomize codons. In order to avoid selective hybridization of the oligonucleotides that contain bases which are complementary to the wild-type ssDNA substrate, a previous round of mutagenesis is performed that deletes the region of desired randomization. The construct containing the deletion is used to prepare a ssDNA substrate for the insertion randomization mutagenesis. A fragment encoding a portion of the Oct-1 POU domain [amino acids 345–441 (Sturm et al., 1988)] was cloned into the vector pBS+ (Stratagene). ssDNA was prepared using VCSM13 helper phage (Stratagene) and XL1-Blue host strain according to the manufacturer's protocol. This ssDNA was the substrate for mutagenesis performed using the oligonucleotide-directed mutagenesis system version 2 (Amersham) according to the manufacturer's instructions. Deletion mutagenesis, designed to delete the region of DNA encoding POU homeodomain residues 45–53 [homeodomain numbering scheme of Qian et al. (1989)] in helix 3, was accomplished using the oligonucleotide 5′-CAATATGGAAAAAGAGGTGCAGAAA-GAAAAAAGAATCC-3′. From the resulting construct, Δ45–53, ssDNA was generated and used as a template for insertional mutagenesis designed to replace the deleted region with the codons for homeodomain residues 47 and 51 randomized as NNG/C. The mutagenic oligonucleotide used for this insertion was 5′-ATATGGAAAAAGAGGTGAT-TCGTNNG/CTGGTTCTGTNNG/CCGCCGCCAGAAA-GAAAAAAGAATCAACC-3′. The products of the insertional mutagenesis reaction were used for transformation of the XL1-Blue strain. A total of 9500 transformed colonies were scraped from LB-agarose ampicillin (100 μg/mL) plates and pooled, and plasmids were prepared by standard protocols.

The fragment encoding the POU domain was excised from the plasmid pool and ligated into vector pRIT2T (Pharmacia) so as to generate an in-frame fusion of the full POU domain [residues 270–411 (Sturm et al., 1988)] with the protein A gene product of *Staphylococcus aureus*. The ligation products were used to transform *Escherichia coli* strain N4830 which was plated on LB-agar ampicillin plates for library screening. A random distribution of nucleotides at homeodomain codons 47 and 51 was confirmed by dideoxy sequencing of a number of individual expression plasmids and by sequencing a sample of the total pooled expression library.

After plating, colonies were allowed to grow for 24 h at 30 °C before transfer to nitrocellulose filters (132 mm, BA85/23 Schleicher and Schuell). After marking the position of filters on plates, the filters were lifted and incubated colony side up on a new set of plates at 42 °C for 2 h to induce fusion protein expression. Colony lysis was achieved by exposure to chloroform vapor for 10 min and then immersion in lysis buffer [100 mM Tris-HCl (pH 7.8), 150 mM NaCl, 5 mM $MgCl_2$, 1.5% bovine serum albumin Fraction V (heat shock), 400 μg/mL lysozyme] (25 mL/filter) for 10 min at 25 °C. After air-drying, filters were then processed through a denaturation/renaturation cycle. Filters were gently shaken in buffer J [25 mM Hepes (pH 7.9), 25 mM NaCl, 5 mM $MgCl_2$, 0.5 mM DTT] plus 6 M guanidine hydrochloride for 5 min at 4 °C. The guanidine hydrochloride concentration was diluted 2-fold in each of 5 steps with the removal of half of the buffer and the addition of an equal volume of buffer J, after which filters were washed twice with buffer J. Each step was incubated for 5 min at 4 °C. Filters were then incubated in blocking buffer [50 mM Tris-HCl (pH 7.9), 50 mM NaCl, 1 mM EDTA, 1 mM DTT, 5% BSA Fraction V] (50 mL/filter), with gentle shaking for 1 h at 25 °C, and then rinsed twice, 5 min each, with binding buffer [20 mM Hepes (pH 7.9), 50 mM KCl, 1 mM EDTA, 0.7 mM DTT, 0.025% NP40]. During the lysis, denaturation/renaturation, and blocking steps, each filter was placed in a separate Petri dish for the incubation. Filters were then gently shaken in binding buffer plus 5 μg/mL denatured sonicated salmon sperm DNA and ^{32}P-labeled probe at a final 1–2 × 10^6 cpm/mL (~10^{-10} M) for 1 h at 25 °C. Filters were then washed twice for a total of 15 min with binding buffer (500 mL for up to 4 filters), blotted dry, and then exposed to Kodak X-OMAT AR film at −70 °C with an intensifying screen for 12–24 h. Positive colonies were picked, patched onto new plates, and rescreened. Plasmids were isolated from colonies which rescreened positive and sequenced by dideoxy sequencing to determine the residues encoded at positions 47 and 51 of the homeodomain. Probes used for screening were derived from cloning the fragment

5'-GATCCTATGCAANNGACC-3'
3'-GGATACGTTNNCTGGAGCT-5'

into the *Xho*I and *Bam*HI sites of pBSKII+ (Stratagene). All 16 variants were obtained, verified by dideoxy sequencing, and excised for use as probes by digesting with *Xba*I and Asp718. Fragments were labeled by the large (Klenow) fragment of *E. coli* DNA Pol I in the presence of dGTP, dCTP, dTTP, and [α-^{32}P]dATP, and then gel-purified on nondenaturing polyacrylamide gels.

A total of 5000 colonies were screened with the probe containing the octamer consensus sequence. For screening with the pools of mutant probes, 4000 colonies were screened by each probe pool, with each of the four probes in the pool at an equal concentration of 1–2 × 10^6 cpm/mL (~10^{-10} M).

Expression of Fusion Proteins. Selected variants were expressed in *E. coli* N4830 strain and were purified by affinity chromatography on IgG–Sepharose as described previously (Kristie & Sharp, 1990). The concentration of each PA fusion protein was determined by densitometric analysis of Coomassie-stained SDS–PAGE-resolved proteins using bovine serum albumin (Boehringer Mannheim) as standard.

Electrophoretic Mobility Shift Assays. Relative affinities of variants were determined by electrophoretic mobility shift assays using the same probes as those used for colony screening. DNA–protein-binding reactions contained 3–30 pg of DNA probe, 10 ng of poly[d(I-C)]/poly[d(I-C)], 10 mM Hepes (pH 7.9), 0.5 mM EDTA, 50 mM KCl, 0.75 mM DTT, 4% Ficoll-400, 300 μg/mL of bovine serum albumin, and 1–8000 pg of PA-Oct-1 POU domain variant in a total volume of 10 μL. The concentration of DNA probe was always at least 1 order of magnitude below the apparent dissociation constant. Reactions were incubated at 30 °C for 30 min and resolved in 4% nondenaturing polyacrylamide gels (Pomerantz et al., 1992). The protein–DNA and free DNA complexes were quantitated using a Molecular Dynamics Phosphorimager with ImageQuant 3.15 and 3.22 software. Apparent dissociation constants were determined as the inverse of the slope of the line derived from plotting fraction of probe bound/fraction of probe unbound vs total PA-POU domain protein concentration. In all cases the lines consisted of at least four points. C1 complex formation assays were performed using the HSVα0 probe (HSV α/IE element: 5'-GTGCATGCTAAT GATATTCTTTGGGG-3') (Kristie et al., 1989) or a mutated version (HSV α/IE "GG" element: 5'-GTGCATGCTA**GG**-GATATTCTTTGGGG-3') that was generated by the oligonucleotide-directed mutagenesis system version 2 (Amersham) according to the manufacturer's instructions. DNA–protein-binding reactions were performed as described above using 0.4–0.8 ng of DNA probe, 300 ng of poly[d(I-C)]/poly[d(I-C)], and, where indicated, 15 ng of PA-αTIF and 1 μL of a chromatographic fraction containing the HeLa cell C1 factor (Pomerantz et al., 1992).

Random Binding Site Selection. The probe used for random binding site selection was generated by annealing the following two oligonucleotides and polymerizing with Klenow in the presence of dGTP, dCTP, dTTP, and [α-^{32}P]dATP: primer R: 5'-GGCTGAGTCTGAACGGATCCN$_{13}$CCTCGAGA-CTGAGCGTCG-3'; primer A: 5'-CGACGCTCAGTCTC-GAGG-3'. For the first round of selection 50 pg of PA-POU domain variant was incubated with 5 ng of probe in DNA–protein-binding reactions under the conditions described above in the absence of any poly[d(I-C)]/poly[d(I-C)]. In each round, reactions were electrophoresed as described above, gels were dried and exposed to film, and the DNA–protein complexes were excised from the dried gel for elution and PCR amplification of bound fragments (Blackwell & Weintraub, 1990) using primer A above and primer B: 5'-GGCTGAGTCTGAACGGATCC-3'. A contamination control was processed in parallel starting from a gel slice containing no protein–DNA complex. In all cases this control did not produce any detectable PCR products. Approximately 1 ng of amplified product was used in the binding reaction of the next round of selection. For the second and third rounds, 50 pg of PA-POU domain variant was used; 10 pg was used in the fourth round. The amplified products of the fourth round of selection were digested with *Bam*HI and *Xho*I and ligated into the vector pBSKII+. Plasmids were derived from transformants and sequenced by dideoxy sequencing.

RESULTS

A bacterial expression library was generated in which POU homeodomain residues 47 and 51 were simultaneously

Table 1: Results of Screening with 5′-ATGCAAATGA-3′

47	51	no. of isolates[a]	no. predicted[b]	isol/pred	affinity[c]
V[d]	N[d]	8	10	0.80	1
R	N	5	15	0.33	3.0
T	N	2	10	0.20	nd
N	N	1	5	0.20	2.8
I	N	1	5	0.20	nd
C	N	2	5	0.40	nd
G	N	1	10	0.10	6.5
H	N	1	5	0.20	4.7
R	R	13	44	0.30	32
G	R	7	30	0.23	23
S	R	2	44	0.05	nd
N	R	1	15	0.07	nd
A	R	2	30	0.07	nd
C	R	1	15	0.07	nd
H	R	1	15	0.07	nd
T	R	2	20	0.10	nd
V	Q	2	10	0.20	1100

[a] ~5000 colonies screened. [b] The expected frequency of that variant in 5000 colonies of the library. Codons 47 and 51 were randomized as NNG/C; 32 possible codons at each position, total complexity of library is 1024. [c] Relative dissociation constant, normalized to that of the wild-type protein. [d] Residue found in wild-type protein.

randomized. In this library, the Oct-1 POU domain was fused to protein A (PA) of *S. aureus* to facilitate the expression, purification, and analysis of individual variants. The library was screened for the binding of radioactive DNA probes using conditions similar to those originally described for the screening of cDNA expression libraries by Singh et al. (1988) and Vinson et al. (1988) (see Materials and Methods). A similar procedure has been described by Lorimer et al. (1992).

The nature of residues at positions 47 and 51 which can participate in the specific recognition of an octamer site was determined by probing the library with a DNA fragment containing the octamer consensus sequence (5′-ATGCAAATGA-3′). As shown in Table 1, the combination of residues that was most efficiently detected in the screen was the combination found in the wild-type protein, valine at position 47 and asparagine at position 51. All of the variants detected can be segregated into classes according to the residue at position 51. Surprisingly, in addition to a class containing asparagine at position 51, a class containing arginine at this position was also detected, as well as a single variant containing glutamine at position 51 and valine at position 47.

The relative effects of amino acid substitutions on binding affinity were determined by comparison of a representative panel of variants to the wild-type protein (Table 1). Of those analyzed, the wild-type protein exhibited the highest affinity for the fragment. Variants containing only substitutions of V47 had minor reductions in affinity, the largest a 6.5-fold reduction for the G47 N51 variant. The greatest reduction in binding affinity was observed for the V47 Q51 variant, which had a 1100-fold reduction in affinity. Two representatives of the R51 class of variants, G47 R51 and R47 R51, had intermediate affinities with reductions of 23- and 32-fold, respectively, compared to the wild-type protein.

The large variability of residues at position 47 detected in the screen suggested only a moderate contribution of V47 to the binding affinity and specificity of the POU domain. The minimal effects on affinity observed for its substitution with chemically diverse side chains were consistent with this notion. On the other hand, the greater apparent selectivity at position 51 suggested a more critical role for N51, which was supported by the 1100-fold reduction in binding affinity upon substitution with the chemically similar glutamine. The intermediate affinity of variants containing arginine at 51 suggested that their specificity might be quite different from that of the wild-type protein. Therefore, the specificity of variants with substitutions at 51 for the adenine residue (adenine 7, Figure 1) contacted by N51 of the wild-type protein (5′-ATGCAAATGA-3′) was examined.

Table 2: Specificity of Variants for Nucleotide 7

	V47 N51[a]	V47 Q51[a]	R47 R51[a]	G47 R51[a]
5′-ATGCAAATGA-3′	1	1100	32	23
5′-ATGCAAGTGA-3′	300	54000	44	69
5′-ATGCAACTGA-3′	45000	96000	610	500
5′-ATGCAATTGA-3′	1100	83000	650	700

[a] Relative dissociation constant, normalized to that of the wild-type protein for the octamer consensus sequence.

The V47 Q51, R47 R51, and G47 R51 variants were compared to the wild-type protein (V47 N51) for binding to fragments containing nucleotide substitutions of adenine 7 (Table 2). All variants exhibited a preference for adenine at nucleotide 7; however, the wild-type protein had a much higher specificity for this residue. Specifically, the substitution of adenine with guanine, thymine, and cytosine resulted in reductions in affinity of 300-, 1100-, and 45000-fold, respectively, for the wild-type protein. In comparison, for the V47 Q51 variant, these substitutions reduced affinity by 49-, 75-, and 87-fold, respectively, while for the R47 R51 variant, reductions of 1.4-, 20-, and 19-fold, respectively, were observed. Thus, asparagine at position 51 is not only required for the highest affinity interaction with the octamer site, it also determines the greatest selectivity for adenine at nucleotide position 7.

The randomized expression library was also used to explore what other combinations of homeodomain residues at 47 and 51 and nucleotides at positions 7 and 8 could provide high-affinity homeodomain–DNA interactions. The library was probed with fragments of DNA containing nucleotide substitutions at these positions. Since much of the specificity of binding was determined by interactions mediated by nucleotide 7, pools of probes were grouped according to the identity of this residue. For example, the "CN" pool consisted of probes 5′-ATGCAACAGA-3′, 5′-ATGCAACTGA-3′, 5′-ATGCAACCGA-3′, and 5′-ATGCAACGGA-3′. Interestingly, when the library was probed with either the "CN" pool or the "TN" pool, no positive colonies were observed. This suggested that neither cytosine nor thymine at position 7 could mediate an interaction that could contribute enough affinity for detection by this screen. When the library was probed with the "GN" pool, a single variant, R47 R51, was detectable at a low frequency [4 positives, 35 predicted (4000 screened)]. This variant was characterized further.

It was possible that the arginine substitutions at positions 47 and 51 in the R47 R51 variant mediated novel interactions with nucleotides other than those at positions 7 and 8. This variant had much less selectivity for nucleotide 7 than the wild-type protein (Table 2) and yet was detected in both the "GN" pool screen and in the initial screen with the unsubstituted octamer consensus sequence. To directly compare its nucleotide preferences at all positions in the binding site with that of the wild-type protein, a random binding site selection assay was employed (Blackwell & Weintraub, 1990; Pollock & Triesman, 1990). The wild-type protein (V47 N51) and the R47 R51 variant were challenged in four rounds of binding site selection, along with the R47 N51 variant as a control for the effects of the substitution of V47 with arginine. At least 25 sequences were determined for each protein from the pool of sites that were selected in the fourth round (Figure

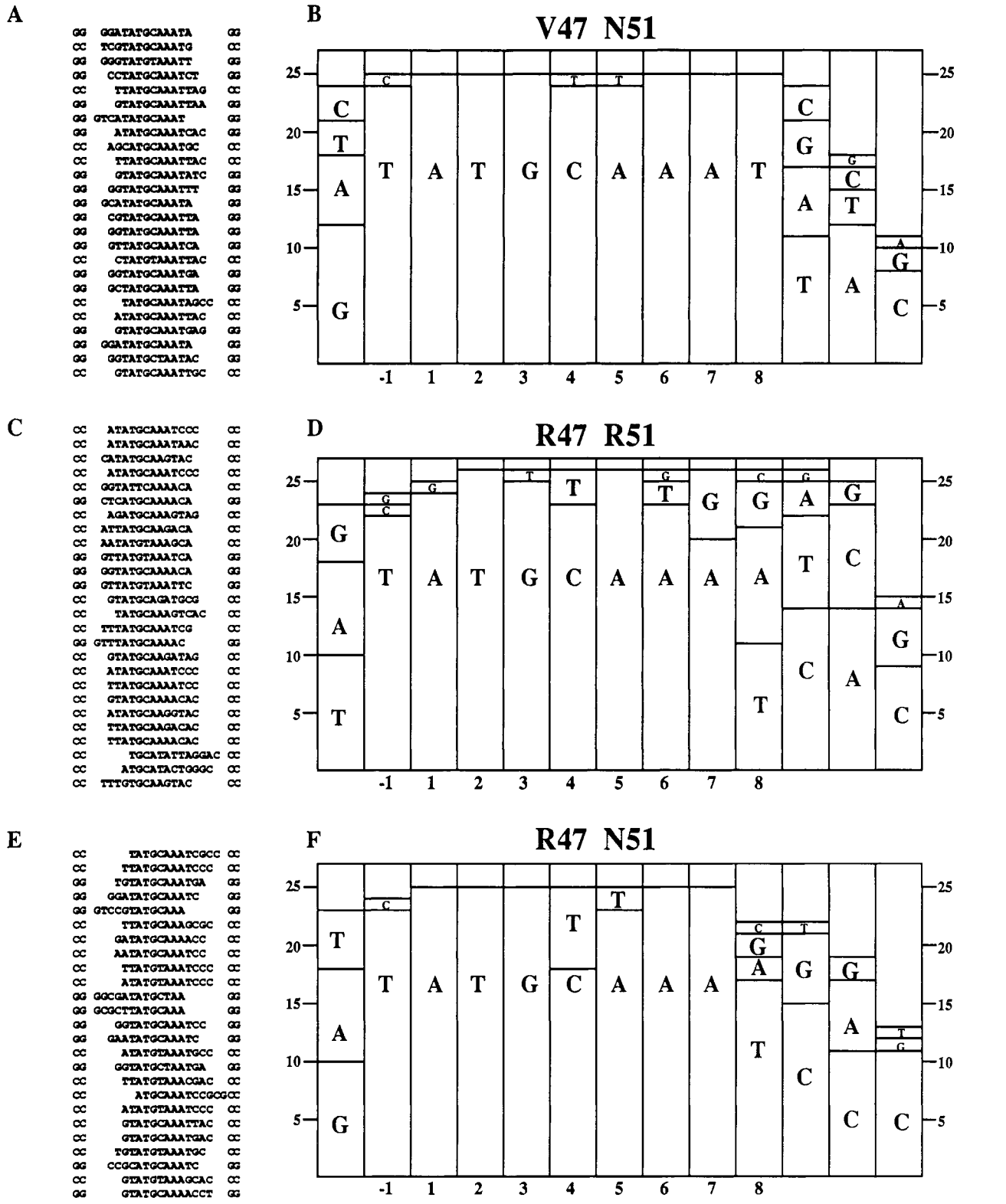

FIGURE 2: Binding site sequences selected by the wild-type Oct-1 POU domain (V47 N51) and the R47 R51 and R47 N51 variants. (A, C, E) Sequences of sites isolated after four rounds of selection. (B, D, F) Consensograms (Wilson, D., et al., 1993) derived from the selected sequences.

2A,C,E). Consensograms (Wilson, D., et al., 1993) were derived from these sequences (Figure 2B,D,F). As shown previously (Verrijzer et al., 1992), the wild-type POU domain selected a consensus octamer site (5′-TATGCAAAT-3′) with strict preferences for nucleotides at positions −1 through 8 (Figure 2A,B). In contrast, the R47 R51 variant selected sites with reduced stringency, especially at positions 7 and 8 (Figure 2C,D). Although this variant prefers adenine at position 7, this preference is not as strong as it is for the wild-type protein: guanosine was selected at that position 6/26 times for R47 R51 and 0/25 for V47 N51. Thus, substitution of N51 with arginine results in a change in the "adenine requirement" at position 7 to a "purine requirement". The R47 R51 variant also exhibits much less stringency of selection at position 8 as compared to the wild-type. Adenine and thymine were selected equally by the variant with less preference for guanosine and cytosine while only thymine was selected by the wild-type protein. The effect at position 8 is

Table 3: Relative Specificity of V47 N51 and R47 R51

	V47 N51[a]	R47 R51[a]	VN/RR[b]
5′-ATGCAAATGA-3′	1	32	0.031
GT	300	44	6.8
CT	45000	610	74
TT	1100	650	1.7
AC	5.3	120	0.04
AA	7.1	79	0.09
AG	16	130	0.12
GC	37000	1000	37
GA	3900	100	39
GG	39000	150	260
CC	29000	130	220
CA	39000	810	48
CG	75000	710	110
TC	1700	1100	1.5
TA	5300	980	5.4
TG	4200	1200	3.5

[a] Relative dissociation constant, normalized to that of the wild-type protein binding to the octamer consensus. [b] The dissociation constant for the V47 N51 protein divided by that for the R47 R51 variant.

partially attributable to the substitution of V47 with arginine (Figure 2E,F). It appears, from the sequences selected, that the R47 R51 variant did not select any nucleotide at any position with greater stringency than did the wild-type protein. This suggests that arginine, either at position 47 or 51, does not make unique nucleotide-specific contacts at any position in the selected site.

The relaxed specificity of the R47 R51 variant suggested its potential to redirect the formation of Oct-1-dependent transcriptional regulatory complexes to novel sequences that are not efficiently recognized by the wild-type protein. We sought to test this possibility using as a model system the formation of the multiprotein C1 complex on the herpes simplex virus (HSV) α or immediate-early (α/IE) enhancer element. Formation of the C1 complex is dependent on the binding of the element by the Oct-1 POU domain and the viral αTIF protein (VP16, Vmw65, ICP25) (McKnight et al., 1987; Gerster & Roeder, 1988; O'Hare & Goding, 1988; Preston et al., 1988; Kristie et al., 1989; Stern et al., 1989), on the specific recognition of the Oct-1 POU homeodomain surface by αTIF (Pomerantz et al., 1992; Lai et al., 1992), and on the presence of the cellular C1 factor (HCF) (Kristie & Sharp, 1993; Wilson, A. C., et al., 1993). Some nucleotide substitutions in the octamer-related sequence in this element should impair binding of the wild-type Oct-1 protein while having only a minimal effect on the binding of the R47 R51 variant. Formation of the regulatory complex on this novel element would then be dependent on the R47 R51 variant, and not possible with the wild-type protein. The combination of nucleotide substitutions at positions 7 and 8 which would provide the greatest discrimination in binding between the wild-type protein and the R47 R51 variant was determined. The R47 R51 variant was directly compared to the wild-type protein for binding to 16 probes containing all combinations of nucleotides at positions 7 and 8. As shown in Table 3, the sequence with the greatest preference (260-fold) for the variant contained guanosine at both positions 7 and 8 (5′ATG-CAAGGGA-3′).

The double guanosine substitution was incorporated into the HSV α/IE element [5′-ATGCTAGGGATATTCTTTGG-3′ (HSV α/IE "GG")] and tested for the formation of the C1 complex in the presence of αTIF, the C1 factor, and either the wild-type or variant Oct-1 POU domain. The R47 R51 POU domain variant was first compared to the wild-type protein for formation of the C1 complex on the unsubstituted

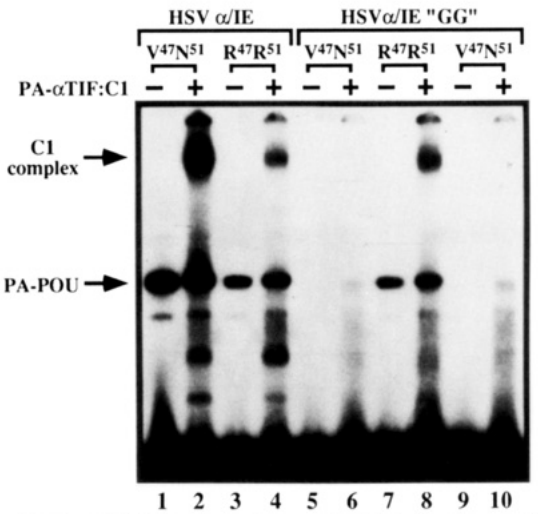

FIGURE 3: The R47 R51 variant changes the DNA-binding specificity of an Oct-1-dependent regulatory complex. Wild-type (V47 N51, 25 pg in lanes 1, 2, 5, and 6, 350 pg in lanes 9 and 10) and variant (R47 R51, 350 pg) PA-Oct-1 POU domain fusion proteins were incubated in DNA–protein-binding reactions in the absence (–) or presence (+) of 15 ng of PA-αTIF and 1 μL of a chromatographic fraction containing the HeLa cell C1 factor as indicated. Reactions included either the unsubstituted HSV α/IE element (5′-ATGCTAAT-GATATTCTTTGG-3′) (lanes 1–4) or the double guanosine substituted HSV α/IE "GG" element (5′-ATGCTAGGGATAT-TCTTTGG-3′) (lanes 5–10) as probe. The positions of the multiprotein C1 and POU domain–DNA complexes are indicated with arrows.

HSV α/IE element (Figure 3). The R47 R51 variant was clearly capable of forming the complex (cf. lanes 2 and 4), although 14-fold more protein was required to attain similar levels of DNA binding. When normalized on the basis of DNA binding activity, the variant was found to have only a 2.4-fold lower ability to form the complex than the wild-type protein on the unsubstituted element. When the HSV α/IE "GG" element was used, no binding of the wild-type POU domain was detectable even at concentrations 14-fold higher than that required to bind the unsubstituted element (lanes 5 and 9), and upon addition of αTIF and the C1 factor, no complex formation was evident (lanes 6 and 10). In contrast, the R47 R51 variant bound the HSV α/IE "GG" element to an extent similar to that for the unsubstituted element (lane 7) and was capable of efficiently forming the C1 complex with αTIF and the C1 factor (lane 8). Therefore, the R47 R51 variant can be used to alter the DNA-binding specificity of a transcriptional regulatory complex which depends upon Oct-1 for DNA recognition.

DISCUSSION

Transcription factors utilize structural modules to recognize specific DNA sequences (Pabo & Sauer, 1992). Some modules, such as the zinc finger originally discovered in TFIIIA (Miller et al., 1985), accommodate different sets of amino acid determinants so that DNA-binding specificity can vary within the framework of a conserved domain structure (Pavletich & Pabo, 1991, 1993; Fairall et al., 1993). Other modules appear to be specialized for the recognition of particular sequences, such as those of the basic–helix–loop–helix (CANNTG) (Murre & Baltimore, 1992) and nuclear hormone receptor [AGNNCA (half-site)] (Evans, 1988) families. The homeodomain has apparently evolved with a constrained specificity for a binding site with a TAAT core. The Oct-1 POU homeodomain has been used to test the

stringency of determinants that mediate recognition of the TAAT core in the major groove.

Among all possible interactions determined by Oct-1 POU homeodomain residues 47 and 51 and the nucleotides at position 7 and 8 in the octamer binding sequence (5′-ATGCAAAT-3′), the N51–adenine 7 and V47–thymine 8 interactions provide the highest affinity and greatest degree of specificity. The structure of the homeodomain and its mode of docking against DNA probably impose greater constraints on the amino acid 51–nucleotide 7 interaction than on the amino acid 47–nucleotide 8 interaction. Fewer residues were detected in the screens at position 51 than at 47, and substitution of either N51 or adenine 7 had a larger effect on affinity and specificity than did substitution of V47 or thymine 8. The fact that all variants analyzed had higher affinities for sites containing adenine at position 7 than those containing other bases at that position suggests that adenine at the third position of the homeodomain subsite (AAAT) is most compatible with the spatial architecture of the homeodomain even when residue 51 is not the highly conserved asparagine. In accord with these results, Botfield et al. (1994) have found that, in the homologous Oct-2 POU homeodomain, N51 provides the highest affinity interaction with the wild-type octamer sequence when compared to 19 substituted variants.

The ability to achieve reasonable affinity ($\sim 10^{-9}$ M) for DNA with arginine at position 51 was surprising, given its lack of chemical similarity to asparagine. In addition, arginine is not found in this position in any natural homeodomain. Arginine at 51 is probably not making a nucleotide-specific contact, and it is possible that this residue contributes to binding affinity by ionic interaction with the sugar–phosphate backbone. Alternatively, the purine requirement at position 7 that is observed for the R47 R51 variant may reflect an interaction of arginine with the N7 of the base. A similar specificity, determined by an arginine positioned in the major groove by an α helix, has been invoked based upon the Hin recombinase–DNA recombination half-site crystal structure (Feng et al., 1994). The relaxed sequence specificity of the R47 R51 variant allows it to nucleate the formation of a transcriptional regulatory complex on a DNA sequence element which is not efficiently bound by wild-type Oct-1. Since formation of the C1 complex is critically dependent on the presentation of residues on the surface of helices 1 and 2 (Pomerantz et al., 1992; Lai et al., 1992), we conclude that substitution of residues 47 and 51 with arginine does not drastically perturb homeodomain structure.

Since the R47 R51 variant binds with reasonable affinity and can participate in the formation of regulatory complexes, it is reasonable to question the absence of this variant set of amino acids in the known sequences of homeodomains. Both the reduced affinity and specificity of the variant, relative to the wild-type protein, may preclude its utility in vivo. The R47 R51 variant binds the octamer sequence with 30-fold lower affinity than the wild-type protein. In addition, its reduced specificity allows it to recognize many more sequence variants at a given protein concentration than the greater specificity of the wild-type protein would allow. The partitioning of the variant between its "specific" and "nonspecific" sites may make the occupancy of an individual target sequence without inappropriate action at other sites impossible at physiological levels of expression of the protein.

Other studies have altered the binding specificity of homeodomains by substitution of residue 50 since this residue is an important determinant of specificity for nucleotides 3′ to the TAAT core (Hanes & Brent, 1989, 1991; Treisman et al., 1989; Percival-Smith et al., 1990). However, for POU domain proteins the residue at this position (cysteine) does not confer sequence selectivity and can be substituted without effect (Ingraham et al., 1990; Verrijzer et al., 1992). Consistent with this is that the POU domain does not select nucleotides 3′ to the octamer site with a high degree of preference. The 3′ boundary of sequence recognition by POU homeodomains appears to be much more critically dependent on residues 47 and 51. It is presently unclear why cysteine 50 is absolutely conserved among POU domain proteins.

The interactions mediated by residues 47 and 51 do not appear to be independent. For example, the effect of the substitution of N51 with arginine is dependent on what residue is at position 47. This substitution reduces the affinity for the octamer consensus sequence by 11-fold when residue 47 is arginine but only by 3.5-fold when residue 47 is glycine (Table 1). In addition, the residue at position 51 also influences which nucleotide is selected at position 8 when residue 47 is arginine. The R47 N51 variant has less selectivity for thymine at position 8 than does the wild-type protein, and this selectivity is further reduced for the R47 R51 variant (Figure 2). We also note the lack of detection of a variant containing valine at position 47 and arginine at position 51. The frequency of this variant in the library is predicted (29/5000) to be larger than most of the variants that were detected in the screen performed with the unsubstituted octamer sequence as probe. This implies that this variant should have been detected if its affinity (off-rate) for the octamer sequence fell within the large range of affinities exhibited by those variants that were detected. The combination of valine at position 47 and arginine at position 51 may somehow be incompatible with homeodomain structure or DNA binding.

The interaction between residues 47 and 51 is probably related to the steric constraints that are imposed upon them as they pack into the major groove. The advantage of having wild-type residues V47 and N51 is not only due to their individual potentials for base-specific contacts but also to their ability to be sterically accommodated in the protein–DNA complex. It is striking that the V47 Q51 variant, which has the identical functional groups for nucleotide recognition and which differs from the wild-type only by a methylene group, has a 1100-fold reduced affinity for the octamer consensus. The substitution of either residue 47 or 51 with the much larger arginine may introduce steric interactions that reduce the density of packing of the side chains of helix 3 into the major groove and, therefore, alter the complementarity of the protein surface to the DNA. This is evident in the consensograms of Figure 2. The wild-type consensogram presents a profile of selected nucleotides with nearly absolute preferences and clear boundaries on either side of the binding site. In contrast, the consensograms of the R47 N51 and R47 R51 variants exhibit a breakdown of binding site stringency. The arginine substitutions affect the selection of nucleotides at positions other than 7 and 8, which may reflect the inability to pack the arginine side chain in a way which does not perturb other interactions. For example, even the nucleotide selected at position 4, which is within the POU-specific domain subsite, is influenced by the substitution of 47 and 51. This implies that the packing of residues in helix 3 also impinges upon the binding specificity of the POU-specific domain and reveals an interdependence of the POU-specific and POU-homeo subdomains that must arise from the spacing of subsites that is strictly preferred by the POU domain. Although there are

no protein–protein interactions observed between the subdomains in the Oct-1 POU domain–DNA complex, and the linker between them is disordered, changing the spacing of subsites reduces the affinity of the POU domain 10–100-fold (Klemm et al., 1994).

The surprising detection of R51 variants attests to the utility of the randomization/screening approach. This technique relies upon the detection of direct binding to radiolabeled DNA probes and allows for the screening of a large number of proteins without competition between variants for binding to the probe. Such an approach should complement others (Youderian et al., 1983; Rebar & Pabo, 1994) for the isolation of transcription factors with new DNA-binding specificities that will provide useful tools and advance the understanding of protein–DNA interactions.

We conclude that it is probably not possible to alter homeodomain DNA-binding specificity by the substitution of residues 47 and 51 without sacrificing affinity and the ability to discriminate between sites. The homeodomain emerges as a module which is limited in its ability to recognize different DNA sequences and which has evolved the capacity to participate in other mechanisms of regulatory specificity such as protein–protein interactions. The constraints on homeodomain residues 47 and 51 emphasize that, in protein–DNA recognition, the potential interactions that a particular amino acid residue may specify are critically dependent on the structural context determined by the folding and DNA docking of the module in which it is found. Therefore, the spatial architecture of a DNA-binding domain may greatly constrain which amino acid residues can serve as its determinants of DNA recognition.

ACKNOWLEDGMENT

We thank Juli Klemm and Carl Pabo for the coordinates of the Oct-1–DNA complex and for provocative discussions; Carl Pabo, Andrew MacMillan, Lee Lim, Juli Klemm, John Crispino, and Dan Chasman for critical reading of the manuscript; Tom Kristie and members of the Sharp lab for their continual support; M. Siafaca for her ever present assistance; and R. Issner and Y. Qiu for indispensable technical assistance.

REFERENCES

Blackwell, T. K., & Weintraub, H. (1990) *Science 250*, 1104–1110.
Botfield, M. C., Jancso, A., & Weiss, M. A. (1994) *Biochemistry 33*, 6177–6185.
Evans, R. M. (1988) *Science 240*, 889–895.
Fairall, L., Schwabe, J. W. R., Chapman, L., Finch, J. T., & Rhodes, D. (1993) *Nature 366*, 483–487.
Feng, J.-A., Johnson, R. C., & Dickerson, R. E. (1994) *Science 263*, 348–355.
Gerster, T., & Roeder, R. G. (1988) *Proc. Natl. Acad. Sci. U.S.A. 85*, 6347–6351.
Hanes, S. D., & Brent, R. (1989) *Cell 57*, 1275–1283.
Hanes, S. D., & Brent, R. (1991) *Science 251*, 426–430.
Hayashi, S., & Scott, M. P. (1990) *Cell 63*, 883–894.
Herr, W., Sturm, R. A., Clerc, R. G., Corcoran, L. M., Baltimore, D., Sharp, P. A., Ingraham, H. A., Rosenfeld, M. G., Finney, M., Ruvkin, G., & Horvitz, H. R. (1988) *Genes Dev. 2*, 1513–1516.
Ingraham, H. A., Flynn, S. E., Voss, J. W., Albert, V. R., Kapiloff, M. S., Wilson, L., & Rosenfeld, M. G. (1990) *Cell 61*, 1021–1033.
Kissinger, C. R., Liu, B., Martin-Blanco, E., Kornberg, T. B., & Pabo, C. O. (1990) *Cell 63*, 579–590.
Klemm, J. D., Rould, M. A., Aurora, R., Herr, W., & Pabo, C. O. (1994) *Cell 77*, 21–32.
Kraulis, P. J. (1991) *J. Appl. Crystallogr. 24*, 946–950.
Kristie, T. M., & Sharp, P. A. (1990) *Genes Dev. 4*, 2383–2396.
Kristie, T. M., & Sharp, P. A. (1993) *J. Biol. Chem. 268*, 6525–6534.
Kristie, T. M., LeBowitz, J. H., & Sharp, P. A. (1989) *EMBO J. 8*, 4229–4238.
Lai, J.-S., Cleary, M. A., & Herr, W. (1992) *Genes Dev. 6*, 2058–2065.
Laughon, A. (1991) *Biochemistry 30*, 11357–11367.
Lorimer, I. A. J., Ho, C.-Y., & Smith, M. (1992) *BioTechniques 12*, 536–543.
McKnight, J. L. C., Kristie, T. M., & Roizman, B. (1987) *Proc. Natl. Acad. Sci. U.S.A. 84*, 7061–7065.
Miller, J., McLachlan, A. D., & Klug, A. (1985) *EMBO J. 4*, 1609–1614.
Murre, C., & Baltimore, D. (1992) in *Transcriptional Regulation* (McKnight, S. L., & Yamamoto, K. R., Eds.) pp 861–879, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
O'Hare, P., & Goding, C. R. (1988) *Cell 52*, 435–445.
Otting, G., Qian, Y. Q., Billeter, M., Muller, M., Affolter, M., Gehring, W. J., & Wuthrich, K. (1990) *EMBO J. 9*, 3085–3092.
Pabo, C. O., & Sauer, R. T. (1992) *Annu. Rev. Biochem. 61*, 1053–1095.
Pavletich, N. P., & Pabo, C. O. (1991) *Science 252*, 809–817.
Pavletich, N. P., & Pabo, C. O. (1993) *Science 261*, 1701–1707.
Percival-Smith, A., Muller, M., Affolter, M., & Gehring, W. (1990) *EMBO J. 9*, 3967–3974.
Pollock, R., & Treisman, R. (1990) *Nucleic Acids Res. 18*, 6197–6204.
Pomerantz, J. L., Kristie, T. M., & Sharp, P. A. (1992) *Genes Dev. 6*, 2047–2057.
Preston, C. M., Frame, M. C., & Campbell, M. E. M. (1988) *Cell 52*, 425–434.
Qian, Y. Q., Billeter, M., Otting, G., Mueller, M., Gehring, W. J., & Wuthrich, K. (1989) *Cell 59*, 573–580.
Rebar, E. J., & Pabo, C. O. (1994) *Science 263*, 671–673.
Sayers, J. R., Krekel, C., & Eckstein, F. (1992) *BioTechniques 13*, 592–596.
Singh, H., LeBowitz, J. H., Baldwin, A. S., & Sharp, P. A. (1988) *Cell 52*, 415–423.
Smith, D. L., & Johnson, A. D. (1992) *Cell 68*, 133–142.
Stern, S., Tanaka, M., & Herr, W. (1989) *Nature 341*, 624–630.
Sturm, R. A., Das, G., & Herr, W. (1988) *Genes Dev. 2*, 1582–1599.
Treisman, J., Gonczy, P., Vashishtha, M., Harris, E., & Desplan, C. (1989) *Cell 59*, 553–562.
Verrijzer, C. P., Alkema, M. J., van Weperen, W. W., Van Leeuwen, H. C., Strating, M. J. J., & van der Vliet, P. C. (1992) *EMBO J. 11*, 4993–5003.
Vershon, A. K., & Johnson, A. D. (1993) *Cell 72*, 1–20.
Vinson, C. R., LaMarco, K. L., Johnson, P. F., Landschulz, W. H., & McKnight, S. L. (1988) *Genes Dev. 2*, 801–806.
Wilson, A. C., LaMarco, K., Peterson, M. G., & Herr, W. (1993) *Cell 74*, 115–125.
Wilson, D., Sheng, G., Lecuit, T., Dostatni, N., & Desplan, C. (1993) *Genes Dev. 7*, 2120–2134.
Wolberger, C., Vershon, A. K., Liu, B., Johnson, A. D., & Pabo, C. O. (1991) *Cell 67*, 517–528.
Youderian, P., Vershon, A., Bouvier, S., Sauer, R. T., & Susskind, M. M. (1983) *Cell 35*, 777–783.