© 1993 Oxford University Press Nucleic Acids Research, 1993, Vol. 21, No. 6 1439–1447

Identification of a DNA structural motif that includes the binding sites for Sp1, p53 and GA-binding protein

Michael C.MacLeod
Department of Carcinogenesis, Science Park-Research Division, University of Texas M.D.Anderson Cancer Center, Smithville, TX 78957, USA

Received December 4, 1992; Revised and Accepted February 17, 1993

ABSTRACT

We have analyzed predicted helical twist angles in the 21-bp repeat region of the SV40 genome, using a semi-empirical model previously shown to accurately predict backbone conformations. Unexpectedly, the pattern of twist angles characteristic of the six GC-boxes is repeated an additional five times at positions that are regularly interspersed with the six GC-box sequences. These patterns of helical twist angles are associated with a second, imperfectly-repeated sequence motif, the TR-box 5′-RRNTRGG. Unrelated DNA sequences that interact with trans-acting factors (p53 and GABP) exhibit similar twist angle patterns, due to elements of the general form 5′-RRRYRRR that occur as interspersed arrays with a spacing of 10–11 bp and an offset of 4–6 bp. Arrays of these elements, which we call pyrimidine sandwich elements (PSEs), may play an important role in the interaction of trans-acting factors with DNA control regions. In 13 human proto-oncogenes analyzed, we identified 31 PSE arrays, 11 of which were in the 5′-flanking regions of the genes. The most extensive array was found in the promoter region of the K-ras gene. Extending over 80 bp of DNA, it contained 16 PSEs that showed an average deviation from the SV40 criterion pattern of angles of only 1.2°.

INTRODUCTION

Much of our present knowledge of gene expression in eukaryotes is based on the *cis*-regulatory element: *trans*-acting factor paradigm (1,2). In this general model, a DNA sequence element located variable distances from the transcriptional start point of a gene is the target for the specific binding of a protein factor, and binding of this factor to its target sequence modulates the activity of the gene. Often, multiple elements are involved in regulation of a single gene and the elements may be either upstream or downstream from the transcriptional start point. Numerous such factor:sequence motif pairs have been discovered (1–4).

A well-studied and instructive example of this paradigm is found in the DNA virus SV40. Expression of the early and late transcription units of SV40 depends on a series of three tandem 21 bp repeats that contain 6 copies of the sequence GGGCGG (5–8), known as the GC-box motif. These motifs are targets for the zinc-finger transcription factor Sp1, and binding of this trans-acting protein to the GC-boxes appears to be necessary for transcription (5–8). Although the molecular details of the interaction between Sp1 and its cognate sequence have yet to be worked out, methylation interference studies indicate contacts are made with particular dG residues of the purine-rich strand in the major groove (5–8). It is assumed that contacts between the protein and functional groups in the DNA sequence impart sequence-specificity to this interaction.

Although the sequence-specific pattern of functional groups available for protein interaction in the major groove of the GC-box is certainly an important aspect of the specificity of interaction, the backbone conformation of the double helix may also play a role. Since the first crystal structure of a duplex oligonucleotide was determined (9), it has been clear that individual base pairs adopt conformations that differ significantly from the average properties of DNA. For example, the helical twist angle between adjacent base pairs in the dodecamer studied by Dickerson and Drew (9) varied from 27.4 to 40.3°. The most important sequence-dependent contribution to these conformational differences appears to be the ‘purine–purine clash’ which results from the propeller twist of the base pairs and the fact that purines are ‘longer’ than pyrimidines and therefore extend more than halfway across the interior of the double helix (10,11). Thus, if all base pairs adopted the average B-DNA conformation, adjacent purines in opposite strands in all 5′-YpR steps would severely overlap. One of the ways in which this ‘purine–purine’ clash is accommodated is by alterations in helical twist angles. The two base pairs involved adopt a low twist angle, while the twist angles with their nearest neighbors are higher than average. Sizeable changes in roll angle (the overall tilt of a base pair with respect to the long axis of the helix) also have been correlated with 5′-YpR steps (9,12). Sequence-dependent differences in conformational parameters such as helical twist and roll angle obviously contribute to the overall three dimensional shape of the DNA surface, and presumably thereby to the ability of DNA-binding proteins to recognize specific sequences.

In the present study we analyzed theoretical helical twist angle patterns in several DNA sequences involved in the *trans*-activation of gene expression and identified a degenerate sequence motif found in periodic arrays in several such DNA regions. Such arrays are shown to occur in a nonrandom manner in human oncogene sequences, suggesting that they may have some important function in the regulation of gene expression.

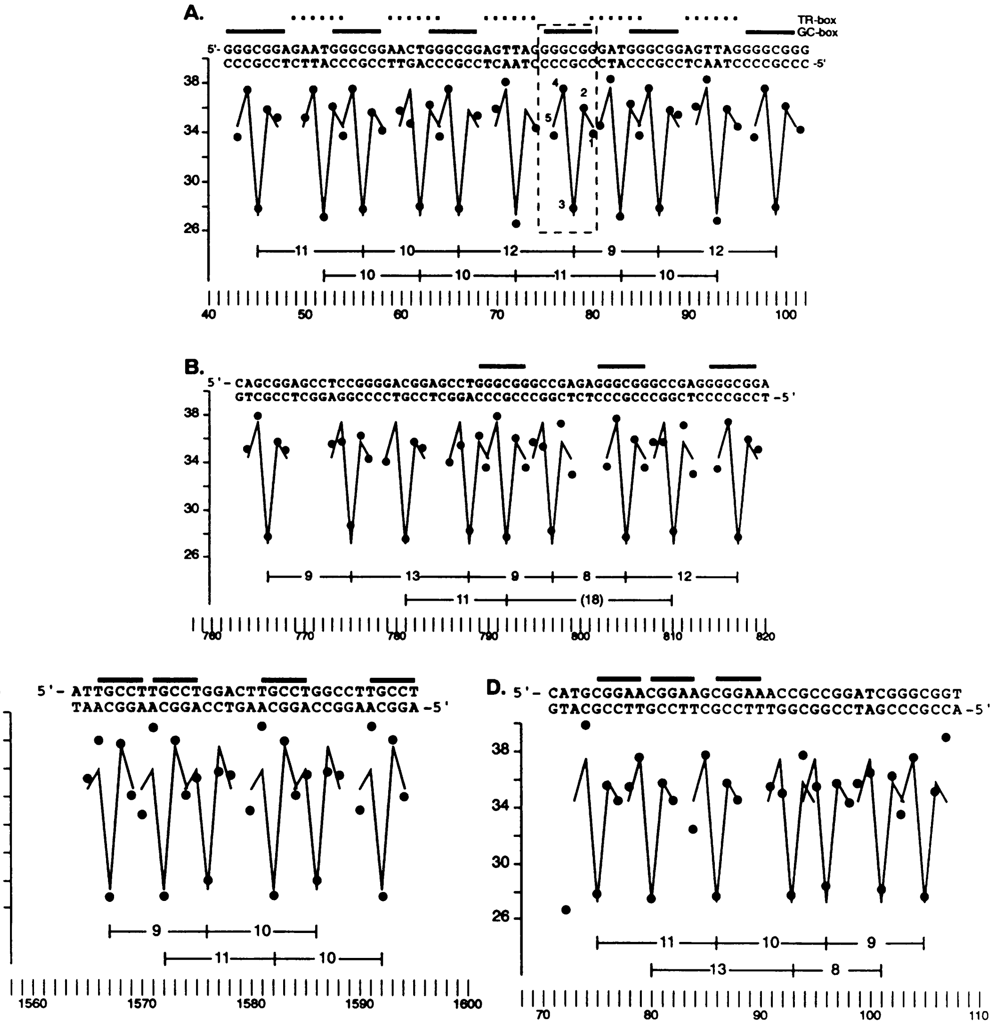

Figure 1. Patterns of twist angles in PSE arrays. Helix parameters were calculated according to the Tung–Harvey model (12) using the computer program AUGUR and assuming an average twist angle of 34° and B-form DNA. For each sequence, the individual angles are plotted (●) as a function of position in the sequence. (**A**) Data calculated for the 21-bp repeat region of SV40, viewed from the pyrimidine-rich strand, are plotted. The solid line superimposed on the individual elements represents the average values for the 11 elements found in the 21 bp repeats of SV40 (SV40 criterion angles, Table 1) in this and subsequent panels. Solid bars above the graph indicate the positions of the GC-boxes; dashed bars indicate the interspersed 'TR-boxes'. The distance between adjacent elements in each sub-array is indicated below the graph. (**B**) A portion of the mouse HPRT promoter is similarly analyzed; GC-boxes are indicated with solid bars. (**C**) A portion of the human ribosomal DNA repeat, recently shown to bind the p53 gene product (24), is analyzed. Solid bars indicate the positions of the repeated motif 5'-TGCCT . (**D**) The pattern of twist angles in the region of herpes simplex virus containing the ICP4 enhancer is shown, viewed from the pyrimidine-rich strand. Solid bars indicate the positions of the repeated CGGAAR motifs (here viewed from the pyrimidine-rich strand) previously shown to be footprinted by the GABP complex (27).

RESULTS AND DISCUSSION

Analysis of the 21 bp repeats of SV40

Several empirical models that predict helix conformation parameters from base sequence have been described that provide fairly good agreement with crystallographic data (10–13). We analyzed the twist angles between adjacent base pairs predicted by one of these models (12) in the 21 bp repeat region of the SV40 genome. The pattern of twist angles predicted for a single GC-box (Fig. 1A, dotted rectangle) is dominated by the low twist (angle 3, 27.9°) needed to relieve the purine–purine clash in the central 5'-CpG step. This is balanced by higher than average twist angles in the adjacent steps (angles 2 and 4, 35.9–37.7°) and close to average twist in the next nearest neighbors (angles 1 and 5). As expected from the sequence repetition, this pattern of twist angles is repeated six times with an average spacing of 10.6 bp. This periodicity positions homologous bases of each GC-box (Fig. 1A, solid bars) on the same side of the double helix, as previously noted (6,14). Unexpectedly, a very similar pattern of twist angles is found interspersed with the GC-box pattern centered over five occurrences of a 5'-TpR step with an average spacing of 10.3 bp (Fig. 1A, dashed bars). This set of 'TR-boxes' is offset an average of 4.6 bp from the GC-box pattern, placing the analogous bases on the opposite face of the double helix from the GC-boxes. Although a 5 bp periodicity in the occurrence of dG residues in the 21 bp-repeats of SV40 has been noted (14), the presence of the interspersed TR-box sequence motif has not to our knowledge been identified previously.

The sets of predicted average twist angles for each of these arrays (Table 1) are very similar, probably within the margin of error of the empirical model used for the analysis.

Table 1. Patterns of predicted twist angles in the 21 bp repeats of SV40.

Array	Angle 1	Angle 2	Angle 3	Angle 4	Angle 5
GC-boxes	34.8	36.0	27.9	37.7	33.7
TR-boxes	34.0	36.2	27.1	37.5	35.6
Both	34.4 (±0.7)	36.1 (±0.2)	27.6 (±0.5)	37.6 (±0.9)	34.6 (±1.1)

In the set of PSEs shown in Fig. 1A, the twist angle in each CpG step for the GC-box elements or in each TpR step for the TR-box elements was designated angle 3, and the preceding and subsequent angles were numbered sequentially from 3′ to 5′ (see Fig. 1A, dotted rectangle). The average twist at each position was calculated for each of the interspersed arrays or for the entire array; standard deviations are given in parentheses for the combined array.

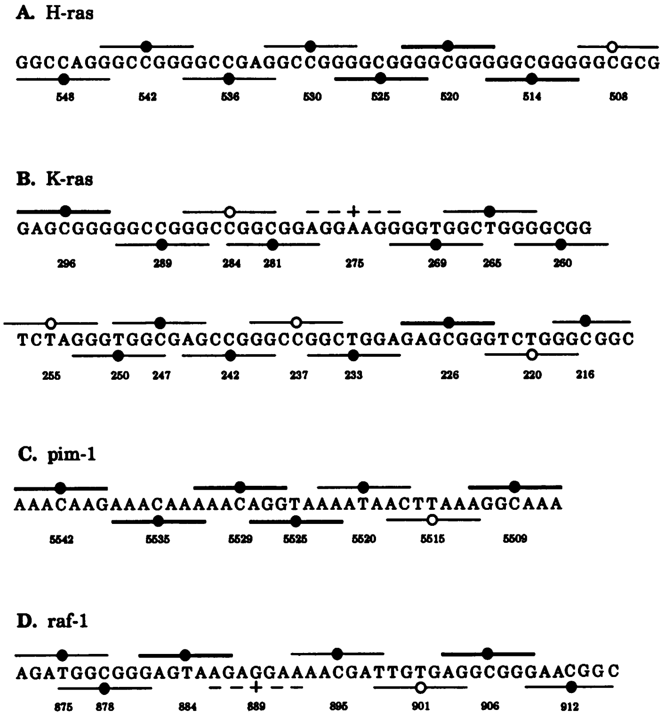
A. H-ras
GGCCAGGGCCGGGGCCGAGGCCGGGGCGGGGCGGGGGCGGGGGCGCG
548 542 536 530 525 520 514 508
B. K-ras
GAGCGGGGGCCGGGCCGGCGGAGGAAGGGGTGGCTGGGGCGG
296 289 284 281 275 269 265 260
TCTAGGGTGGCGAGCCGGGCCGGCTGGAGAGCGGGTCTGGGCGGC
255 250 247 242 237 233 226 220 216
C. pim-1
AAACAAGAAACAAAAACAGGTAAAATAACTTAAAGGCAAA
5542 5535 5529 5525 5520 5515 5509
D. raf-1
AGATGGCGGGAGTAAGAGGAAAACGATTGTGAGGCGGGAACGGC
875 878 884 889 895 901 906 912

Figure 2. PSE arrays in human proto-oncogenes. PSE arrays in selected oncogenes were detected as described for Figure 1. The sequences of portions of (A) Ha-ras, (B) K-ras, (C) pim-1, and (D) raf-1 are shown. In each case the sequence of the purine-rich strand is presented from 5′ to 3′; sequence numbering is from GenBank. Individual PSEs are indicated by the horizontal lines above and below the sequence, with a circle at the central pyrimidine of each element. The different line types and circles are explained below.

Consequently, we have been encouraged to treat both the GC-box and the TR-box arrays as members of a more fundamental pattern characterized by the general sequence motif 5′-RRRYRRR. The average values of the five twist angles for this pyrimidine sandwich element (PSE) as defined by the SV40 repeats are also given in Table 1; they will be referred to below as the SV40 criterion angles. When the 21-bp repeats are viewed in this framework, the two interspersed sequence motifs are: GGGCGGR and RRNTRGG. The six bases that flank the central pyrimidines are predominantly purines (63/66=95%), with the three occurrences of a pyrimidine in the flanking bases all at the position immediately 5′ to the central pyrimidine.

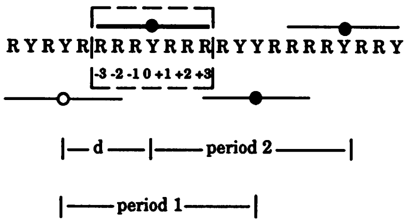

Figure 3. Hypothetical PSE array. The conventions used for numbering the internal positions of each PSE and for the calculation of offset and periodicity are diagrammed.

Table 2. Deviations of PSE-classes from the SV40-criterion angles.

Sequence Class	Position of Mismatch	Average RMS deviation	Range
RRRYRRR	–	0.9	0.3–2.4
YRRYRRR	−3	1.1	0.4–2.5
RRYYRRR	−1	1.3	0.3–2.1
RRRYRRY	+3	1.5	0.5–2.5
RYRYRRR	−2	3.9	3.3–4.7
RRRYRYR	+2	4.0	3.3–4.9

A set of 3706 PSEs with 0 or 1 mismatch from the consensus derived from the human oncogene dataset described below, were analyzed for their helical twist angle patterns. The average rms deviations from the SV40-criterion angles (and the ranges) were determined as a function of the position of the mismatch within the 7 nt window defined in Figure 3.

Analysis of promoters of housekeeping genes

The GC-box motif has been identified in the 5′-flanking regions of numerous mammalian 'housekeeping' genes (15–21), and has been implicated as a functional promoter in several genes including the hamster *APRT* gene (22). Since the GC-boxes of SV40 occur in a periodic, PSE array, we wondered whether similar structures are found surrounding other functionally important GC-boxes. Consequently, we analyzed twist angle patterns in several such genes by looking for close matches to the SV40-defined pattern. Small PSE arrays were found in the putative promoter regions of the mouse *HPRT*, human *APRT* and hamster *APRT* genes. For example, three matches to the SV40-criterion angles, spaced 10 bp apart and including two GC-boxes, were found in the hamster *APRT* promoter; a fourth PSE was interspersed within this region, offset by 5 bp. A more extensive PSE array was detected in the mouse *HPRT* promoter between nucleotides 767 and 818: six matches to the SV40 criterion angles, including three GC-boxes, were spaced an average of 10.2 bp, and three matches were interspersed with this array with an average offset of 6.3 bp (Fig. 1B).

In addition to the promoters of housekeeping genes, several human proto-oncogenes are known to contain GC-boxes in their promoters. In the case of the Ha-ras gene, 3 GC-boxes are contained in the functional promoter (23), adjacent to 4 sequences that differ by one or two bases from the canonical GC-box motif (Figure 2A). The latter 4 PSEs exhibit twist angle patterns very similar to the SV40 criterion angles (average rms deviation = 1.1°) and are arranged with the canonical GC boxes to make two interspersed subarrays with periodicities of about 11 bp. In the case of the K-ras promoter (Figure 2B), canonical GC boxes at positions 216 and 260 are contained within an extensive PSE array extending from 216 to 296 and containing 16 elements. The overall periodicity is 10.2 bp and the rms deviation from the SV40 criterion angles is 1.2° (Figure 2B).

PSE arrays at p53- and GABP-binding sites

Since the identification of Sp1 as the GC-box binding protein, numerous other proteins that activate transcription by interacting with specific DNA sequences have been described (1–4). We analyzed the DNA target sequences of several such *trans*-acting factors for the appearance of twist angle patterns that matched the SV40-defined PSE. Most of the target sequences examined, including those for TFIII-A, c-myc, AP-1, and AP-2, did not exhibit homologies (data not shown). However, target sequences for two known DNA-binding proteins did contain PSE arrays.

The product of a human tumor suppressor gene, p53, interacts with the 21-bp repeats of SV40 and with a region of the human rRNA gene cluster that contains several copies of the sequence 5′-AGGCA(24,25). However, upon inspection of the target sequence each of these can be seen to be embedded in a PSE of the form 5′-AGGCARR. Furthermore, as seen in Fig. 1C, the four AGGCA copies are contained in two interspersed PSE arrays with a total of six members; the average separation within each array is 9.8 bp, and the offset between the two arrays is 4.3 bp. When compared with the set of five SV40 criterion angles (Table 1), the overall rms deviation for the angles in these six elements is 1.4°. Two copies of the AGGCARR sequence are also present in the p53-responsive region of the murine muscle-specific creatine kinase promoter–enhancer (26). Analysis of twist angles in this sequence indicates that the p53-binding motifs are embedded in a 6-member PSE array with a basic periodicity of 11.0 bp (data not shown).

A third kind of target sequence that we found to contain PSE arrays is a *cis*-regulatory region of herpes simplex virus needed for immediate early transcription, the ICP4 enhancer (27). A heteromeric protein complex designated GA-binding protein (GABP) binds to this DNA sequence in vitro (27); one polypeptide in the complex is related to the Ets family of nuclear DNA-binding proteins, while the second polypeptide contains 4 imperfect copies of the ankyrin repeat motif. Using both DNase I footprinting and methylation interference assays to define the binding site, the sequence motif CGGAAR was suggested to be involved in binding of GABP (27). Our analysis of this DNA region (Fig. 1D) indicates that the three occurrences of this motif are contained in PSEs of the form 5′-RNRCGGA, and that a total of 7 PSEs are contained in a primary array with periodicity 10.2 bp, interspersed with a second array with periodicity 10.5 bp, offset from the first by 4.3 bp. The RMS deviation of angles in this array from the SV-40 criterion angles is 1.9°; much of this variability is contributed by the first and last elements of the array.

Description of the PSE array search algorithm

To more rigorously analyze the occurrence of arrays of PSEs in genomic DNA sequences, we developed a semi-automated, score-based sequence analysis to identify potential arrays. In Figure 3, the prototype PSE element is shown (surrounded by a dotted line) along with the internal numbering convention we have used. The most striking feature of the SV40 criterion angles is the low helical twist angle between positions 0 and +1, caused by the purine–purine clash inherent to all 5′-YpR steps. Therefore, in searching for arrays of PSEs, we required a pyrimidine in position 0 and a purine in position +1. Using a simple computer program, a 7-bp window was moved along a given sequence, and each position at which the sequence was identical to the prototype PSE, 5′-RRRYRRR, or contained one or two mismatches in positions other than 0 and +1 was identified as a match. In the hypothetical sequence of Figure 3, four such matches are identified by the horizontal lines above and below the sequence; the circles in the center of these lines indicate position 0 of each element. Clusters of 6 or more matches within 41 bp were identified as potential arrays, and were further analyzed. Each strand of a given sequence was searched independently, but only matches with the same polarity were used to construct a cluster.

We developed an arbitrary but consistent set of rules to eliminate from a given set of potential arrays those that were most unlike the arrays seen in the examples given above. This set of rules governed both the allowable spacings between elements, and the range of mismatches allowed. Spacings between elements, defined here as the distance in nucleotides between the central pyrimidines of adjacent PSEs (d in Figure 3), were allowed to have values between 3 and 7. For potential elements that overlapped severely ($d=2$), only one of the elements was considered. When adjacent elements had $d>8$, a 'gap' was introduced at the midpoint between the two elements. To eliminate potential arrays in which one of the interspersed subarrays was not convincingly periodic, sequences that had two adjacent gaps or had the structure 'gap–PSE–gap' were removed from consideration. Furthermore gaps adjacent to the element at either end of an array were not allowed; in such cases the end element was deleted. In addition, end elements that poorly matched the criterion angles (rms > 3) were deleted. Every other element (or gap) was assigned to one of two subarrays. These are indicated in Figure 3 as horizontal lines below (subarray 1) or above (subarray 2) the sequence, and the distance between adjacent elements of each subarray was used to determine the periodicity of the array.

Once the spacing of a potential array was determined, criteria governing the overall amount of mismatch from the prototype were applied. We first examined a large set of PSE matches found in a human sequence dataset (described more fully below), looking at the degree to which individual elements matched the SV40 criterion. We noted that all sequences that exactly matched the PSE prototype gave a predicted set of helical twist angles that very closely matched (average rms deviation <1.0°) those of the SV40 criterion set (Table 2). Among the set of sequences with one mismatch from the prototype, the two classes with mismatches at positions −2 and +2 gave helical twist angles with the greatest deviation from the criterion, (average rms ~4°) while sequences of the form 5′-RRYYRRR, -RRRYRRY and -YRRYRRR gave fairly close agreement (average rms 1.1–1.5°). The latter classes were designated 'good' matches, and all other classes were designated 'fair' matches. In Figures 2 and 3, the perfect matches are denoted by a thick line with a filled circle, the 'good' matches by a thin line with a filled circle, and the 'fair' matches by a thin line with an open circle. With these definitions, each potential array was checked for the following criteria: (i) the number of elements had to be at least 6; (ii) the number of perfectly matched elements had to be at least two; (iii) the sum of the 'perfect' and 'good' elements (as defined above) had to be at least 70% of the total number of elements plus 'gaps' in the array. If any of these criteria were not met, the potential array was disqualified. If a 'gap' had been introduced, the disqualified array was split at the gap and each

Table 3. PSE arrays in human proto-oncogenes

Gene	Location[a]	No. of PSEs	Deviation (°)[b]	Periodicity[c]
Ha-ras	508–548 (5′-flank)	8	1.2	11.3
Ha-ras	4348–4412 (3′-flank)	12	1.1	10.6
Ha-ras	694–735 (5′-flank)	9	1.4	9.3
Ha-ras	2378–2404 (exon)	6	1.1	8.3
myc	306–345 (5′-flank)	8	0.9	11.3
myc	3408–3443 (5′untrans)	7	0.9	10.1
myc	3005–3075 (5′-untrans)	14	1.5	10.9
myc	3924–3961 (5′-untrans.)	8	1.2	10.6
myc	3996–4040 (5′-untrans.)	9	1.5	11.0
myc	4688–4733 (exon)	10	1.0	10.4
myc	5600–5629 (intron)	7	2.1	9.4
myc	7906–7939 (3′-flank)	7	1.4	10.3
K-ras	216–296 (5′-flank)	16	1.2	10.2
fos	810–879 (5′-untrans.)	14	1.4	9.9
fos	77–112 (5′-flank)	7	1.4	11.8
p53	194–224 (5′-flank)	6	1.7	10.4
jun-a	307–354 (5′-untrans.)	11	1.4	9.4
jun-a	2844–2872 (3′untrans)	7	1.2	8.8
sis	9–59 (5′-flank)	10	1.0	9.8
sis	1614–1656 (5′-flank)	8	1.3	9.6
sis	4071–4150 (5′-flank)	15	1.4	9.4
sis	488–516 (5′flank)	7	1.1	10.5
sis	5014–5036 (5′-flank)	6	1.8	8.4
sis-5′	666–702 (5′untrans)	10	1.0	8.1
pim	5509–5542 (3′-untrans)	7	1.3	10.6
pim	5589–5622 (3′-untrans)	7	1.5	9.5
pim	5039–5080 (3′-untrans.)	8	0.9	10.3
raf-1	875–912 (5′untrans)	7	1.5	10.9
int-1	2634–2669 (intron)	7	1.2	10.0
int-1	3514–3543 (3′-untrans)	6	1.1	9.3
int-1	4089–4124 (3′-untrans)	6	0.8	11.7

Sequences were retrieved from GenBank, and a personal computer was used to calculate helix parameters (12) and identify positions that closely matched the consensus sequence (5′-RRRYRRR). PSE arrays were identified as described in the text.

[a] The locations of the central YpR steps of the first and last elements in each array are given using the GenBank numbering. untrans, untranslated.

[b] For each array, the average RMS deviation of the five angles for all PSEs of the array from the SV40-criterion angles is given.

[c] For all elements of each array, the fraction of the six nucleotides that flank each central pyrimidine which are purines is given.

Table 4. Length distribution of PSE arrays

		# of arrays of given length					
Dataset	Length Searched (kbp)	Total	n=6	n=7	n=8	n=9	n≥0
Oncogenes[a]	48.5	31	4	10	5	3	9
ras shuffled[b]	51.6	15	6	6	1	2	0
myc shuffled[c]	48.5	8	2	4	0	0	2
cDNA[d]	48.8	13	5	5	2	0	1
E.coli[e]	48.9	8	3	2	2	1	0

[a] The following sequences were searched: Ha-ras (J00277), c-myc (J00120), dbl (J03639), c-erb (M16892), int-l (X03072), jun-a (J04111), pim-1 (M27903), c-sis (Y00326, X03493), fos (K00650), K-ras (X07918), p53 (M26864), egfr (M11234), and raf-1 (M38134).

[b] The human Ha-ras sequence was shuffled 8 times and the resulting random sequences were searched.

[c] The human c-myc sequence was shuffled 6 times and the resulting random sequences were searched.

[d] The following sequences were searched: Hum25asyn (X04371), Hum3oct (X14813), Hum4cola (J05070), Huma2m (M11313), Humacadm (M16827), Humacp5 (J04430), Humada (K02567), Humadh21c (D00137), Humadh1ca (M12963), Humadppo (M17081), Humagalar (X05790), Humaiceb (J04144), Humalad (M13928), Humald (M19922), Humalda (M11560), Humaldb (K01177), Humaldh1 (K03000), Humalfuc (M29877), Humamipep (M22324), Humang (K02215), and Humargl (M14502).

[e] The following sequences were searched: Ecoace (V04198), Ecoada (M10211), Ecoadk (X03038), Ecoafr1 (M32083), Ecoair (X12982), Ecoalka (K02498), Ecoapt (M14040), Ecoaroc (M27714), Ecoarod (X04306), Ecodcma (M32307), Ecodld (X01067), Ecodnab (K01174), Ecodsda (J01603), Ecoentc (X12670), Ecogpta (M13422), Ecophoab (M29663), Ecopola (J01663), Ecouvraa (M13495), Ecouvrc (X03691), and Ecouvrd (X00738).

half was separately evaluated, according to the above criteria. This process was repeated until all portions of the potential array were either allowed or disqualified.

Analysis of human oncogenes

The scoring system just described was applied to several sequence datasets, each comprising a total sequence length of between 48,500 and 51,600 bp. Because both strands were searched, each dataset amounted to about 100,000 nt. Since preliminary studies of the human c-Ha-ras gene and the c-Ki-ras promoter showed these genes to contain several PSE arrays, we initially decided to study human oncogene sequences. We chose a set of eight human proto-oncogenes for which fairly complete genomic sequences information, including both 5′ and 3′ flanking regions,

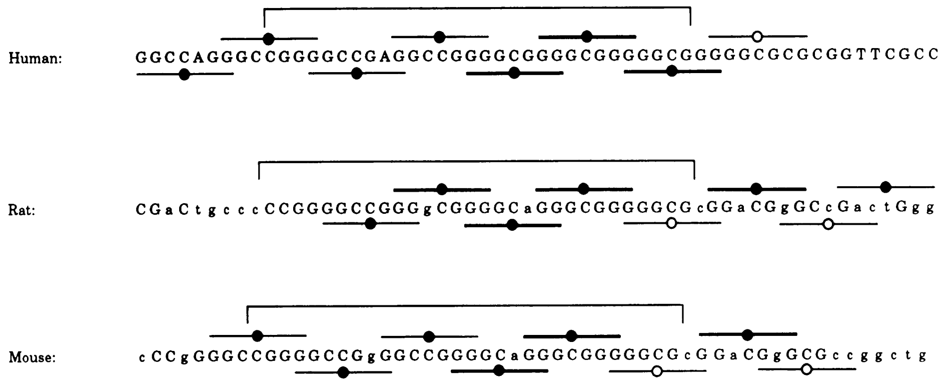
Human:
GGCCAGGGCCGGGGCCGAGGCCGGGGCGGGGCGGGGGCGGGGGCGCGCGGTTCGCC
Rat:
CGaCtgcccCCGGGGCCGGGgCGGGGCaGGGCGGGGGCGcGGaCGgGCcGactGgg
Mouse:
cCCgGGGCCGGGGCCGgGGCCGGGGCaGGGCGGGGGCGcGGaCGgGCGccggctg

Figure 4. Comparison of human, mouse and rat c-Ha-ras arrays. The sequences are all shown as the purine-rich strand and are derived as follows: human: nucleotides 551 to 496, GenBank J00277; rat: nucleotides −884 to −939, ref. 29; and mouse: nucleotides 129 to 75, ref. 30. Brackets indicate a core sequence that is highly conserved. Lower case symbols indicate nucleotide changes from the human sequence. Other symbols are as in Figures 2 and 3.

was available in GenBank, and five more for which genomic sequences including the 5′ flanking regions but only part of the coding region was available. Analysis of this dataset (48,500 bp) yielded a total of 31 PSE arrays distributed among 10 of the total of 13 proto-oncogenes. The locations of these 31 arrays are given in Table 3. In order to evaluate this finding, we constructed and searched two random datasets of approximately equivalent length. These were prepared using the Shuffle facility of the gcg software package (28) which randomly rearranges the order of a given nucleotide sequence. The datasets were prepared by 'shuffling' the c-Ha-ras sequence eight times and the c-myc sequence six times, respectively. The total number of arrays identified was 15 for the shuffled ras sequences and 8 for the shuffled myc sequences, factors of two to four lower than the number of arrays recovered in the oncogene dataset. This difference in the total number of arrays of length 6 or greater between the oncogene and shuffled datasets was statistically highly significant (chi-square, $p<0.01$).

The differences between the sets of arrays found in the oncogene and shuffled datasets were more pronounced if the data were considered in terms of the length of the arrays. In Table 4, the frequencies of occurrence of arrays containing different numbers of elements are compared. As one would expect for randomly occurring arrays, the length distribution in the shuffled datasets is skewed, with over three fourths of the arrays containing 6 or 7 elements. In contrast, less than half of the oncogene arrays are this short. If only arrays of 8 or more are considered, the frequency of occurrence in the oncogene dataset is about 7 fold higher than that in the shuffled datasets. Comparison of the two datasets in terms of arrays with 6 or 7 elements versus those with 8 or more indicates a highly significant difference between the oncogene data and the random data ($p=0.0141$, Fisher's exact test). We conclude that the occurrence of PSE arrays with 8 or more elements in the oncogene dataset is much greater than would be expected by chance alone.

A possible, trivial explanation for the higher number of PSE arrays in the oncogene dataset could be a higher frequency of occurrence of the PSEs themselves. However, examination of the datasets revealed that this was not the case. In fact the total frequency of PSEs in the shuffled datasets (.125) was slightly higher than in the oncogene dataset (.106). A second possibility was that coding sequences, which are of course contained in the oncogene set but not in the shuffled sets, tend to contain arrays of PSEs. To test this, we constructed a dataset of 25 human cDNA sequences containing a total sequence length of 48,800 bp, and searched it for PSE arrays as described above. As shown in Table 4, recovery of arrays in the cDNA dataset was similar to that in the randomly shuffled datasets; the total number of arrays was significantly smaller that those in the oncogene dataset (chi-square, $p<.01$).

Finally, it could be argued that the most important feature of the PSE arrays we have identified is a 'core' of two or three GC-boxes such as those illustrated in Fig. 1B for the mouse HPRT promoter or in Fig. 2A for the human c-Ha-ras gene. The occurrence of PSE arrays surrounding such features might then be expected to occur at a higher frequency than that measured in the shuffled oncogene dataset. However, although several of the sequences given in Figures 1 and 2 as examples of PSE arrays are rich in GC-boxes, this was by no means a consistent feature of the data. In fact, of the 31 PSE arrays identified in the oncogene dataset, 20 contained no GC-boxes and only 4 of the arrays contained two or more GC-boxes.

Furthermore, if the occurrence of arrays surrounding such GC-box clusters is a random occurrence, one would not expect such arrays to be conserved. In contrast to this expectation, the PSE array structure in the 5′-flanking region of the c-Ha-ras gene is conserved through mammalian evolution. The homologous regions of the human, rat and mouse genes are shown in Fig. 4. There is a central 30 base pair region (bracketed in Fig. 4) that is 93% identical between the three species. The sequence of the surrounding 25−26 nucleotides is only about 50% conserved in pairwise comparisons. However, in each sequence the 30 base pair core is part of an 8-membered PSE array that has a relatively uniform spacing between elements (11 ± 1) and contains no gaps. Because of the divergence at the nucleotide level, the position of the left-most PSE in the human and mouse sequences is shifted ±6 bp with respect to the rat sequence. This conservation of array structure over and above the conservation of GC-box motifs strongly suggests that the array structure itself makes an important contribution to the function of these DNA sequences.

Some of the basic parameters characterizing the set of oncogene PSE arrays are given in Table 3. The concordance of the twist angle patterns for the oncogene PSE arrays with the SV40 criterion, as measured by the RMS deviation for each array, ranged from 0.8° for a six-element array in the 3′-untranslated region of the int-1 oncogene to 2.1° for a seven-element array flanking the p53 gene; the average RMS deviation for the 272 PSEs contained in the 31 arrays was 1.3°. For arrays with 8 or more elements, the range of rms deviations was smaller (0.9 to 1.5°). This concordance is striking since only 13 of the 272 elements analyzed matched the GC box consensus sequence. A further 8 elements matched the 5′-AGGCA sequence, characteristic of the p53 protein binding site. The average periodicity of the interspersed arrays was 10.1 ± 0.9 bp and the offset between interspersed arrays was 5.1 ± 1.5 bp. This is illustrated by the schematic diagrams of several representative arrays given in Fig. 2.

Distribution of PSE arrays within oncogenes

In addition to the fine structure of the PSE arrays, the distribution of these arrays within the genes was also of interest. Of the 31 arrays identified (Table 3), 11 were 5′ to the major transcriptional start sites of the genes, 8 were in the 5′ untranslated regions of the major transcripts, and one was 3′ to the poly(A) addition site. Only four of the arrays were within exon sequences or internal introns. Eight of the 13 genes analyzed contained at least one PSE array 5′ to the coding sequences. This finding is consistent with a general role of PSE arrays in gene regulation, and suggests that searches for such arrays may prove useful in large-scale sequencing projects for finding the 5′-ends of genes.

Low frequency of occurrence of PSE arrays in prokaryotes

The majority of the data presented so far has dealt with human and viral sequences, although examples of mouse and hamster PSE arrays have been given. We wondered to what extent the occurrence of PSE arrays is common to all organisms. Analyses of selected genes from the chicken, fruitfly (*D.melanogaster*), nematode (*C.elegans*) and yeast (*S.pombe*) indicated the presence of fairly long PSE arrays in a wide spectrum of eukaryotes (data not shown). However, preliminary analysis of several bacterial genes failed to indicate long PSE arrays. Accordingly, we constructed and searched a 48,900 bp dataset, consisting of 20 genes from *E.coli*. As shown in Table 4, the occurrence and length distribution of PSE arrays in this dataset appeared similar to that in the randomly shuffled datasets. Comparison of the total number of arrays in the *E.coli* and oncogene datasets yielded a highly significant difference (chi-square, $p < .001$).

Information content of PSE arrays

Inspection of the oncogene PSE arrays for sequence motifs repeated within a single array indicated a wide variability. For example, three arrays (in the Ha-ras, K-ras and c-myc genes) that contained two or three copies of the GC-box consensus sequence also contained three to eight PSEs with a single base change from the consensus; two of these are illustrated in Figure 2A and B. The array shown in Figure 2C, from the pim proto-oncogene, contains two copies of the sequence AAACAA (positions 5542 and 5535) as well as two sequences one base change removed (AAACAG, position 5529, and AAATAA, position 5520). On the other hand, some arrays, such as the 7-member array from the promoter of raf-1 shown in Figure 2D, seem to show very little homology between elements other than the general conformance to the PSE motif.

In an attempt to quantify the similarities between elements within arrays, we used a method previously worked out to determine optimal alignments of protein-binding sequences (31,32). The method calculates an information content for each position in a sequence according to the formula:

$$\mathrm{I} = \sum_{b=A}^{T} f_b \bullet \log_2\left(\frac{f_b}{p_b}\right)$$

where the summation runs over the four normal bases (A, C, G, and T), f_b is the fractional occurrence of base b and p_b is the probability of occurrence of base b, taken to be 0.25 for all b. In cases where the actual frequency of a base was zero, an estimated frequency of 0.5/n was substituted (32), where n is the number of elements in the array. The value of this function is low when each of the 4 bases occurs at equal frequency at a given position and high when only a single base is found at that position. We applied this algorithm to each of the seven positions of the PSE element for each of the 31 arrays found in the oncogene dataset. For comparison, artificial arrays were analyzed by picking a position in a shuffled ras sequence that was known not to contain an array, and assigning every fifth or sixth nucleotide downstream from that position as the center of an 'element.' Twenty such pseudo-arrays with eight 'elements' each were analyzed to gain some feeling for the values of the information content to be expected for random 'arrays.'

Initially, we averaged the information content at each of the seven positions and compared the values so obtained for the random and oncogene arrays. The mean value for the overall information content of the oncogene arrays (0.74 ± 0.18), was approximately twice the value found for the random arrays (0.34 ± 0.11). This difference was highly significant (t-test, $p < .0001$). Thus, overall there are significant biases towards particular nucleotides at particular positions in the oncogene arrays. However, there is by no means an overwhelming bias, on average, since the maximal values for information content of our sequences (i.e. if each element of an array were identical) ranged between 1.61 and 1.74, depending on the number of elements. Thus, the average observed value was a little less than half the maximal value.

Given this finding, it was also of interest to compare the information contents at the different positions of the PSE. The information contents at each of the 7 positions of the PSE for the 31 oncogene arrays were compared by ANOVA, and the difference between positions was found to be significant ($p < .0001$). Comparison of the mean values for the seven positions indicated relatively low information content at positions −1 and +3 (0.53 ± 0.29 and 0.56 ± 0.33, respectively) and higher values at the other positions. Post-hoc analysis by the Scheffe F test, indicated that positions −1 and +3 had significantly ($p < .05$) lower information content than positions +1 and +2 (0.93 ± 0.26 and 0.86 ± 0.30, respectively).

CONCLUSIONS

If the occurrence of PSE arrays is not due to chance, then their widespread occurrence suggests a functional importance. Since we have noted above several such arrays that occur in sequences that are targets for DNA-binding proteins (viz. Spl, p53 and GABP), we suggest that the PSE array is a fundamental pattern that can be important in the interaction of *trans*-acting factors with their target sequences. Several potential roles can be envisioned. 1) The presence of multiple PSEs may reflect the complexity of regulation of the genes examined, with each array consisting of several sequence motifs that bind different *trans*-acting factors. The close spacing of the PSEs could facilitate binding of heteromeric dimers, or could allow modulation of binding affinity at a given site based on occupancy of adjacent binding sites (33,34). The recent identification of a family of Sp1-like proteins (35) with overlapping DNA-binding specificities adds a further dimension to this possibility. 2) The presence of an array of structurally homologous sites surrounding a single binding site could effectively increase the probability that the binding site is occupied by its cognate factor if the factor exhibits lower affinity binding to other sites in the array. Binding to the secondary sites would have the effect of increasing the local concentration of the factor, facilitating binding to the primary site. The transcription factor Sp1 in particular is known to bind to several different sequences related to the GC-box consensus (35–38). As an example, the PSE array from the K-ras promoter (Figure 2B) contains two consensus GC-boxes (GGGCGG) and 8 PSEs that differ by 1 nucleotide from the consensus, 2 of which are identical with the retinoblastoma control element (GGGTGGC), which was recently demonstrated to bind Sp1 (38). 3) The longer range structure of an array may determine the probability of formation of a nucleosome covering the binding site, or may determine the precise positioning of a nucleosome. Long PSE arrays were found in all eukaryotes examined but not in prokaryotic genes (Table 4). This finding is consistent with the suggestion that PSE arrays are somehow involved in nucleosome positioning, since prokaryotes lack nucleosomes. By analogy to the TFIII-A/5S-DNA system (39), nucleosome formation and positive regulation via factor-binding may be alternative, stable states in many genes. Inhibition of transcription factor binding by nucleosome positioning is by no means universal, however, as evidenced by the binding of glucocorticoid receptor (40–43) and GAL-4 (44) to positioned nucleosomes. Indeed, it has been suggested that in some cases precise rotational positioning of a nucleosome ensures access of a factor to its DNA target displayed on the surface of a nucleosome (42). It is worth noting that the average periodicity of the PSE arrays identified in the human oncogene dataset (10.1°) is similar to the rotational periodicity of DNA on the surface of nucleosomes (45–47), and it will be interesting to determine to what extent PSE arrays influence nucleosome positioning.

Many of the conceptual advances made in the past 10 years in our understanding of transcriptional regulation have relied on a linear model, namely the search for motifs in the primary DNA sequence that are associated with specific protein binding. The results presented here reinforce the importance of the 3-dimensional surface presented by the DNA in these interactions and suggest the need for non-linear tools in analyzing DNA sequence data for possible regulatory sequences. The extent of several of the PSE arrays identified and their apparent inherent redundancy suggests that they may have multiple functions, particularly in the complex regulation of pivotal cellular genes such as the proto-oncogenes. Dissection of these functions may necessitate non-linear approaches (for example, changes in the phasing or arrangement of PSEs) in addition to the standard (linear) approach of site-directed mutagenesis. Similar approaches are currently in use in the dissection of DNA sequence contributions to nucleosome positioning (48,49). The alterations in DNA conformational parameters that result from 'purine–purine clash' and which served as the starting point for the present analysis have been suggested as important factors in nucleosome positioning (50). This raises the intriguing possibility that PSE arrays have a direct role in determining rotational positioning of nucleosomes.

ACKNOWLEDGEMENTS

Supported by NIH grants CA-35581 and ES-03602. We thank our colleagues W.Wood, R.Nairn, C.Conti, and C.Aldaz for critical reading of the manuscript, D.Johnston for statistical advice, R.Tan for supplying the AUGUR program, V.Conti and P.Goertz for technical assistance, M.Gardiner for manuscript preparation, and J.Riley for excellent graphics support.

REFERENCES

1. Johnson,P.F. and McKnight,S.L. (1989) *Annu. Rev. Biochem.*, **58**, 799–839.
2. Wingender,E. (1988) *Nucleic Acids Res.*, **16**, 1879–1902.
3. Dynan,W.S., and Tjian,R. (1985) *Nature*, **316**, 774–778.
4. Lee,W., Mitchell,P., and Tjian,R. (1987) *Cell*, **49**, 741–752.
5. Dynan,W.S., and Tjian,R. (1983) *Cell* , **35**, 79–87.
6. Gidoni,D., Dynan,W. S. and Tjian,R. (1984) *Nature*, **312**, 409–413.
7. Gidoni,D., Kadonaga,J.T., Barrera-Saldaña,H., Takahashi,K., Chambon,P., and Tjian,R. (1985) *Science*, **230**, 511–517.
8. Albrecht,G.R., Cavallini,B. and Davidson,I. (1989) *Nucleic Acids Res.*, **17**, 7945–7963.
9. Dickerson,R.E., and Drew,H.R. (1981) *J. Mol. Biol.*, **149**, 761–786.
10. Calladine,C.R. (1982) *J. Mol. Biol.*, **161**, 343–352.
11. Dickerson,R.E. (1983) *J. Mol. Biol.*, **166**, 419–441.
12. Tung,C.-S., and Harvey,S.C. (1986) *J. Biol. Chem.*, **261**, 3700–3709.
13. Kabsch,W., Sander,C. and Trifonov,E.N. (1982) *Nucleic Acids Res.*, **10**, 1097–1104.
14. Rhodes,D., and Klug,A. (1986) *Cell*, **46**, 123–132.
15. Reynolds,G.A., Basu,S.K., Osborne,T.F., Chin,D.J., Gil,G., Brown,M.S., Goldstein,J.L. and Luskey,K.L. (1984) *Cell*, **38**, 275–285.
16. Dush,M.K., Sikela,J.M. Khan,S.A. Tischfield,J.A. and Stambrook,P.J. (1985) *Proc. Natl. Acad. Sci.*, USA **82**, 2731–2735.
17. Mitchell,P.J., Carothers,A.M., Han,J.H., Harding,J.D., Kas,E., Venolia,L. and Chasin,L.A. (1986) *Mol. Cell. Biol.*, **6**, 425–440.
18. Azizkhan,J. C. Vaughn,J. P. Christy, R. J. and Hamlin,J. L. (1986) *Biochemistry*, **25**, 6228–6236.
19. Nalbantoglu,J., Phear,G. A. and Meuth,M. (1986) *Nucleic Acids Res.*, **14**, 1914.
20. Broderick,T.P., Schaff,D.A., Bertino,A.M., Dush,M.K., Tischfield,J.A. and Stambrook,P.J. (1987) *Proc. Natl. Acad. Sci.* USA **84**, 3349–3353.
21. Edwards,A., Voss,H., Rice,P., Civittello,A., Stegemann,J., Schwager,C., Zimmermann,J., Erfle,H., Caskey,C.T., and Ansorge,W. (1990) *Genomics* 6, 593.
22. Park,J.-H. and Taylor,M. W. (1988) *Mol. Cell. Biol.*, **8**, 2536–2544.
23. Ishii,S., Kadonaga,J.T., Tjian,R., Brady,J.N., Merline,G.T. and Pastan,I. (1986) *Science*, **232**, 1410–1413.
24. Kern,S.E., Kinzler,K.W., Bruskin,A., Jarosz,D., Friedman,P., Prives,C. and Vogelstein,B. (1991) *Science*, 252, 1708–1711.
25. Bargonetti,J., Friedman,P.N., Kern,S.E., Vogelstein,B. and Prives,C. (1991) *Cell*, **65**, 1083–1091.
26. Zambetti,G.P., Bargonetti,J., Walker,K., Prives,C., and Levine,A.J. (1992) *Genes and Development*, **6**, 1143–1152.
27. Thompson,C.C., Brown,T.A. and McKnight,S.L. (1991) *Science*, **253**, 762–768.
28. Devereuz,J., Haeberli,P. and Smithies,O. (1984) *Nucleic Acids Res.*, **12**, 387–395.
29. Damante, G., Filetti, S. and Rapoport, B. (1987) *Proc. Natl. Acad. Sci.* USA, **84**, 774–778.
30. Plumb, M., Telliez, J.-B., Fee, F., Daubersies, P., Bailleul, B., and Balmain, A. (1991) *Molecular Carcinogenesis*, **4**, 103–111.
31. Schneider,T.D., Stormo,G.D., Gold,L., and Ehrenfecht,A. (1986) *J. Mol. Biol.*, **188**, 415–431.
32. Stormo,G.D., and Hartzell,G.W.,III. (1989) *Proc. Natl. Acad. Sci.*, USA **86**, 1183–1187.
33. Hoch,M., Gerwin,N., Taubert,H., and Jackle,H. (1992) *Science*, **256**, 94–97.
34. Lehmann,J.M., Zhang,X.-K., and Pfahl, M. (1992) *Mol. Cell. Biol.*, **12**, 2976–2985.
35. Kingsley,C., and Winoto,A. (1992) *Mol. Cell. Biol.*, **12**, 4251–4261.
36. Li,R., Knight,J.D., Jackson,S.P., Tjian,R. and Botchan,M.R. (1991) *Cell*, **65**, 493–505.
37. Pitluk,Z.W., and Ward,D.C. (1991) *J. Virol.*, **65**, 6661–6670.

38. Kim,S.-J., Onwuta,V.S., Lee,Y.I., Li,R., Botchan,M.R., and Robbins,P.D. (1992) *Mol. Cell. Biol.*, **12**, 2455–2463.
39. Wolffe,A. P., and Brown,D. D. (1988) *Science*, **241**, 1626–1632.
40. Richard-Foy, H., and Hager,G.L. (1987) *EMBO J.*, **6**, 2321–2328.
41. Perlmann,T., and Wrange,W. (1988) *EMBO J.*, **7**, 3073–3079.
42. Pina,B., Bruggemeier, U., and Beato,M. (1990) *Cell*, **60**, 719–731.
43. Archer,T.K., Cordingley,M.D., Wolford,R.G., and Hager,G.L. (1991) *Mol. Cell Biol.*, **11**, 688–698.
44. Taylor,I.C.A., Workman,J.L., Schuetz,T.J. and Kingston,R.E. (1991) *Genes and Development*, **5**, 1285–1298.
45. Lutter,L.C. (1978) *J. Mol. Biol.*, **124**, 391–420.
46. Drew,H.R., and Calladine,C.R. (1987) *J. Mol. Biol.*, **195**, 143–173.
47. Pina, B., Truss, M., Ohlenbusch,H., Postma,J., and Beato,M. (1990) *Nucleic Acids Res.*, **18**, 6981–6987.
48. Shrader,T.E., and Crothers,D.M. (1989) *Proc. Natl. Acad. Sci.*, USA **86**, 7418–7422.
49. Shrader,T.E., and Crothers,D.M. (1990) *J. Mol. Biol.*, **216**, 69–84.
50. Calladine,C.R., and Drew,H.R. (1986) *J. Mol. Biol.*, **192**, 907–918.