Sequence Signals in Eukaryotic Upstream Regions

Ruth Nussinov

Two DNA sequence elements are known to recur frequently upstream of eukaryotic polymerase II-transcribed genes. The TATAAA, at position −40, specifies the transcription initiation site. The GGCCAATCT is less frequent around −80. Sequence analysis of upstream regions reveals that the underlined yeast UAS2 consensus sequence, TGATTGGT, is also very frequent at −80 in higher polymerase II-transcribed animal sequences. The underlined CCAAT box and yeast UAS sequences are complementary. Structural analysis suggests some symmetry in their DNA structures. Upstream of the TATA-AT-rich region there is an abundance of GC seqeunces. Analysis of nucleotide tracts indicates that these are preferentially flanked by their complementary nucleotides with a pyrimidine-purine junction, i.e., TTA_N, CCG_n, C_nGG, T_nAA. Here, I discuss DNA structrual considerations in upstream regions along with protein readout of the major and minor groove information content. These sequence-structure aspects are put in the general context of protein (factors)-DNA (elements) recognition and regulation.

I. INTRODUCTION

Regions upstream of mRNA start sites in eukaryotic polymerase II-transcribed genes contain relatively short signal sequences. These DNA signals, termed "elements", are recognized by a variety of protein "factors". Upstream sequences contain many recognition elements, signaling the position of transcription initiation and regulating its efficiency. The different DNA elements are recognized by their corresponding sequence-specific protein factors. A growing number of these elements and factors have been identified, and a fascinating general scheme begins to emerge. Although the title of this review is *"Sequence Signals in Eukaryotic Upstream Regions"*, one needs to put these signals into the overall perspective. Signal sequences are only pieces in a jigsaw puzzle.

One of the basic emerging principles is the modular nature of DNA elements and of protein factors. For upstream/enhancer elements, this enables exchange either of different elements within the same upstream region or between different species. Both changes of locations and orientations can occur. Protein factors commonly contain at least two modules — one for binding DNA, the other for protein-protein interaction. Here, too, modules from different proteins can be exchanged.

The main advantage of this modular nature is the increased number of possibilities for gene regulation. Currently compiled evidence is consistent with this explanation. Thus, heterologous protein factor subunits can be assembled (e.g., HAP2/HAP3). Also, a protein-protein activation domain is itself apparently composed of several subdomains (e.g., in the Sp1) where, depending on distance and orientation, different ones can be used. This review is studded with such examples. For viruses, such a system is particularly useful, since it allows them to grow in a wide host range and under different conditions.

Given these advantages and the long evolutionary time required to perfect the system, it is not surprising that the regulation of polymerase II transcription initiation has been so well conserved in eukaryotes from yeast to man.

The protein factors do not read the nucleotide *letter* sequence as such. Rather, they sense the three-dimensional structure and the potential hydrogen bonding groups and hydrophobic interactions embedded in it. These, along with another parameter, DNA flexibility, playing a crucial role in DNA protein recognition, are sequence dependent. In this review, we discuss these aspects in general. We focus on the particular ways by which they pertain to upstream regions and to protein-mediated regulation of DNA expression.

II. DNA AND PROTEIN MODULARITY AND EVOLUTIONARY CONSERVATION

Perhaps the most distinctive property associated with eukaryotic polymerase II promoters is the modular character of their elements.[1,2] On the DNA side, each module may contain a putative DNA-protein recognition site, often in the form of a relatively short consensus sequence, specifying some regulatory function of the promoter. Elements in two promoters may then play equivalent roles and can be exchanged. The possibility of interchanging such elements without seriously affecting the functioning of the DNA is the best direct proof of the modular character. A classic example of DNA modules is furnished by the three distinct sequence elements that have been located for the thymidine kinase (tk) gene of the herpes virus.[1,3] The first is the TATA box, located 40 base pairs (bp) upstream from transcription initiation, which is needed for accurate initiation. The second, middle region, at −50 to −70, and the third, distal regions, at −80 to −110, are needed for efficient transcription initiation. The β-globin gene also has three analogous elements. Thus, a functional promoter can be constructed by joining the distal component of the tk gene to the middle and TATA component of the β-globin gene.

On the protein side we have corresponding modular "factors". Initiation and regulation of eukaryotic messenger RNA synthesis by RNA polymerase II requires transcriptional acti-

R. Nussinov received a B.Sc. degree from the University of Washington in Seattle and M.S. and Ph.D. degrees from Rutgers University in New Brunswick, New Jersey. Dr. Nussinov is currently an Associate Professor in the Department of Molecular Medicine at Tel Aviv University in Tel Aviv, Israel. Dr. Nussinov is also currently associated with the Laboratory of Mathematical Biology at the National Cancer Institute of the National Institutes of Health in Bethesda, Maryland.

These features are also characteristic of the yeast transcription factors having two subunits, HAP2 and HAP3.[65] Indeed, both the CCAAT box factors and the yeast factors make identical contacts with the DNA.[63,66] This organization again shows the functional superiority of the heteromeric protein system. A single protein composed of two or more subunits, where each can be independently regulated in response to tissue signals,[64] has more regulatory options. Moreover, as noted above, the interchange of DNA binding and transcription activating subunits among such proteins provides an additional level of combinational complexity to gene regulation. Different subunit combinations may increase or decrease the specificity of a factor in transcriptional activation. This can explain how some molecules that appear to lack intrinsic DNA binding activity are able to modulate the transcription activity of DNA binding proteins.[67]

Another fascinating aspect of the modular approach recently discovered in human context and which has apparently been adopted by the cellular machinery is worth noting.[67,68] Individual members of the human CCAAT box binding proteins, CTF/NF-I, are implicated in both eukaryotic transcription activation as well as *cis* control for DNA replication.[29] Further, molecular analysis of these human clones reveals that the corresponding mRNA species coding for the above arise from transcripts of one gene that have been spliced different ways, yielding alternate coding regions.

B. Upstream and Enhancer Elements: The SV40 Example

Both upstream and enhancer elements stimulate transcription independently of their position and orientation. As Wasylyk[26] notes in his recent review, it appears that many aspects of the mechanism of transcription activation by proximal (upstream promoter) and distal (enhancer) elements is similar. The difference may lie just in the advantages that proximity itself can confer. The implications of the distance will be considered in some detail in later sections.

The SV40 early promoter has been the subject of particularly extensive deletion, insertion, base substitutions, and rearrangement studies. Also, some of the specific protein-DNA interactions that take place in a eukaryotic control region have been first identified in the SV40.[3] A schematic diagram of the 300-bp region of the SV40 control region containing a number of regulatory elements is given in Figure 1.

Starting at the left-hand side of the diagram (the T-antigen gene) and proceeding to the right, the first regulatory element encountered is the origin of DNA replication.[3,70-72] Three T-antigen binding sites[3,72-74] and the sites of early transcription initiation[3,76-78] are located within the 70-bp domain containing this origin of replication. Abutting the origin of replication is an AT-rich stretch, homologous to the TATA box sequence. However, its deletion only slightly affects transcription.[3,79-81] Adjacent to the AT-rich sequence, there are three 21-bp-long direct repeats. Each 21-bp repeat contains two copies of the hexanucleotide sequence GGGCGG. Altogether, we have the six GC motifs, I, II, ... VI. As might be expected, complete deletion of the three 21-bp repeats severely limits early promoter function.[3,82-84] However, viral mutants having only two of the GC hexanucleotides are viable. Also, the 21 bp are functional even when inverted.[84,85]

One of the earliest identified transcription factors, the Sp1, binds to the GC hexanucleotide in a sequence-specific manner.[86-88] Systematic mutagenesis has shown that the order of binding affinities of the six GC motifs to the Sp1 protein is III>V>II>VI>I>IV.[26,85,89] It may result from differences in the sequences flanking the GC hexanucleotides and from steric constraints.[3,26,85,89] The latter apparently limit the Sp1 interaction to five sites at any one time. The weak Sp1 binding site I is most important for early transcription, possibly because of its proximity to the start site.[26] As discussed in Section V, the Sp1 appears to bind in the major groove of the DNA helix. Sp1 also activates cellular genes[88,90,91] containing the GC hexanucleotide core. Interestingly, the Sp1 has been recently shown[92] to bind DNA and activate transcription even when the binding site is CpG methylated.

The SV40 enhancer has two distinct levels of organization.[27] The first level contains three discrete 15- to 20-bp-long elements, called A, B, and C, that cooperate with one another or duplicates of themselves to enhance transcription[27,93-95] (Figure 1). These enhancer elements function autonomously when present in two or more copies and display unique patterns of cell-specific enhancer activity.[27,96,97] In the second level of organization the A and B elements themselves are viewed as bipartite, containing subunits that correlate either with the core

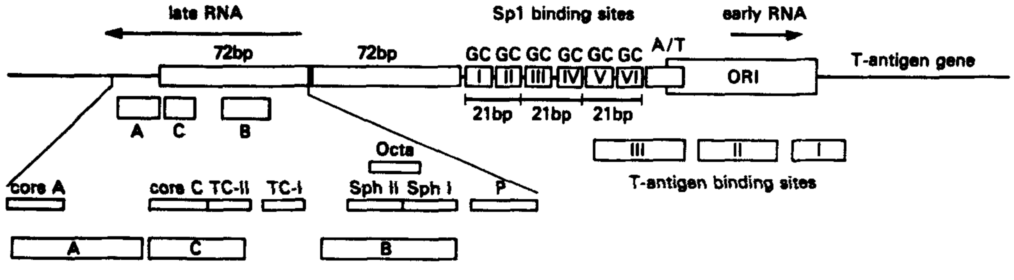

FIGURE 1. A schematic diagram of the SV40 control region. (Redrawn from References 3 and 27.)

(GTGG$^{\text{AAA}}_{\text{TTT}}$G)[98] or with the sph (AAG$^{\text{T}}_{\text{C}}$ATGCA)[99] sequence motifs. Ondek et al.[27] show that these subunits, termed enhansons, can be duplicated or interchanged to create new enhancer elements.

Current data suggest that the SV40 enhancer contains multiple levels of binary organization. This organization is evident in the operative, synthetic SV40 enhancer-β-globin construct. The SV40 enhancer is separated by more than 2 kb from the β-globin upstream promoter region. The enhancer contains two B elements (rather than A, B, and C as in the original SV40 enhancer). Each of the B elements contains one pair of sph enhansons.[27] It is intriguing to note that successive levels exhibit increased sensitivity to spacing between their components.[27] Thus, the enhancer, as a whole, is capable of activating transcription when separated by as much as 10 kb[100] from the proximal promoter elements — the CCAAT and TATA boxes. The enhancer is, in turn, composed of elements that can enhance transcription when separated by up to 100 bp. Finally, a pair of enhansons must be in very close proximity to create a functional enhancer element.[27]

As for the promoters, an enhancer structure that contains multiple levels of binary organization allows a greater degree of transcriptional regulation with a smaller number of factors. As Ondek et al.[27] suggest, the coreA/coreA enhanson pair needs only the core A recognizing factor. A sphII/sphII enhanson pair also needs solely the sphII binding factor. The sphII/coreA enhancer element is, however, dependent on two factors. For the enhanson simplest binary-motif level, there are $2^2 = 4$ possibilities. For the enhancer binary element level, there are $4^2 = 16$ possibilities. Clearly, for each level more than two motifs are possible, yielding many combinations, some of which may be inactive. The high degree of positional freedom enjoyed by the enhancers increases the probability that spontaneous mutations would occur in this longer DNA span creating new enhancers.[27] If, on the other hand, enhancers were structurally simple, such mutations might change them frequently, creating chaos in the cell expression.

In their experiments Ondek et al.[27] used the CoreA (termed GT-IIC motif in References 99,101-103), sphI, and sphII motifs. However, additional motifs (e.g., GT-I, TC-II, Oct, p) are present in the A and B elements of the SV40 enhancer. Davidson et al.[102] have recently purified the TEF-1 protein that specifically binds to two SV40 enhancer *un*related sequence motifs (GT-IIC and sph). TEF-1 binds cooperatively to DNA templates containing tandem repeats of either of the cognate motifs. It does not, however, bind to inverted or spaced motifs.

Based on these and other enhancer protein binding studies, Davidson et al.[102] and Fromental et al.[101] distinguish between four different classes of enhancer factors. Class A (containing TEF-1) corresponds to factors that bind cooperatively to tandem repeats of their cognate sequences. When bound as dimers, such class A enhancer factors can act in concert to generate enhancer activity.[102] In contrast, class B (containing TEF-2) corresponds to factors that do not bind cooperatively to tandem repeats of their cognate motifs, and on their own generate minimal enhancer activity.[101,102] However, the association of a class B (TEF-2) with a class A (TEF-1) factor results in a functional unit increasing enhancer activity.[101] The motifs need not be close to each other. Class C contains the lymphoid-specific octamer binding protein[102,104] and the NF-kB protein that binds to the TC-II SV40 enhancer motif.[105,106] Also, in this class the factors act in concert, even when the motifs are not close to each other.[102] Examples of class D are the steroid/thyroid hormone and retinoic acid nuclear receptor family. These factors promote enhancer activity also when bound to single copies of their cognate motifs. The redundancy of information in enhancers as a principle of mammalian transcription control is further reviewed by Schaffner et al.[107]

Additional detailed discussion of upstream and enhancer elements is given in a recent review of the transcription elements and their binding factors.[26] A detailed review[108] along with some further recent results of the interactions between promoter elements and their factors in adenovirus E1A-dependent transactivation are given by Kovesdi et al.,[109,111] Raychaudhuri et al.,[110] and Simon et al.[112] Upstream polyoma DNA element factor regulation is described by Kovesdi et al.[113]

IV. SPACING AND LOOPING

In eukaryotes, *cis*-acting upstream and enhancer elements can activate transcription at a distance. In prokaryotes, transcription of one DNA region affects the initiation of replication at a distant site. How does an enhancer sequence affect transcription of a gene thousands of base pairs away?

In a recent review, Wang and Giaever[114] discuss three classes of mechanisms for action at a distance: tracking, or movement of a protein along the DNA; looping, by bringing two sites together; and distal action, through influence on the topology of a DNA loop. The modes of action at a distance are often interrelated. For example, tracking can induce loop formation. This happens if one protein domain is bound to the DNA and a second domain tracks through an increasingly longer DNA sequence. Alternatively, one protein can be bound to the DNA through one of its DNA binding domains. Its second domain may interact with another protein. Tracking of the DNA by the DNA binding domain of this protein would also involve loop formation. Tracking along a DNA might be affected by DNA sequences and structures or by DNA-bound proteins encountered along the way. In eukaryotes, the nucleosomal core particles[114-118] or other regulatory proteins may impede the tracking. The directionality of the tracking stems from the asymmetry of the protein. The binding of a protein with a dyad symmetry to an asymmetric DNA sequence might also bias the direction of tracking.[114] Here, we discuss protein tracking along the DNA. For large protein complexes, such as RNA polymerase or the multifactor promoter complex, it might be more likely that it is the DNA that tracks through the protein.

The DNA loop is held together by proteins and forms a "topological domain". If the protein is rigid, the DNA loop is subject to the same topological constraints as those in a covalently closed DNA ring.[114] Cooperative binding of protein factors to a pair of DNA elements has been observed if the centers of the DNA sites are separated by an integral number of helical turns (5 or 6 in the phage λ repressor), but not if they are separated by 4.6, 5.5, or 6.4 turns.[114,119-121] These results are consistent with the model where two protein moluecules, bound on the same side of the DNA double helix, come together and form a smooth bend in between. For relatively short DNA sequences, this aspect may become critical. Changes in spacing by half integral helical turns places the proteins on opposite sides of the helix, necessitating twisting or writhing of the DNA between the two sites. This is thermodynamically less favorable than a smooth bend in a plane.[114]

Loop formation is expected to be less likely for shorter overall distance of the two sites at its ends.[26] The DNA is far less rigid than was assumed until recently. Yet, it still is increasingly more difficult to bring two DNA sites together by looping as the chain length between them decreases below about 250 bp.[26] Indeed, in the prokaryotic synthetic lac construct, when the distance between the two operators was decreased from 200 to 120 bp, a significant drop in the repression effectiveness was observed.[122] This diminishing effect at short distances[123] is a key observation supporting the DNA looping model.

Potential flexibility of the protein components should not, however, be overlooked.[122] Such a flexibility may explain the small difference between cooperative protein binding effects when the interoperator spacing is reduced from 63 to 52 bp.[119,122] The protein flexibility can potentially play an important role also in the eukaryotic RNA polymerase II transcription initiation system, where upstream DNA control elements may also be present at short range. This could conceivably happen as follows. The activating domains of the protein transcription factors that are bound to the DNA elements contain a "negative noodles" region. The latter can interact with the carboxyl-terminal domain of the RNA polymerase II, which contains a universal recurring motif of a tandemly repeated seven-amino acid sequence.[6,122-125] Indeed, this flexibility might be the key role of the "negative noodle" recurring motif, frequently found in sequence-specific DNA binding regulatory proteins. The apparent latitude in the "noodle" structure observed in a number of active chimeric protein constructs might also be related to this potential flexibility. This whole fascinating subject is revisited in some detail in Section VI.

Wang and Giaever[114] also consider the configurational entropy of bringing together two widely separated sites. Association between proteins bound to sites far apart on the same DNA molecule is unfavorable *in vitro*. *In vivo*, however, the probability of the coalescence of the two is strongly influenced by the binding of other proteins to the intervening sequence. This binding might then yield a supra-nucleo-protein structure.[114] Such a structure might be involved in gene repression via a higher order structure of nucleosomes.

A critical factor in a control system may be the relatively constant, high effective concentrations of proteins bound to their spaced cognate sequences.[122] This relative concentration might be regulated temporally by external factors or by constraints on the flexibility of the intervening DNA. The supercoiling concentrates the cognate sequences and might increase the tendency for loop formation, while rigid structure might decrease it.[122,123] (For a recent review of DNA looping see Schleif.)[126]

Recently an intriguing twin-supercoiled domain model of transcription has been proposed.[318-321] According to this model, under some conditions, when the resistance to the rotational motion of the transcription complex is large, the advancing polymerase generates positive supercoils in the DNA template ahead of it and negative supercoils behind it. Mutual annihilation of the positively and negatively supercoiled regions may be prohibited if points on the DNA in both regions are anchored to large protein complexes. Such twin-supercoiled domains would also arise as the RNA polymerase, with its interacting regulatory proteins, tracks along the DNA. If these proteins are bound to a specific sequence of the DNA, the DNA loop generated by the tracking constitutes an independent domain. If the regulatory protein is bound to a point in front of the advancing polymerase (e.g., when the enhancer is downstream of the polymerase, as is the case in some immunoglobulin genes) the looped DNA would be positively supercoiled. If the regulatory proteins are bound upstream of the advancing polymerase, the growing loop is negatively supercoiled. Supercoils of one sense or another may accumulate if the rate at which they are formed is faster than those of the processes that remove them.

The twin-transcriptional loop model raises the possibility that transcription of one gene may regulate transcription of another gene located *in cis* through template supercoiling. The degree of supercoiling may activate or deactivate an adjacent gene. The degree (and sign) of supercoiling will also affect the local structure of the DNA. Regulatory sequence elements may or may not be recognized by their cognate protein factors.

Since in many cases DNA elements are recognized by their respective factors both *in vivo* and *in vitro* in different synthetic constructs, it appears that the DNA structural alteration is not very large.

V. DNA STRUCTURE: SOME GENERAL CONSIDERATIONS

A. DNA Helical Structure and Its Sequence Dependence

It is well established that DNA is a conformational polymorphic molecule. It can assume various conformations, depending on base sequence and environment.[127] Structural studies on double-helical DNA indicate that there are three major conformational families, the A, B, and Z forms of DNA. Environmental factors can induce transitions among different conformers.[128] Double-helical DNA, in contrast to residues in globular proteins, is highly exposed to solvent. Besides these major classes, DNA exhibits smaller conformational variations, depending on the sequence. For example, rotational relaxation times indicate that poly (dA·dT) is significantly more stable than random sequence DNA.[129,130] Such conformational variability also suggests that DNA flexibility should also be sequence dependent. Protein binding could be related to such effects. As an example, the co-crystal of EcoRI with DNA[131] shows distortions at several specific positions in the DNA when compared with the crystal structure of the DNA alone.[127]

Double-helical DNA plays the essential role of storing genetic information in a stable form, and its stability affects the control of gene expression.[132] The classic B-DNA consists of a right-handed double helix with two intertwined polynucleotide strands. The successive base pairs stack on top of each other with a relative rotation near 36°. The double-stranded structure is stabilized by complementary hydrogen bonding, by stacking of the flat surfaces of the base pairs, and by exposing the polar edges of phosphates to solvent. Although the structures of the DNA have been studied in some detail,[128,133-135] it is not well understood why the DNA is stabilized in a particular helical conformation. Hydrogen bonding between the bases leads to double-stranded forms, but not to a specific type of helix. Depending upon the base sequence, DNA exhibits local conformational variations,[134,135] varying from small base pair orientational changes to DNA bending.[136,137] The possible implications for biological function have recently been a focus of intensive research.[138-141]

Because of the very complex interactions in macromolecules, the physical reasons for this conformational polymorphism are not understood. To date, perhaps the most popular model of the sequence dependence of B-form variants was proposed by Calladine.[142] It assumes that the steric clash between base pairs may be responsible for the conformational variation. In order to identify the important interactions responsible for the stabilization of particular conformations of a complex macromolecule like the DNA, it is useful to first consider its separate components, base pairs and backbone. Interestingly, the basic double-helical structure of DNA and its local conformational variations may apparently result largely from the charge pattern in the base pairs and the electrostatic interactions between them,[132] with the backbone playing a rather passive role. From this point of view, the specific bases determine not only the sequence-dependent conformational variations, but also the overall structure.

Consider the two base pairs shown in Figure 2. The coordinate system we have used to describe their geometry is as follows. The X axis connects atoms C8 and C6 in the base pair and is oriented as shown. The Y axis points along the helix direction from the center of the line between C8 and C6.

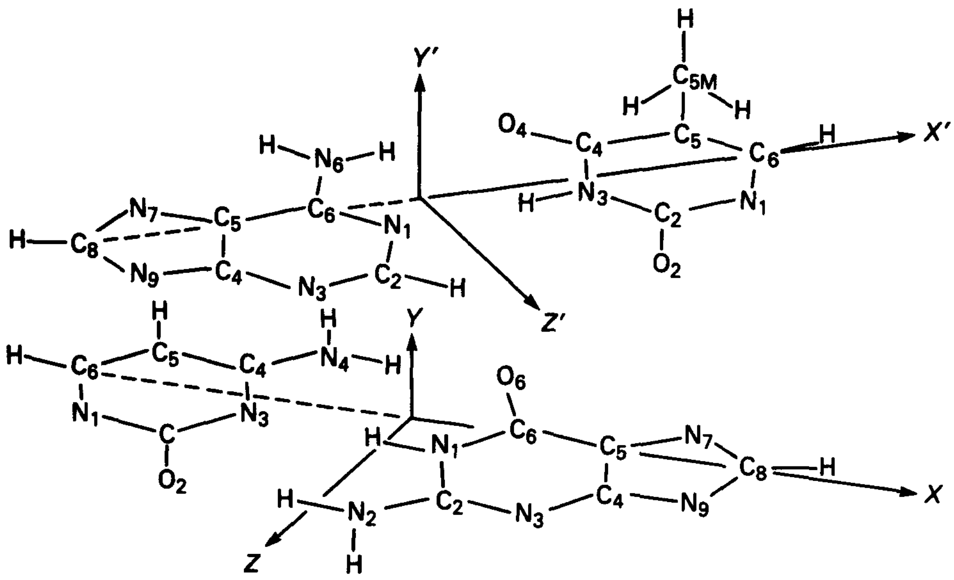

FIGURE 2. Coordinate system of two base pairs. Shown here are the CG and AT base pairs in a CA step. (From Sarai, A., Mazur, J., Nussinov, R., and Jernigan, R. L., *Biochemistry*, 27, 8499, 1988. With permission.)

Finally, the Z axis completes the right-handed orthogonal coordinate system, pointing toward the minor groove. Twist is the angle between X′ and X. Roll is defined as 90° — (angle between Z′ and Y). Thus, the sign of roll is defined as positive if the major groove is compressed. Tilt is similarly defined as 90° — (angle between X′ and Y). (See also Srinivasan et al.[143])

We also consider the intra-base pair degree of freedom, "propeller twist", in which matched bases rotate with respect to each other about the X axis. (For more detailed discussion see Sarai et al.[24,132])

DNA crystal structures have shown that values of twist, roll, tilt, and propeller twist are distributed about the average in specific patterns according to the actual sequence. Calladine[142] proposed that the local conformational variation is caused by steric clash between purines on opposite helix strands at adjacent base pairs. Most of the base pairs are positively propeller twisted (in screw direction) in order to improve the stacking interaction between adjacent base pairs. This brings the two purines in the YR step closer than van der Waals contact in the minor groove. Calladine[142] suggested that such clashes would be relieved by flattening the propeller twist, by rolling the base pair into the major groove, by sliding the base pair, and by decreasing twist. The clash can also occur between purines in the RY step in the major groove, but the clash is not as severe as in the minor groove of the YR step. Dickerson[144] quantified Calladine's model by introducing parameters according to the severity of the clash, and was able to reproduce many features of the dodecamer X-ray structure. Although the Calladine-Dickerson model elegantly explains the experimental results, it does not prove that the steric clash caused by the propeller twist is responsible for the conformational variation. The observed DNA structure could be determined by a balance among many attractive and repulsive forces acting between base pairs. By looking at the structure, it is difficult to see whether the clash is the cause or the result of such a balance. Although that model has been extended to A-DNA structures with partial success,[144] it does not distinguish between the two purines or between the two pyrimidines and fails to explain the DNA bending observed by gel electrophoresis[136,137] for various sequences containing A runs.

In order to test the importance of "propeller twist-induced steric clashes" in conformational variation, energy calculations were performed with artificially fixing the propeller twist at zero.[132] This eliminates steric clash between parallel, planar base pairs. Although the magnitude of the variation is slightly diminished in the absence of the propeller twist, calculations both with and without propeller twist show the same features of sequence-dependent variations observed in the dodecamer d(CGCGAATTCGCG) X-ray structure. The result clearly demonstrates that the steric clash induced by propeller twist is not essential to qualitatively account for the sequence-dependent conformational variation of DNA. Note in Table 1A that YR steps tend to be undertwisted. The CG step is undertwisted by more than 10° compared with a standard twist value of 36°. This is the same as in the dodecamer calculation, where twist at the CG step is 26 to 27°. This feature is due to eletrostatic interactions that also make important contributions to the variations of roll and tilt. Analysis of the atoms responsible for the conformational variations reveals that in the case of the CG step, C2 and N2 of G have unfavorable interactions, both hard-core repulsive and electrostatic interactions, with atoms of the adjacent G in the minor groove. On the other hand, O6 of G and C4 of C have favorable electrostatic interactions on the major groove side. These interactions contribute to rolling the base pairs in the CG step into the major groove, as observed in the crystal structure. These results suggest that electrostatic interactions among bases are a major driving force of the sequence-dependent conformational variation. Propeller twisting enhances this effect by positioning the relevant favored atom pairs closer and by introducing additional repulsive forces. Steric clash itself is not the largest force effecting conformational change, and certainly not the only force. The propeller twists should improve the stacking of adjacent bases in the same strand, but they are unlikely to be strong enough to cause large changes of roll and tilt, which might weaken the stacking stability. The propeller twist is unlikely to be the cause of all

Table 1a

Base step	Average twist angle	Average roll angle	Average tilt angle	Stacking energies
AA(TT)	38	0.53	0.48	−1.9
GG(CC)	40	1.9	−0.9	−3.1
AT	39	−0.3	0.04	−1.5
TA	24	5	−0.8	−0.9
GC	38	0.04	0.53	−3.1
CG	23	−8.4	1.6	−3.6
CA(TG)	29	6.4	−0.9	−1.9
AC(GT)	40	0.3	0.14	−1.3
CT(AG)	35	−0.4	0.6	−1.6
TC(GA)	38	−2.3	−0.14	−1.6
Mean	35	0.3	0	−2

Table 1b

Base step	Twist	Roll	Tilt	Propeller twist
AA	36	3.6	1.1	7.6
AG	36	1.4	1.4	5.1
AC	37	−1.2	1.1	7.6
AT	36	−1.2	0.2	9.1
GA	36	1.2	0.2	5.2
GG	37	−1.7	1.4	4.7
GC	36	−1.8	1.0	7.4
GT	36	1.3	−1.6	11.0
CA	35	4.4	0.5	9.2
CG	33	−7.9	−3.8	7.9
CC	37	−1.9	0.1	6.8
CT	36	−1.1	−0.9	5.2
TA	34	3.5	0.3	8.8
TG	35	3.3	−2.7	7.1
TC	36	−2.0	0.4	6.0
TT	37	3.4	−2.9	7.9

the conformational variation, as proposed previously,[142,144] but rather it may play a passive role in affecting the sequence-dependent conformational variation, enhancing the electrostatic effect.

The present calculation also provides useful information about the sequence-dependent flexibility of DNA (see Table 1A). The apparent difference between calculated and experimental twist angles at the AT step is probably due to the very flat potential surface of this step. The calculated conformational fluctuation about twist at this step is 15° RMS deviation from the mean angle. Thus the difference is well within the range of thermal fluctuations. Such a large flexibility of the AT step is consistent with the structure reported for the complex between EcoRI and DNA containing the AATT segment,[131] in which the AT step is strongly undertwisted by about 25°.[132]

Whereas Table 1A has been calculated without harmonic forces, they have been included in the calculations in Tables 1B and 1C. While the harmonic forces constrain the twist angles, making them closer to 36π, these forces are to some extent artificially incorporated. Tables 1B and 1C differ in the parameters used in the calculations (see the table caption).

B. DNA Structure in Upstream Regions

Tables 1A through 1C sum the preferred twist, roll, and tilt angles calculated with different parameters (see Table 1 caption) for all base pair steps.[24,132] Using the values in Tables 1B and 1C, tentative structures of the control region of the SV40 and of some of the SV40 composite elements have been computed (Plate 1*). It should be emphasized, however, that

Table 1c

Base step	Twist	Roll	Tilt	Propeller twist
AA	36	2.0	0.3	8.7
AG	36	1.0	0.7	5.6
AC	36	−0.9	0.6	8.3
AT	36	−1.0	0.2	10.0
GA	36	2.3	−0.4	7.1
GG	36	0.3	0.3	5.8
GC	36	−0.6	0.2	7.9
GT	36	−1.0	−1.0	11.0
CA	35	6.1	−0.8	9.8
CG	34	6.9	2.3	7.8
CC	36	0	−0.6	7.3
TA	35	5.0	−1.8	7.5
TC	36	−0.4	0.4	7.4
TT	36	0.3	−0.9	9.5

Note: Preferred angular parameters and calorimetric values for all base pair steps. The average twist, roll, and tilt angles are calculated at 300 K. Sequences in parentheses are the sequences of the matching strand. Although those represent the same structure, we calculate both complementary sequences. The numbers given here are the averages of the two. In Table 1a calculations are performed on the tetramer sequence NNMM by varying the twist (or roll, or tilt) at the central base step. Although the angles are mainly determined by the nearest-neighbor interactions, the values are affected slightly by the flanking sequences. This can be rectified by averaging over all possible flanking sequences (Tables 1b and 1c). The conformations of the 256 quartets have thus been calculated for Tables 1b and 1c. Table 1a is calculated without harmonic forces. In Tables 1b and 1c the harmonics are incorporated. In Table 1b $\epsilon = 2$, $\alpha = 0.75$. In Tables 1a and 1c $\epsilon = 5$, $\alpha = 1$. The twist values in 1a are taken from Reference 132, with $\alpha = 1.2$. Positive roll means major groove compression. Negative roll implies minor groove compression. Further details of the computations are given in References 24, 127, 132. The stacking energies in Table 1a are taken from Reference 301.

* Plates follow page 198.

even if our assumptions were true for short oligonucleotide structural calculations, for long DNA sequences we may have compounding of the errors.

Maroun and Olson[130] and Srinivasan[145] have developed a rigorous program that translates the DNA primary base sequence into a three-dimensional structure.[130,145] This program appears to be very successful in structural computations, in that there is good agreement between its predictions and the experimental results.[146,147] Using this method, the structure of the SV40 regulatory region has also been computed[322] (see Plate 2).

What aspects of the DNA regulatory sequence are unique and might aid in protein recognition and binding? The finding that oligo (dA)·oligo (dT) tracts cause DNA bending may be an example of a relevant feature. Indeed, such tracts have been located upstream of promoters,[138,148] inside promoters, in protein binding sites, and in origins of replication.[140,141,150] Methylation and ethylation interference experiments with the SV40 T-antigen[130] have shown that while the sequences flanking the oligo (dA)·oligo (dT) run are bound to this protein, the (dA)·(dT) run itself is not. Nonetheless, mutations in the (dA)·(dT) run impair the protein binding. Apparently, then, the tract plays a role in bringing the flanking sequences into proper juxtaposition, enabling T-antigen binding. Bent SV40 DNA has also been detected by Milton and Gesteland.[146] Inokuchi et al.[151] note that there are bends of the DNA helical axis at the upstream activation sites of α cell-specific genes in yeast. The effect of the three possible substitutions at nucleotide 5258 in the polyoma enhancer has also been studied.[152] The wild-type sequence appears to be more curved than the sequence with one of the point mutations. The two other point mutations apparently abolish DNA curvature.

Oligo (dG)·(dC) have also been detected upstream of genes.[153,154] In particular, the upstream sequence of the chicken adult β^A globin gene containing contiguous Gs has been well studied.[155] Fragments containing oligo (dG)·(dC) inserts migrate normally on polyacrylamide gels. Supercoiling, a prerequisite for efficient eukarytoic transcription, may alter the DNA conformations of these sequences. These can manifest via hypersensitivity to S1 nuclease[153,156] and increased reactivity with bromoacetaldehyde,[155] suggesting some single-strand characteristics. Short oligo (dG)·oligo (dC) are often present upstream of transcription-initiation sites (see Section VIII). Some of these are the GC boxes (the GGGCGG), shown to be the Sp1 protein binding sites (see Section III). In her recent review, Olson[152] (and references therein) notes that there are many examples of GC-rich curved duplexes. Furthermore, GC base pairs are also found with comparable, if not greater, frequency than AT residues at bends in oligomeric X-ray structures. These GC base pairs appear to be the major contributors to the perturbation found in computer simulations of DNA bending.[157] Short GC and AT runs are also thought to be involved in the folding of the DNA around the nucleosomes. GC base pairs might thus be, alongside TA, essential bending elements of curved DNA.[157] Studies of Milton and Gesteland[146] corroborate the latter point. Mutagenesis was used to identify the critical bases involved in bends in the SV40 DNA. One of their mutations (at position 110), interrupted a short run of Gs with no neighboring As. As a result, the DNA fragment migrated faster on the polyacrylamide gels, presumably due to straightening. Psoralen preferentially photoreacts with 5′TA and to a lesser extent with 5′AT. A psoralen-hypersensitive region was identified in the SV40.[158] This region extends from 150 bp on the late side of the enhancers to the early promoter boundary. This may also be indicative of a sequence-directed structural alteration.[158]

C. Some Clues to the Structure of the Promoter and Enhancer Elements

Although the detailed structure of the enhancer and promoter elements is unknown, there are already some data bearing on this subject. Phase-sensitive, two-dimensional nuclear Overhauser effect (NOE) spectra of $[d(GGTATACC)]_2$ containing the TATA sequence has been carried out.[159] In the structural model the GG and CC are in the B-DNA form, whereas the TATA moiety is in the wrinkled D (WD) form. The WD[160] structure is characterized by the following properties:[161] (1) 8 bp per helical turn; (2) a much narrower minor groove, creating a hydration tunnel; (3) significant cross-strand hydrophobic interactions; and (4) alternating torsion values at TA and AT steps.

The GGCCAATCT "CCAAT box" is composed of alternating two purines and two pyrimidines. The Calladine[142]-Dickerson[144] rules predict for such a sequence an interesting symmetric pattern of twist and roll angles.[162] The twist and roll angles of Table 1A[24,132] still predict a roughly symmetric pattern (Figures 3A and 3B).

The yeast analog of the CCAAT box is the UAS2 with the TG$^{C}_{A}$TTGGT consensus.[65,163] The yeast HAP2/HAP3 factors bind to this element (see Section VI). It is interesting to note that the ATTGG sequence, contained within this UAS2 consensus, is complementary to the CCAAT box consensus. Further, as discussed in Section VIII and shown in Figure 5E, ATTGG occurs frequently in higher eukaryotes at the same site as the CCAAT. The symmetric patterns of twist and roll angles of this yeast analog are presented in Figures 3C and 3D.

We repeated searches for consensus angular patterns in a variety of RNA polymerase II promoters — previously done with the Calladine[142]-Dickerson[144] values[164,165] — with those of Table 1A. These searches have been unsuccessful so far. Marked deviations from average B-DNA values were found only at recurring consensus *sequences*, i.e., mostly the CA initiator (at position +1) and the TATA (at −30).

Two additional considerations might help elucidate some aspects of the structure of the DNA and, in particular, might

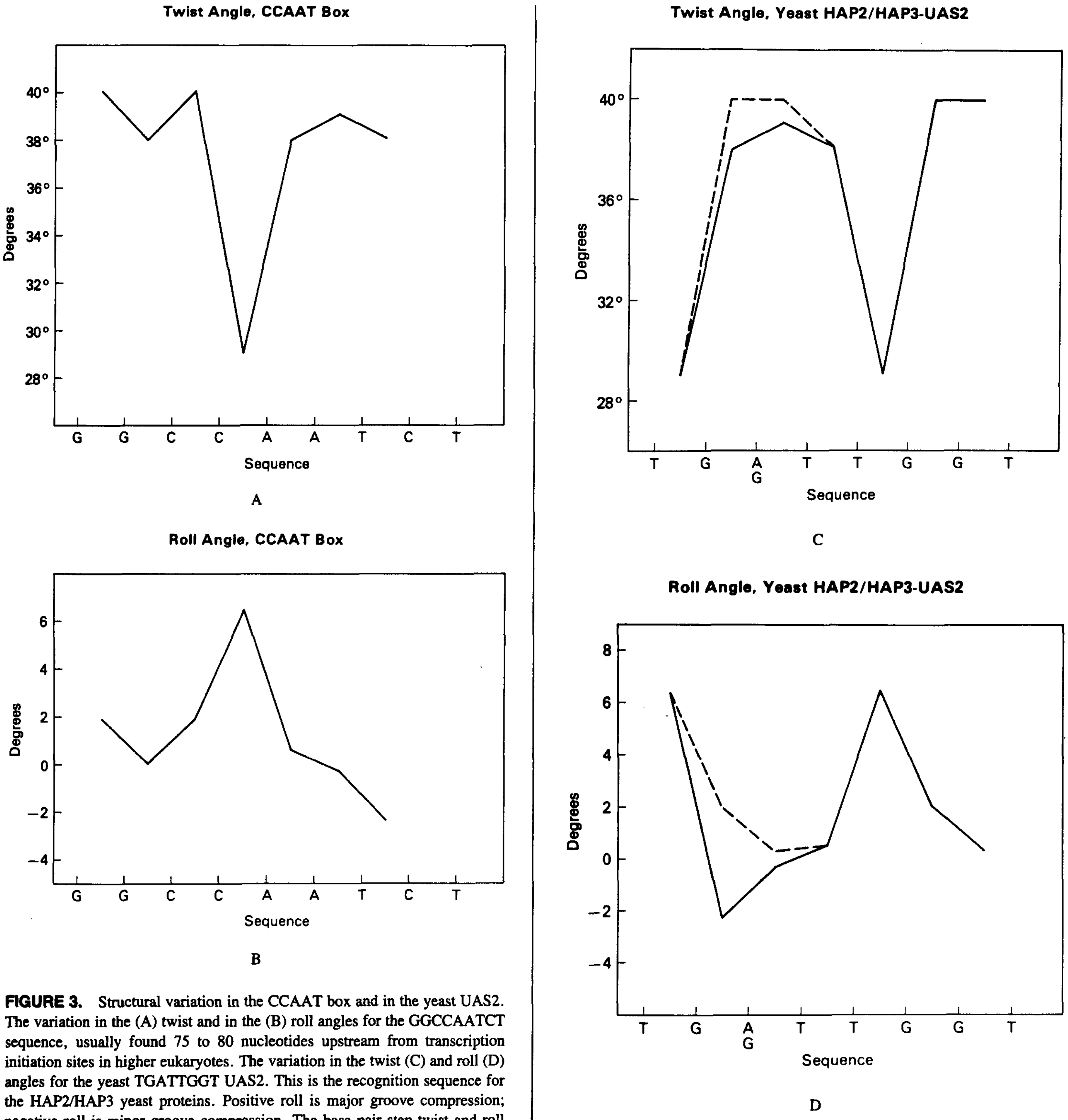

FIGURE 3. Structural variation in the CCAAT box and in the yeast UAS2. The variation in the (A) twist and in the (B) roll angles for the GGCCAATCT sequence, usually found 75 to 80 nucleotides upstream from transcription initiation sites in higher eukaryotes. The variation in the twist (C) and roll (D) angles for the yeast TGATTGGT UAS2. This is the recognition sequence for the HAP2/HAP3 yeast proteins. Positive roll is major groove compression; negative roll is minor groove compression. The base pair step twist and roll values are taken from Table 1a. The twist and roll patterns of the CCAAT box are symmetric. Note that the GGCCAATC sequence of higher eukaryotes is complementary to the yeast G $\frac{A}{G}$ TTGGT. It is intriguing to observe in Figure 5E that in higher eukaryotes the ATTGG is frequently in the same position as CCAAT. Note also that this sequence is composed of alternations of two purines and two pyrimidines, RRYYRRYY, already observed to yield symmetric angular variations.

aid in determining the positioning of nucleosomes and regulatory proteins. These are the groove width and the DNA anisotropic flexibility. Satchwell et al.[166] have noted that GGC/GCC, AAA/TTT, AAT/ATT, and AGC/GCT are periodically phased along nucleosomal eukaryotic DNA. Travers[167] and Travers and Klug[168] have suggested that such oligomers possess distinct conformations. In such conformations AT sequences have a narrow minor groove, whereas GC sequences possess a compressed major groove. There are a number of observations supporting this fact, including DNA crystal structures,[169-171] hydroxyl radical cutting patterns,[172] nucleosome positioning,[173] and NOE studies.[159] Further, in curved DNA, A/T (G/C) sequences tend to be positioned with compressed minor (major) grooves, respectively, facing the protein.[167,168] DNA bending will be in a direction further compressing the AT minor groove and the GC major groove. Thus, scanning upstream (or other) sequences, some predictions of the directionality of DNA bending may be made. In most DNA-protein complexes studied to date (e.g., nucleosomes, gyrase, EcoRI, λ-repressor, phage 434 repressor), the protein prefers to be positioned on the inside of the curve. This point cannot, however, be generalized. The DNAase I crystal structure[174] indicates that the enzyme prefers being positioned on the outside of the curve, with the DNA bent away from the protein. Possible implications for the protein-DNA interactions in regulatory regions are discussed in Sections VI and VII.

D. The Major and Minor Grooves as Carriers of Information

A basic important question is how does a specific regulatory protein or drug "know" where to bind to the DNA chain. The basic information is stored, after all, not in the sugar phosphate backbone, but in the ordering of the bases that are enclosed inside the double helix. In order to recognize different specific sequences, the protein (or drug) must exhibit a physical complementarity to a particular base sequence that it recognizes, either in shape, charge, polarity, or pattern of hydrogen bonding.[175] The base sequence can be sensed only by probes that fit within the major and minor grooves sensing the chemical features of the edges of the Watson-Crick base pairs (Figure 4). These features include the positioning of hydrogen donors and acceptors, the hydrophobic thymine methyl group in the

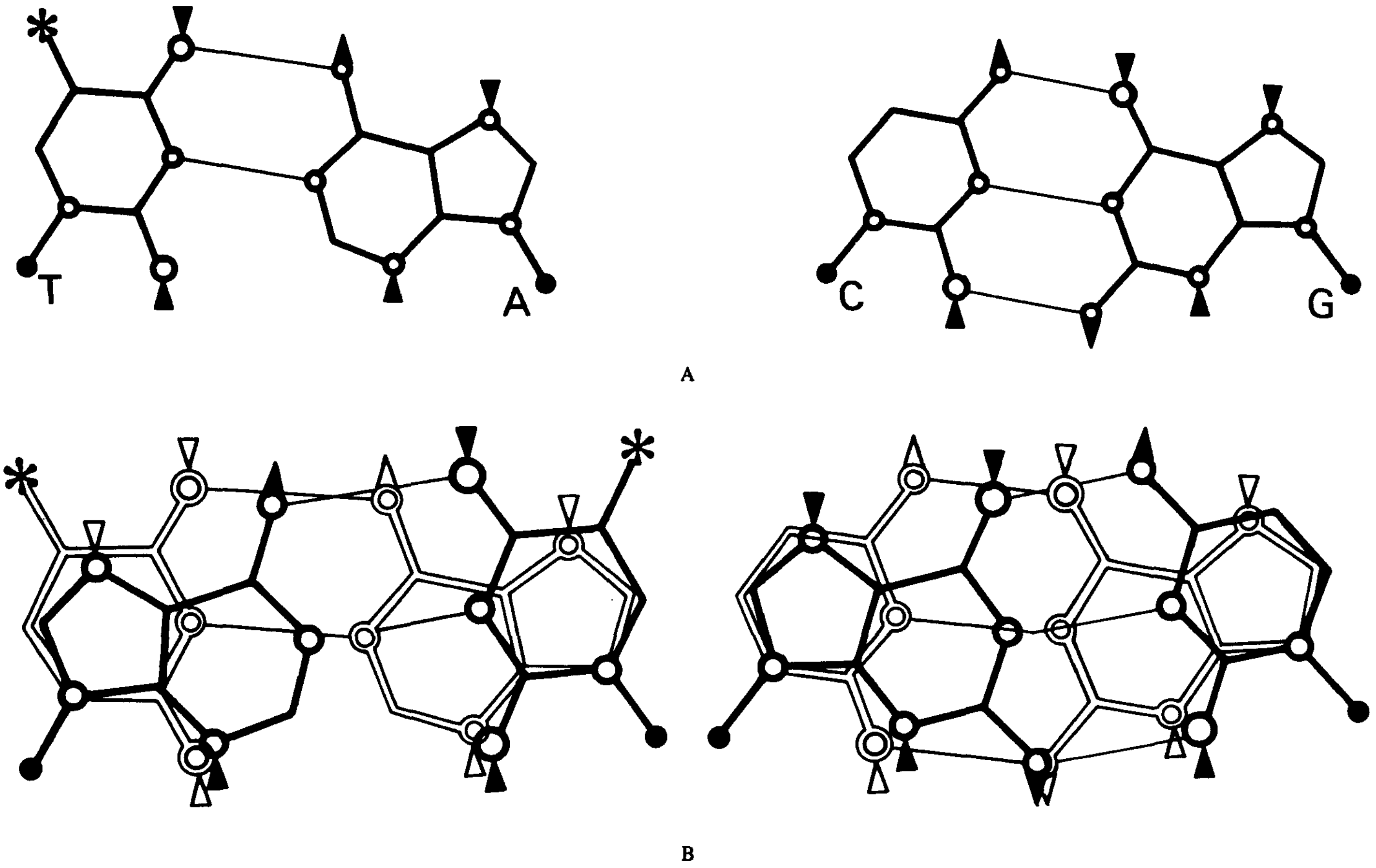

FIGURE 4. (A) T·A and C·G base pairs, showing potential hydrogen bond donors (arrowhead pointing away from base pair) and acceptors (arrowhead pointing toward base pair). Black dots are C1′ atoms of deoxyribose. Small open circles are nitrogens, and large open circles are oxygens. * = Methyl group of thymine. The upper edge of each base pair is the major groove edge, and the lower is the minor groove edge. (B) Reversals of base pairs, showing similarities or differences in hydrogen bonding positions. Left: T·A (open bonds) over A·T (solid bonds). Right: C·G (open) over G·C (solid). Note that hydrogen bonding positions nearly overlap in the minor groove, meaning that outside molecules cannot recognize base pair reversals. (From Dickerson, R. E., Kopka, M. L., and Pjura, P. E., *DNA-Ligand Interactions*, Guschlbauer, W. and Saenger, W., Eds., Plenum Press, NY, 1987, 45. With permission.)

major groove, and the bulky guanine NH_2 group in the minor groove.[175]

Recognizable information is thus stored in the major and minor grooves. Dickerson and colleagues[175] suggest that the major groove has about twice the information content of the minor groove. This rationale is as follows: A-T base pairs differ from G-C ones in either the major or the minor groove. However, as Figure 4 shows, A-T and its T-A reversal have different hydrogen donor-hydrogen acceptor patterns in the major groove. Also, the CH_3 occupies a different position. On the other hand, in the minor groove, A-T and T-A are indistinguishable. An analogous situation exists for G-C vs. C-G reversal. A regulatory molecule that senses the hydrogen bonding pattern in the minor groove cannot discriminate between A-T and T-A, or G-C and C-G. Given the groove size and its information content, it is not surprising that crystal structure analysis indicates that sequence-specific DNA binding proteins interact mainly with the major groove. In the EcoRI co-crystal, the minor groove is almost ignored.[130] The pattern of hydrogen bond donors and acceptors in the major and minor grooves is given in Table 2.

This, however, does not exclude a protein exploitation of the sequence-dependent geometry of the minor groove. Neither does it exclude a protein reading of the spatial phosphate backbone atom coordinates. The latter is also clearly determined by the geometry of the specific base sequence.[176]

In Section VI we discuss in detail some insights into protein-DNA interactions provided by the recent results. As described below, regulatory proteins apparently contain at least two modules. One is a sequence-specific DNA binding domain, and the second interacts with other regulatory proteins or with the RNA polymerase II.

VI. THE MODULAR NATURE OF THE TRANSCRIPTION FACTORS

A theme emerging from recent reports is that the mechanism of RNA polymerase II is remarkably conserved in eukaryotes.[52] There are several examples of sequence and functional similarity between yeast and mammalian transcriptional activators. The yeast transcriptional activator GCN4 uses the same DNA binding site as the chicken oncogene jun[16,22] and its mammalian AP1 homologue.[17] Transcriptional activators that bind to sites containing the CCAAT sequence are heteromeric complexes in both mammals and yeast (HAP2 and HAP3).[64,65,177] Interestingly, even the interactions between subunits of these proteins have been conserved. This is indicated by the fact that hybrid complexes containing a HAP2 (or HAP3) subunit and a mammalian CP1A (or CP1B, respectively) subunit still bind DNA specifically.[66] The mechanism of transcription activation by specific upstream DNA binding proteins has been conserved as well.[52] The yeast GAL4 activator stimulates transcription from mammalian promoters containing the GAL4 recognition sequence.[18,19] Moreover, even the *fos* oncogene, which apparently acts as a transcription activator, promotes transcription in yeast when it is fused to yeast DNA binding domain.[21] The next section focuses on various aspects of transcription regulation in yeast.

Transcription of the yeast GAL4 (galactose metabolizing) genes requires the GAL4 protein.[5] DNAase I footprinting and methylation protection experiments indicate that GAL4 binds to four sites in the upstream regulatory region of the GAL genes.[178,179] Comparison of the corresponding DNA sequences in these sites suggests that a 17-bp sequence of imperfect dyad symmetry is important for GAL4 recognition. The GCN4 protein is required for initiation of the transcription of many enzymes synthesizing amino acids during amino acid starvation.[5] GCN4 binds to specific sequence in the promoter regions of all co-regulated genes.[180,181] The GCN4 DNA binding element is a 9-bp palindrome, with one to two deviations from the consensus ATGACTCAT frequently observed.[5] Transcription of the yeast CYC1 gene, encoding iso -1- cytochrome C, is activated by two adjacent upstream activation sites, UAS1 and UAS2.[65,163] The activity of UAS2 depends on both the HAP2 and HAP3 proteins. Mutations of these proteins inactivate UAS2, suppressing the expression of genes involved in the respiratory metabolism.[65] It is worth noting that the common DNA recognition element, TG$^{G}_{A}$TTGGT, is homologous to the CCAAT box, recognized by higher cell transcription factors.[182,183]

Detailed analysis of transcription factors suggests that the two functions of DNA binding and transcriptional activation originate in two specific, physically separate regions of these proteins. Thus, analysis of truncated versions of the GAL4 and GCN4 indicates that the DNA binding domains are constrained within short regions of the proteins.[5] Also, for the GAL4, containing 881 amino acids, the N-terminal 73 amino acids are sufficient for specific DNA binding.[184] Likewise, the 60 C-terminal amino acids in the GCN4 281 amino acids long are sufficient for DNA binding.[5,8] While these truncated proteins bind to the DNA, neither activates transcription. In fact, the binding of these shortened proteins to their cognate DNA sequences even represses transcription,[8,184] perhaps by saturating these DNA elements. Brent and Ptashne[7] have clearly shown this physical separation of the functions responsible for DNA binding and transcription activation protein modules.[5] Specifically, they have constructed a hybrid protein in which the 73 N-terminal GAL4 DNA binding domain was replaced by the DNA binding domain of the *Escherichia coli* LexA repressor. When a LexA recognition sequence was fused to the GAL4 genes promoter, this hybrid protein activated transcription. In the hybrid, the DNA binding domain is contributed by the LexA protein and the activation domain by the remaining GAL4 domain. As expected from the modular nature of the protein, the hybrid did activate transcription. Hope and Struhl[8] have similarly shown that a LexA-GCN4 hybrid protein, in which the LexA DNA binding domain is fused near the N-terminus of GCN4, can stimulate transcription by binding to sites containing either the GCN4 or LexA DNA recognition sequences.[5] This suggests that the GCN4 transcription activation domain is functional whether a DNA binding domain is located at its natural C-terminal position or at the inverted N-terminal position.[5]

In another set of experiments, Struhl[15] shows that the jun oncoprotein, which causes sarcomas in chicken, efficiently activates transcription in yeast, either through its own or a heterologous DNA binding domain. The jun protein is probably the oncogenic version of the normal AP-1 transcription factor. jun and GCN4 bind to the same DNA consensus sequences[16] ATGA$^{C}_{G}$TCAT,[185] even though the DNA binding domains in jun and GCN4 have only 45% of their amino acid sequence in common. A hybrid protein containing the entire jun amino acid sequence fused to a LexA DNA binding domain activates transcription of yeast amino acid biosynthetic genes. In these genes, the LexA operator has been fused in the upstream region. (In fact, in terms of transcriptional activation function, the jun oncoprotein is nearly as functional in yeast cells as GCN4, a native yeast activator protein[15]). Thus, the modular nature of the protein transcription factors is clearly manifested and, furthermore, it allows elucidation of roles played by a variety of proteins.

The functioning of structurally altered regulatory proteins of RNA polymerase II-transcribed genes has been analyzed.[2,61] This identified a protein domain important for hormone binding, in addition to the DNA recognition and activation domains. The steroid hormone receptor proteins that bind to and activate

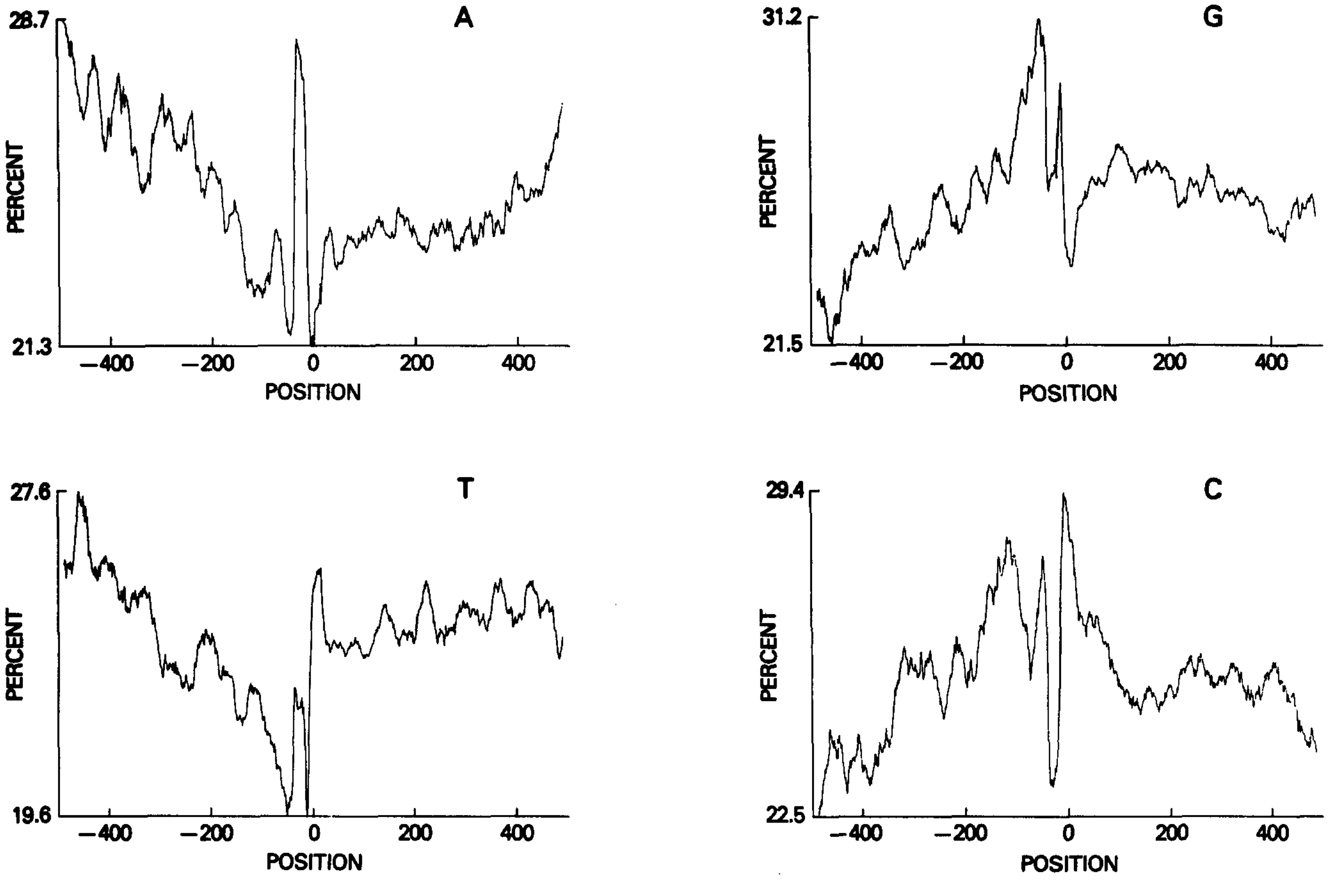

FIGURE 5. The frequencies of occurrence of (A) mononucleotides, (B) doublets, (C) triplets, (D) quartets, and (E) pentamers in 1641 eukaryotic sequences. The sequences are from the following GenBank categories: primates, rodents, other mammals, nonmammalian vertebrates, invertebrates, and viruses. The sequences have been aligned by their transcription initiation site (position 0). Negative values on the X axis are upstream positions. A window (of length 25 for mononucleotides, triplets, and quartets; length 15 for doublets; and length 50 for pentamers) has been shifted by one nucleotide at a time across the aligned sequences. The percent oligomer occurrence in the window is computed. Of particular interest is the G/C abundance upstream of the TATA box, and the complementary CCAAT and ATTGG sequences that peak at the same position in (E). ATTGG is also part of the UAS2 in yeast. (See also the Figure 3 caption.)

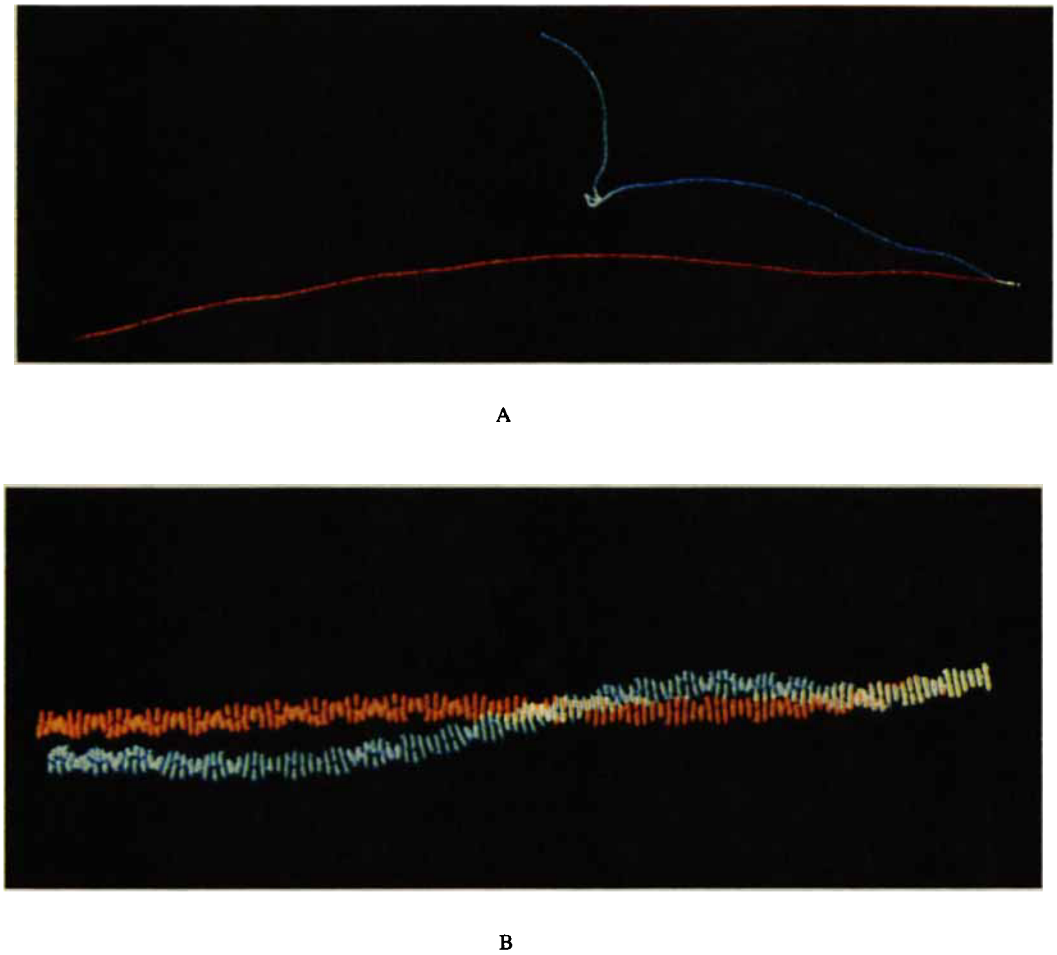

PLATE 1. Putative computer-generated structures of the SV40 control region. The structures have been generated using the angular values of Tables 1a and b. (A) The largest fragment portrayed here, nucleotides 5171-5243, 1-350; (B) T-antigen binding sites, nucleotides 5184-5243, 1-61; (C) T-antigen binding site I, nucleotides 5184-5209; (D) two 21 bp repeats, nucleotides 62-103; and (E) the 72 bp enhancer repeats, nucleotides 107-250. C and D are given in stereo. The GenBank numbering system is used. Two structures have been computed for each region: in blue $\epsilon = 2$, and $\alpha = 0.75$ (Table 1a); in red $\epsilon = 5$, and $\alpha = 1.0$, (Table 1b) where ϵ is the bulk (average) dielectric constant. The 5′ ends of every two corresponding structures of the same region have been superimposed on the Evans and Sutherland computer graphics system. As expected, with a smaller dielectric constant (ϵ), the structures are more bent. The computations have been done on the Cray. (See References 24, 127, and 132.)

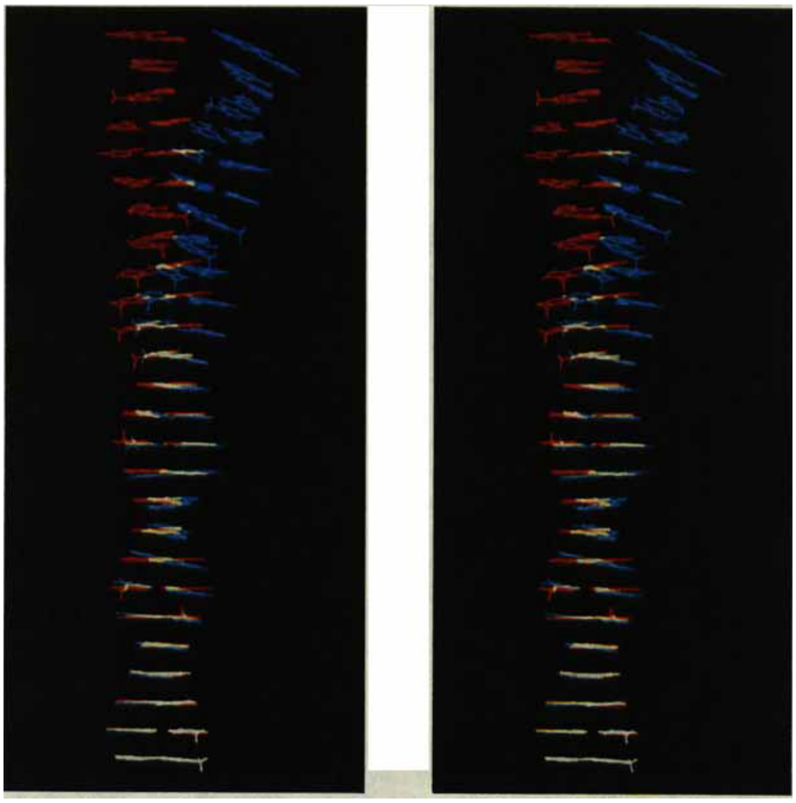

PLATE 1C

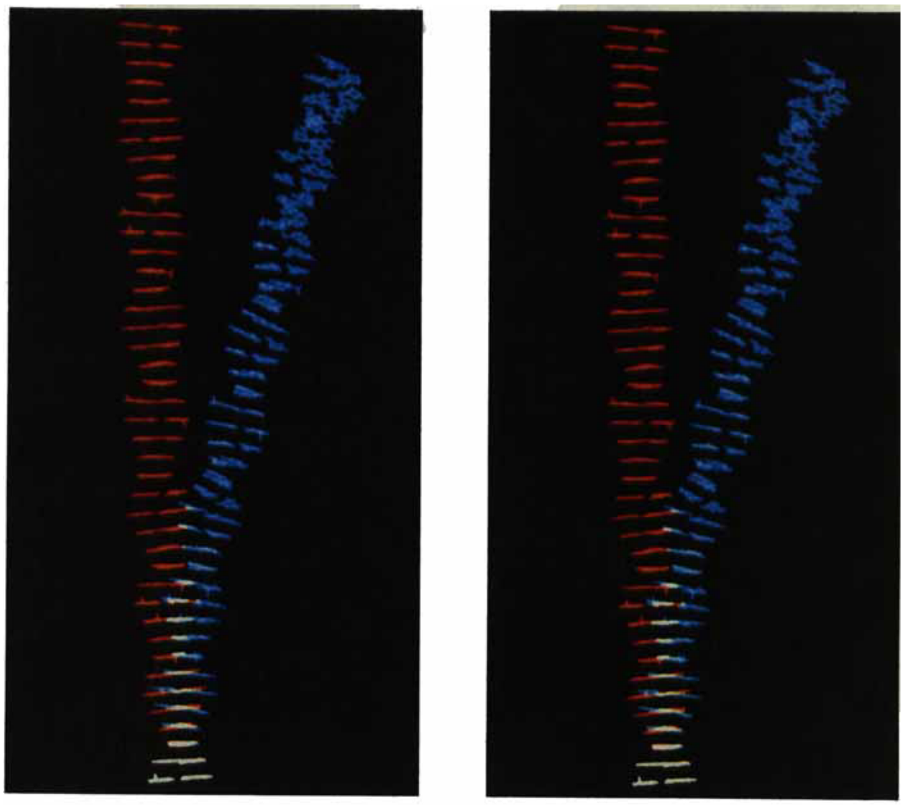

PLATE 1D

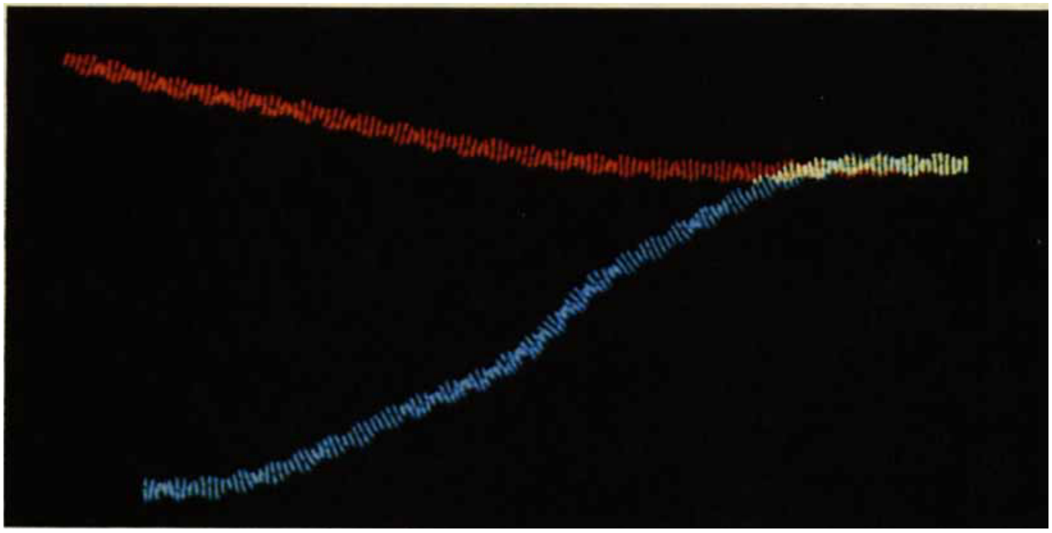

PLATE 1E

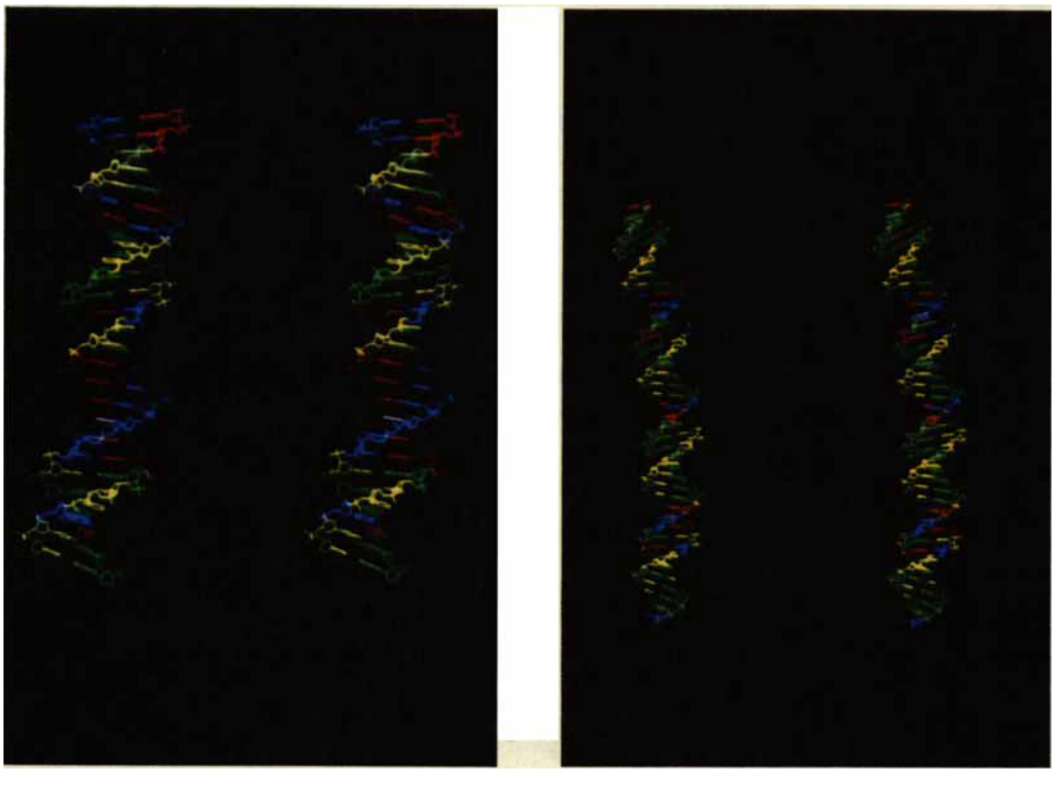

PLATE 2. Putative computer-generated stereo structures for the SV40 control region. These structures have been generated at Rutgers by M. Babcok and Drs. W. Olson and A. Srinivasan. Details of the computations are given in References 129, 130, 143, and 145. Although the methods of the calculations used here and for Plate 1 differ, there is some structural similarity. In our model (Plate 1) the harmonic forces make the backbone closer to ideal, with twist nearer 36°. (The harmonic forces incorporate the backbone effect on twist coordinates.) Here, it was assumed that twist is 36°, regardless of the sequence. (A) T-Antigen binding site I, nucleotides 5184-5209; (B) two 21 bp repeats, nucleotides 62-103. These structures correspond to Plates 1C and D, respectively. The color scheme is as follows: Adenines – red, Cytosine – yellow, Guanines – green, and Thymines – blue.

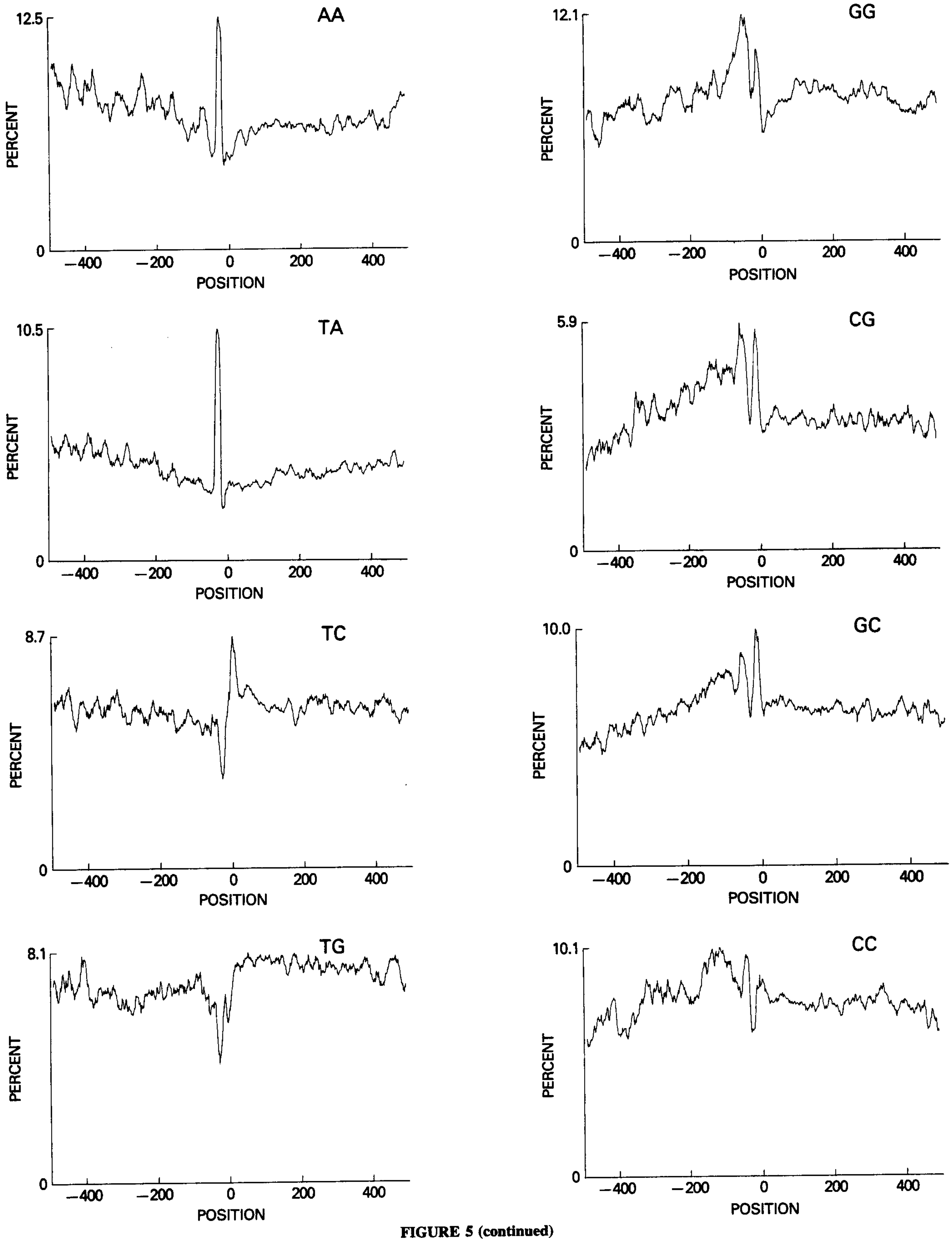

FIGURE 5 (continued)

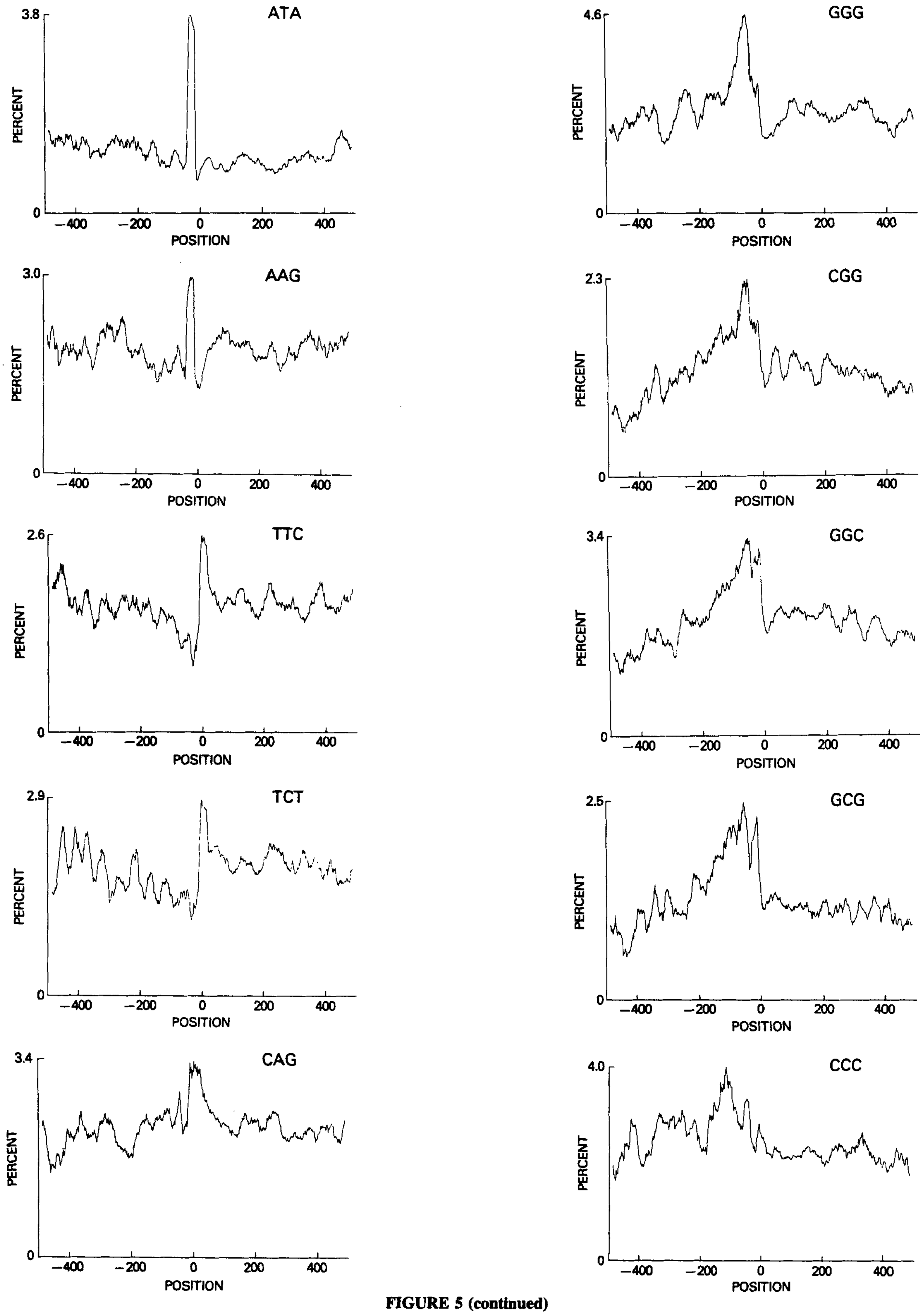

FIGURE 5 (continued)

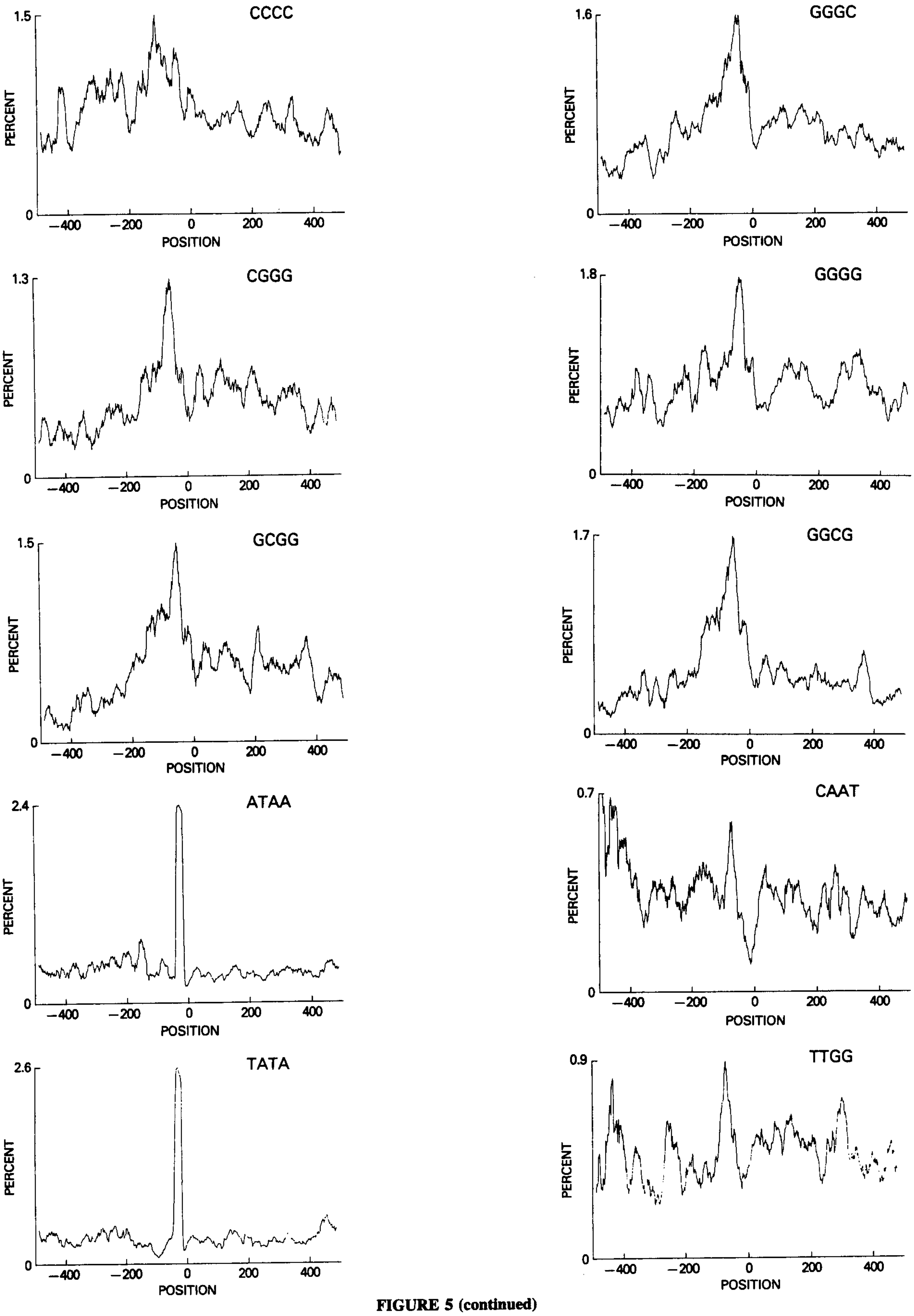

FIGURE 5 (continued)

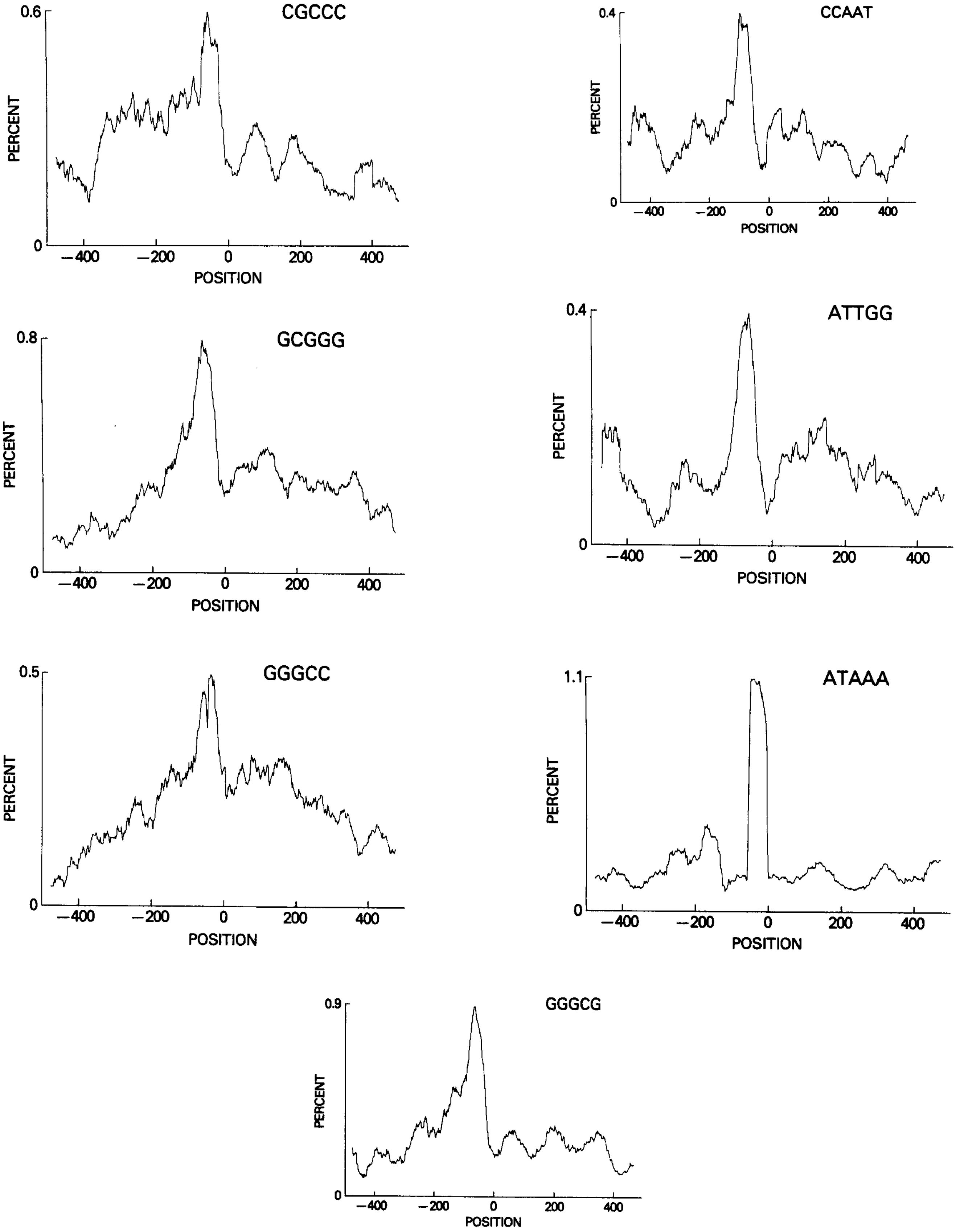

FIGURE 5 (continued)

Table 2

Base pair	Major groove				Minor groove		H-Bond
	H-Bond			Hydrophobic interaction			
C·G	D(NH_2)	A(0)	A(N)		A(0)	D(NH_2)	A(N)
G·C	A(N)	A(0)	D(NH_2)		A(N)	D(NH_2)	A(0)
A·T	A(N)	D(NH_2)	A(0)	CH_3	A(N)	A(0)	
T·A	A(0)	D(NH_2)	A(N)	CH_3	A(0)	A(N)	

Note: The hydrogen donors (D) and acceptors (A) in the major and minor grooves in the DNA. In parentheses is the group that can make hydrogen bonds with the side chains of the amino acids. These amino acids are present in the DNA binding domain of the protein. The methyl group is important for hydrophobic interactions. The order of the hydrogen donors and acceptors noted here is reading across the floor of the groove starting with the purine for A·T and G·C and with the pyrimidine for T·A and C·G.

steroid-inducible enhancers have been studied extensively.[2,20] It was found that the binding site for the receptor-ligand complex coincides with the DNA sequences required for steroid-dependent gene regulation.[2,20,186-197]

There is direct additional evidence that binding of the steroid receptor-ligand complex to the enhancer activates transcription.[2,198-200] Molecular cloning allowed identification of three domains in the proteins: one binds to the DNA, one binds to the steroid, and one that stimulates transcription.[198,199,201-205] The first and third apparently partially overlap. The modular organization of the steroid receptors has been studied by replacing the DNA binding region of the progesterone receptor with the corresponding region of the glucocorticoid receptor.[2] The hybrid protein binds to the glucocorticoid DNA recognition sequence and activates progesterone-dependent genes.[206] Thus, again, the DNA binding and transcription activation domains can be separated from the hormone binding domain.[2] Additional studies with truncated receptor molecules have indicated that steroid binding may induce a conformational change in the receptor. Truncated molecules containing the hormone binding domain interact with the DNA and activate transcription only in the presence of the hormone.[2]

Attempts to further characterize the conserved receptor domains, namely, the DNA binding domain and the functionally independent hormone binding domain, were recently made.[204,205,207] In these experiments, the human hormone binding domain has been removed. Although mutants carrying the defective gene retain only 5% of their wild-type receptor transcriptional activity, they still appear to bind efficiently to estrogen-responsive elements.[19,207] This suggests that the hormone binding domain plays an important role in the activation of the human estrogen receptor. Webster et al.[19] have recently shown that chimeras of the human estrogen receptor DNA binding domain and either GAL4 or GCN4 activating regions can activate a promoter region that is controlled by an estrogen-responsive enhancer. Godowski et al.[197] fused the DNA binding domain of the bacterial LexA repressor to the glucocorticoid receptor derivatives. These mutant proteins lack the original DNA binding domain. The resultant chimeric proteins activate transcription from promoters linked to the LexA operator. Interestingly, regardless of the exact positioning of the LexA DNA binding domain within the hybrid protein, hormone binding was still required for transcription activation. It thus appears that the hormone binding domain in the glucocorticoid receptor can act as a strong — and reversible — inhibitor of receptor activity. This reversible inhibition mechanism is independent of the DNA binding domain.[197] Further, these results also show that the receptor contains at least one activation domain in addition to the one overlapping the DNA binding domain. Apparently, this domain occupies a region near the receptor amino terminus.[197]

As noted earlier, the HAP2/HAP3 proteins regulate transcription of the yeast CYC1 gene by binding to the upstream activation site UAS2. The activity of UAS2 requires the presence of both HAP2 and HAP3. The latter bind to their cognate DNA element in an interdependent manner.[65] Tagging of the HAP2 and HAP3 can be achieved by gene fusion to LexA and β-galactosidase, respectively. This allows following the mechanism of HAP2/HAP3 binding. These experiments show that these two proteins associate in the absence of the DNA to form a multisubunit activation complex. HAP2/HAP3 are analogs to the CCAAT binding proteins in higher cells. Three distinct proteins have been identified by Chodosh et al.,[64] CP1, CP2, and NF-1. All three appear to recognize their binding sites with highest affinity when assembled into a multisubunit complex. Like the HAP2/HAP3 complex, the CP1, CP2, and NF-1 complex is also composed of heterologous subunits.

What is the functional significance of heteromeric DNA-specific binding proteins? A greater variety of regulatory options might be available to a protein composed of two or more subunits that are independently regulated. Moreover, the interchange of subunits among such proteins is likely to provide additional combinatorial complexity to gene regulation.[64] Are large gene families in eukaryotes regulated by complexes with nonidentical subunits?[65] There is some preliminary evidence that points in this direction.[208-210]

A. Activation Domains and Protein-Protein Interactions

The possibility of exchanging transcriptional activation domains between different protein factors without loss of activity suggested that these domains must share common characteristics. The amino acid sequences of these domains vary from one factor to another. There is, however, one common feature — all domains possess an excess of acidic amino acids, resulting in a new negative charge. Mutational studies of the activator domain suggest that its polypeptide is conformationally ill defined, and hence the name "negative noodle". The "noodle" can function almost irrespective of sequence, provided only that there is sufficient excess of acidic residues.[6] The first studies of this type have been performed with the GAL4 and GCN4[5,8,211] transcriptional activators. Portions of the GCN4 have been found to be equally capable of activating the transcription normally done by GCN4. No significant sequence homology has been detected between the GAL4 and GCN4 activation domains. Moreover, acidity is the only noticeable common characteristic of these domains and of random *E. coli* polypeptides fused to GAL4 DNA binding domain. These random chimeric proteins also displayed moderate transcriptional activation in yeast.[212] Furthermore, the strength of the activation is correlated with the preponderance of negative charge;[6,213] mutations of a GAL4 derivative that increase or decrease negative charge usually increase or decrease activation, respectively.[4,213] In addition, the vertebrate jun oncoprotein activates transcription in yeast through its acidic domain.[4,15] The activation of GAL4 through random fusion of *E. coli* segments encoding acidic amino acids may suggest that any negatively charged polypeptide would do.[212] These results, as well as similar studies of the GCN4,[5,8] jun,[15] the herpes simplex virus VP16[214] that activates transcription of immediate early genes, and additional observations (e.g., References 18,19,64,66,197) support the view that the conformation of these "negative noodles" is ill defined.[6] However, there are also data suggesting that the *E. coli* peptides might not consist of independent activating regions. Rather, these regions may be inactive or partially active segments that act in combination with otherwise cryptic GAL4 acidic regions.[15] This apparent need for a "stub" of eukaryotic activator may suggest that some recurring protein motif might nevertheless be required within the "negative noodle".

Intriguing results have recently been presented by Hope et al.[4] These authors have investigated the structure and function of the GCN4 activation region by systematic, high-resolution deletion analysis and by proteolysis of the wild-type and mutated GCN4 proteins. Various deletions between the activating coding region and the DNA binding domains were made. They found that the smallest functional unit consists of 36 amino acids from the acidic region fused directly to the DNA binding domain (containing 89 amino acids). Thus, there is no need for a large flexible region between the transcription activation and DNA binding domain.[4] Internal deletions of gradually increasing lengths did not reveal a position where there is a sudden complete loss of activity. Rather, a series of small stepwise reductions in activity was observed.[4] This can be contrasted with analogous studies on enzymes, where deletions, including the active site, would most likely result in a drastic drop in activity. The direct correlation between the progressive drop in activity and the step-wise deletions provides additional evidence that transcription activation regions do not have a defined tertiary structure. Studies of the spacing between the activation region and the DNA binding domain indicated that the distance between the domains is unimportant.[4] This provides direct evidence that the activation region encodes a domain that is independent from the DNA binding region. Hope et al.[4] show that the GCN4 activity appears to be related to the size of the transcriptional activation region that remains after the deletions. They have not detected a single case where a shorter region activates transcription more efficiently than a longer one. The direct correlation between the length of GCN4 activation region and transcription indicates a repeating structure with additively enhanced action. It has been suggested that such a structure is an amphipathic helix with acidic and hydrophobic residues along different faces.[4] The activation region might be a dimer of two intertwined amphipathic α-helices. Their hydrophobic surfaces would face each other and their negatively charged surfaces would be exposed to the solvent,[4] available for interaction with the RNA polymerase II.[6]

The herpes simplex protein, VP16 and the chimeric GAL4-VP16 construct, containing the GAL4 DNA binding domain and the VP16 activating region, have recently been shown to be particularly powerful activators.[214] Interestingly, in the absence of the GAL4 DNA binding element, this protein, like the native VP16, is a very potent inhibitor of transcription. This is probably the result of its avid binding to the transcription initiation complex, RNA polymerase II, TATA factor, and possibly additional factors. It would be interesting to find out what makes it such a strong activator. Does it, too, contain possible amphipathic helices?

The above amphipathic model proposed to explain the GCN4 results[4] also fits nicely with available data on the RNA polymerase II. The carboxyl-terminal domains of RNA polymerase II large subunit of yeast and mouse are known to contain 27 and 52 heptapeptide repeats, respectively.[6,124,125] Comparison of these and analogous RNA polymerase II domains from additional organisms yielded the canonical heptanucleotide sequences, [tyr — ser — pro — thr — ser — pro — ser]$_n$. The carboxyl-terminal heptanucleotide repeat was termed CT7n by Sigler.[6] Sigler suggests that the large number of the hydroxyl residues implies that this domain projects into the solvent. The proline residues may help in stabilizing the secondary structure of this arm. There is evidence suggesting the CT7n is not required for enzymatic (catalytic) activity in minimal polymerase II systems.[215] There is also evidence that beyond a minimal number of repeating units, its size and exact sequence is unimportant.[216,217] The carboxylates of the acidic, "negative noodle" might, then, form strong hydrogen bonds with the projected "blob" of hydroxyl groups of CT7n.[6] In this way the richness of the potential hydrogen-bonded contacts of the CT7n may allow serving many different transcriptional activation factors and systems. After transcription initiation, release of the polymerase from the activation factors might occur via phosphorylation of the hydroxyl groups in the CT7n domain. Some further potential implications of the interactions of the acidic domain of the transcriptional activators with the RNA polymerase CT7n domain have been discussed in Section IV.

A recent analysis[216] of Sp1 revealed multiple transcriptional activation domains, including a novel glutamine-rich activation motif. With the exception of one domain (domain D, located at the extreme carboxyl-terminal end of the protein), the other transcriptional domains of Sp1 apparently operate independently of one another. The activity of these domains is not critically dependent upon their spacing from each other or from the DNA-binding domain. A possible reason for these multiple independent transcriptional-activation domains is that they may supply the Sp1 with several options for protein-protein interactions at different distances and orientations from the DNA binding domain. This creates greater flexibility in the binding options of the general transcription machinery to the Sp1.[218] The activation module is thus also divided into several modules. This further increases the number of ways in which the RNA polymerase II DNA elements-protein factors can be regulated.

B. Sequence and Structure Motifs in the DNA Binding Protein Domains

An essential part of gene expression and regulation is the binding of a regulatory protein to its recognition DNA element. Embedded in the structure of such proteins is a domain, or "motif", that serves for binding to DNA.[219] Proteins that regulate the various functions of the DNA bind strongly to certain DNA sequences. The sequence specific DNA binding proteins that have been studied most extensively are the gene-activator and repressor proteins of bacteriophages. X-ray crystallographic studies have revealed a structural motif common to several of these proteins.[220,221] This motif is characterized by two α-helices juxtaposed at roughly right angles to each other by a turn of four amino acids. It is termed the "helix-turn-helix" motif.

Undoubtedly, X-ray crystallographic studies of proteins bound to their cognate DNA oligomers are highly desirable. In particular, we are interested in information pertaining to atomic interactions between the protein and the DNA. Since, however, a crystallographic study is a tedious and slow procedure, additional methods yielding valuable insights into the nature of the protein-DNA interactions are frequently used. At least for the λ repressor-operator, there is a good agreement between the recent crystal structure[222] and the results obtained using the other chemical, biochemical, and genetic methods. Ethylation interference experiments are commonly used. These experiments implicate the phosphates involved in the contacts between the protein and the DNA. The pattern of hydroxyl radical cutting is also instructive.[223] The small hydroxyl radical precisely marks the stretch of DNA covered by the protein. This contrasts with the much larger DNAase I endonuclease itself. In the latter case, a longer piece of DNA would appear to be contacted by the protein, owing to an excluded volume between the DNAase I and the analyzed protein.

Protection from chemical methylation is another approach yielding valuable information. The protected DNA region is inferred to be bound to the protein factor. Thus, protection of guanines in the major groove implies major groove protein-DNA contact. Methylation of adenines N3 minor groove implies no protein minor groove contacts. That last effect has been clearly shown by the λ repressor-operator.[221] Compairson of the sequences to which the protein binds indicates the conserved base pairs. These pairs are likely to be important for the protein-DNA interaction. The way in which single amino acid mutations in the proteins affect the binding of the protein to its cognate DNA sequence also provides useful information.

The results obtained using such methods aid in the positioning of the protein on the DNA. One should nonetheless avoid overinterpreting them in detailed (computerized) model building. In the case of the λ-repressor-operator, there is good agreement between the co-crystal and models incorporating the results of these chemical, biochemical, and genetic studies. Probably this agreement stems from the fact that, as seen in the co-crystal of the λ-operator DNA, in this case the DNA is quite straight and B-like. That, however, might not always be the case. There is great variability of possible DNA structures. In particular, DNA flexibility allows the observed large, protein-induced changes in the various DNA angular (twist, roll, etc.) parameters. Several recent examples illustrate this point. In the phage 434 operator-repressor co-crystal the DNA is bent, with marked variation in twist, and the minor groove is compressed in the center.[224] The 434 repressor bends the DNA to create a circular arc of 65 Å. The bending, however, is not smooth. The DNA is relatively straight in the middle of the operator and bends on each of its sides. This permits contacts between the sugar-phosphate backbone and the amino terminus of one of the α helices of the repressor. The minor groove is compressed in the center and gradually widens toward both ends. The sugar-phosphate backbone on either side of the minor groove interacts with the loops between two α helices. Finally, the center of the operator is overwound by several degrees, and its ends are underwound.[224] Protein-induced irregularity of the DNA helix is also observed in the DNAase I co-crystal.[174] Here the 14-mer DNA has a B-type conformation, but with substantial distortion of both local and overall helix parameters. This is induced mainly by tyrosine and arginine interacting in the unusually 3-Å wide minor groove. This widening is coupled to a 21.5-Å bend of the DNA *away* from the bound protein, into the major groove. In another case, the EcoRI restriction enzyme bends the DNA by about 30° into the minor groove.[131] As in the 434 operator, in the nucleosome the DNA is bent in a nonsmooth manner with a 43-Å radius around the histone octamer.[225] Thus, in contrast to the remarkably constant protein structures, the DNA often appears to be significantly distorted. This reflects the intrinsic flexibility of the DNA and hence highlights the difficulty in predicting its interaction with a protein.[174]

While crystals yield the most detailed structural information, it is doubtful whether or not this information truly reflects the structure existing in solution. Thus, the triple, bifurcated hydrogen bonds observed in crystal structures of oligomers containing $(dA)_n$·$(dT)_n$ flanked by mixed sequences[169,170] may exist only in the crystals. No eukaryotic activator protein together with its cognate DNA recognition sequence has been crystallized to date. Nevertheless, the principles of the protein-DNA interaction are the same as those observed in prokaryotic protein-DNA complexes. Analysis of these and of the amino acid sequences of some eukaryotic transcription activators has suggested the existence of several motifs in protein-DNA interactions.

1. Helix-Turn-Helix Motif

Several bacterial repressor proteins, the trp, λ repressor, phage 434 cro and repressor have been studied to date. All these proteins display the so-called "helix-turn-helix" motif in their binding to the DNA.[176,222,224,226] As noted above, this motif consists of two successive α helices. The second α helix is located in the major groove of the DNA. The exact geometry of binding is, however, variable.[227] This is the case even for the 434 repressor and cro,[224,226,227] two proteins that have 50% amino acid sequence identity, very similar three-dimensional structures, and which bind to the same DNA operator sequence. In the Cro, CAP, and λ repressor proteins, this motif binds to the DNA via hydrogen bonding and van der Waals interactions between the side-chains of the proteins and the edges of the base pairs that are exposed in the major groove. This is corroborated by the new crystal structures.[176,222,224,226,227] It has, however, become clear that in addition to the sequence-specific interactions with the bases in the major groove, there are also multiple contacts between the protein and the phosphate backbone. Whereas the first class of sequence-specific interactions discriminate between different DNA sequences, the other interactions with the backbone enhance the complementarity at the protein-DNA interface and increase the overall binding affinity.

The recently obtained crystal structure of the trp repressor[176] provides an example of an indirect "readout" of the DNA sequence-(structure) by the protein. Here the bases are hydrogen bonded to the protein via water molecules. The operator sequence is apparently recognized by the changes it induces in the DNA backbone. It is not clear, however, whether this distortion of the backbone is large enough to allow sequence-specific recognition by the trp repressor.

Recent studies have indicated that homoeo-box proteins are sequence-specific transcription factors.[228-232] The amino acid sequence of the homoeo motif of unc-86 (a *Drosophila* homoeo protein) is similar to those of three transcription factors (Oct-2, Pit-1, and Oct-1). The recurring homoeo motif is apparently a DNA binding domain. Amino acid sequences that could form a helix-turn-helix motif are also present in the homoeo domain of several homoeo proteins.[229-232] This has been corroborated recently by the circular dichroism spectra of the homoeo-domain peptide, showing the presence of a significant amount of α-helical structure.[233] This *Drosophila* homoeo domain peptide is a sequence-specific DNA binding protein, recognizing the TAA repeat. It is interesting to note that unc-86, Oct-2, Oct-1, and Pit-1 have an additional homologous region upstream from the homoeo domain.

2. Zinc Fingers

The zinc finger was originally discovered repeated in tandem in the *Xenopus* transcription factor IIIA (TFIIIA).[234] Since then it has been found in many regulatory proteins. Zinc fingers serve as modules for the building up of a specific nucleic acid binding domain.[219] In the TFIIIA, each finger contains two closely spaced cysteines, followed by two histidines.[235] Biophysical studies indicate that a zinc ion is tetrahedrally coordinated by the cysteine-histidine motif.[236,237] This coordination imparts both stability and DNA sequence specificity.[221] In addition to the two cysteines and two histidines, two aromatic residues and one leucine are conserved as well.

There are at least two classes of finger proteins.[237] Both have nucleic acid sequence-specific recognition capabilities. In the first C_2H_2 (two cysteines and two histidines) class, the pairs of cysteines and histidines are separated by a loop of 12 amino acids. Recently, a complex finger structure has been suggested[239] from structures of known metalloproteins.[238] There are at least two fingers separated by seven or eight amino acids. The transcription factor IIIA and the Sp1 protein[240] that binds to the GGGCGG repeat are examples of this class.[241] The second class, C_x proteins, have a variable number of conserved cysteines. Thus, in the yeast transcriptional activator GAL4 there is a cluster of six invariant cysteines. The steroid receptors contain two fingers, encoded by different exons,[242] with four and five cysteines, respectively.

In the glucocorticoid receptor the location of the metal-binding finger has been defined by site-directed mutagenesis.[243] In fingers with more than four cysteines, multiple ions may share the cysteines[237,244] — a type of interaction observed in metallothionein.[244]

Direct evidence for the existence of zinc fingers in transcription factors is lacking. For several proteins there are, however, indications that DNA binding is conferred by the zinc finger.[237] In a "finger swap", the putative finger region of the estrogen receptor has been replaced by that of the glucocorticoid receptor, and the DNA binding specificity of both has been altered.[206] The Sp1 binds DNA with an isolated finger region.[240]

Frankel and Pabo[245] have cautioned that the definition of "zinc finger" domains has become too loose. Strictly speaking, zinc fingers must contain two cysteines, two histidines, two aromatic residues, and one leucine with proper spacing. DNA-binding proteins containing several cysteines, but not the other finger components, might not have the characteristic three-dimensional structure of a "finger". While the cysteine residues in these nuclear receptors (e.g., adenovirus E1A, GAL4) probably still bind zinc, this cysteine-containing domain might be involved in protein-protein interaction, rather than in DNA recognition.

3. The EcoRI Restriction Endonuclease

Although the binding of an enzyme that cleaves the DNA backbone might differ from that of a regulatory protein, it nevertheless is instructive to inspect its mode of interaction with the DNA. In the EcoRI co-crystal,[131] the two α helices in each subunit touching the DNA are part of a domain in which several α helices surround a five-stranded β sheet.[246] The major α helical regions are oriented to allow their electropositive amino ends to point at the negatively charged DNA, with the negatively charged glutamates also forming hydrogen bonds with several of the same base pairs. Two stranded β sheets wrap around the DNA in the major groove. The EcoRI, which binds specifically only to its GAATTC recognition sequence, has two hydrogen bonds with each base pair. Such a network of hydrogen bonds is not observed with the other, less specific, bacterial repressors.

4. Minor Groove Interactions

As noted above, the major groove is information rich and is thus a prime candidate for DNA-specific protein interaction. Nevertheless, minor groove interactions have been observed as well. A study of the structure of DNAase I crystallized with a self-complementary octa-nucleotide has been reported recently.[174] It shows that an exposed loop of DNAase I binds in the minor groove of B-type DNA. Hydrogen bondings with the backbone are observed as well. While the cutting rates of bovine pancreatic deoxyribonuclease I vary along a given DNA sequence, indicating that the enzyme recognizes sequence-dependent structural variations,[247,248] it is still not a sequence-specific DNA binding protein. It can thus be argued that only a relatively inspecific protein binds to the information-poor minor groove (see Section V). The sequence-dependent variation in the minor (and major) groove width is also likely to play a role. A model of sequence-specific DNA binding in the minor groove has recently been suggested for the λ integration host factor[249] and the HU bacterial chromosomal protein.

5. Steering the Protein into the DNA Grooves

Recent studies have shown that in addition to the direct atomic interactions between the amino acid side chains and the DNA base pairs, there is a limited number of structural motifs that steer the appropriate amino acid side chains of the protein into the DNA grooves.[221] These structural motifs are composed of amino acids that may[222] or may not[221] make direct contact with the DNA. In the former case they interact with the sugar-phosphate backbone. In the latter case they form what Landschulz et al.[221] term "three-dimensional scaffolds" that match the contour of the DNA. The amino acid sequences of these scaffolds in several proteins are similar.[221,238]

The co-crystals of the λ repressor, the phage 434 repressor, and the trp[176,222,224,226,250] have an extensive network of interactions between the proteins and the DNA sugar-phosphate backbone. These contacts aid in the precise positioning of the helix-turn-helix motif on the DNA. Interestingly, most of these contacts, at least for the λ and phage 434 repressors, involve polar side chains or peptide –NH groups.[222] Lysine and arginine are not used to make many of the phosphate contacts. Alternatively, it is possible that hydrogen bonds with polar side chains may have a greater directional specificity when compared with the salt bridges formed with the charged groups of lysine and arginine.[222] Surprisingly, shorter polar side chains are preferred. Jordan and Pabo[222] hypothesize that since the shorter side chains are less flexible they may help in positioning the repressor more precisely. In any case, it appears that these contacts enhance the specificity by positioning the "correct" residues in contact with the "correct" bases. In the λ repressor, the amino acids contacting the backbone are found at the beginning, middle, and end of the helix-turn-helix motif. These precise structural requirements may help explain the strict conservation of this helix-turn-helix motif.[222] Residues from the carboxyl end of helix 1 and the amino terminal of helix 4 also touch the backbone. This may explain the structural homology in these regions between proteins containing the motif.[222,251,252]

A new motif involving protein-protein interactions between paired helices has been described recently. It may create three-dimensional scaffolds that match the contour of the DNA.[221] This intriguing motif consists of a periodic repetition of leucine residues. The leucines appear to extend from an unusually long α helix. Landschulz et al.[221] propose that the leucine side chains of one helix interdigitate with those of a matching helix from a second polypeptide, forming a stable, noncovalent linkage, termed the "leucine zipper". In this class of proteins such paired helices may play a crucial role in arranging the contact surface for sequence-specific interaction with the DNA.

The periodic array of leucines was first noted in a CCAAT binding protein[221] (see Section III.A and references therein). This protein also binds to an enhancer core region that is homologous (i.e., well conserved) in many viral enhancers. The amino acid sequence composing the relevant part of this protein has been arranged on a schematic α helix. It was found that one side of the helix was composed predominantly of hydrophobic amino acids (particularly leucines), while the other side displayed amino acids with charged side chains and uncharged polar side chains.[221] Periodic leucines were later found in additional DNA binding proteins of organisms ranging from yeast to man: GCN4, v-*fos*, v-*jun*, mouse c-*myc*, human N-*myc*, and human L-*myc*. In these proteins, leucine repeats at every seventh position at least four times. These α helices appear to be unusually stable. They are stabilized both by the amphipathic arrangement of hydrophobic amino acids and by the occurrence of oppositely charged amino acids. The latter are arranged in a manner that allows formation of salt bridges, and an exceptionally high density of ion pairs was observed by Landschulz et al.[221]

Based on the above observations, these authors have proposed the leucine zipper model. In this model, the leucine ''spoke'' of the helical ''wheel'' interdigitate with another leucine ''spoke'' coming from a second ''wheel''. That could happen since in each wheel there is only one hydrophobic ''spoke''. The ''thin'' and ''long'' leucines were naturally selected for the interhelical hydrophobic interactions and preferred over the bulky short or stubby geometries of other hydrophobic amino acids. Interdigitation of leucine ''spokes'' with the helices arranged in an antiparallel orientation appear to be optimal for chain interlocking. In the CCAAT binding protein, the region immediately adjacent to the leucine zipper has a high proportion of basic amino acids. Interlocking of the leucine zipper of two polypeptides would juxtapose the positively charged regions of the two polypeptides in a manner suitable for sequence-specific recognition of the DNA.[221] Several papers have appeared in December 1988 corroborating the intriguing ''leucine zipper'' hypothesis.[253-257] Three murine jun proteins, c-jun,[252,253] jun-B, and jun-C[254] bind DNA either as homodimers or heterodimers. DNA binding by jun was markedly increased when c-*fos* was present.[253,254,256] c-*fos* thus appears to be a natural partner of jun. There is direct evidence that a leucine zipper plays a role in the fos-jun interaction.[256,257] Mutagenesis of the fos protein shows that a heptad repeat of leucine residues stabilizes the interaction between fos and jun.[256] Further, the basic residues adjacent to the leucine repeat of fos contribute to the DNA binding potential of the complex.[256] Studies with truncated fos proteins[257] also indicate that the fos leucine zipper domain is necessary for the DNA binding and that other protein regions must be important in conferring the high-affinity binding to the target sequence.*

C. Importance of DNA Stiffness in Protein-DNA Binding Specificity

The importance of DNA stiffness in protein-DNA interactions has recently been shown for the 434 repressor-operator binding.[258] As might be expected, the sequence-dependent DNA structure makes for sequence-dependent stiffness. The latter may be significant whenever a protein (or other ligand) bends or twists the DNA to form its bound complex. The relationship between DNA sequence and flexibility has also been addressed in Section V and VI.[24,132]

VII. NUCLEOSOMAL DNA AND ADDITIONAL CONSIDERATIONS OF NUCLEOPROTEIN COMPLEXES

The bulk of eukaryotic nuclear DNA is folded into a higher-order structure with the nucleosome serving as the basic repeating motif. The nucleosome is formed by a left-handed superhelical winding of about 150 bp of DNA around eight histone molecules. A varying length of linker DNA lies between these histone octamers.[259-262] Some structural considerations as well as experimental evidence suggest that the tightly wound superhelical nucleosomes are at least partially unwound during transcription (reviewed in Reference 262). The heat shock 70 (hsp70) protein genes of *Drosophila* provide a particularly well-studied system. The chromatin structure of the hsp70 genes changes upon transcription activation.[261-265] Unlike most of the hsp70 nucleosomal gene, the ~350-bp promoter proximal control region is hypersensitive to nucleases, and thus apparently nonnucleosomal both prior to and following the heat shock.[262,266-269] The promoters of additional, frequently transcribed genes have also been shown to be nuclease sensitive (e.g., References 270, 271). In particular, the SV40 regulatory region (Figure 1) is known to be nucleosome free.[272,273] Nevertheless, *in vivo* part of the regulatory regions is not naked in solution. These regions are still packaged to varying degrees in higher-order structures. Also, enhancer elements of both the heavy and k-light chain immunoglobulin genes found in the intron between the variable and constant regions[274,275] are likely to be present in a nucleosomal stretch.

In a recent review[276] an important question has thus been raised: Can transcription factors that recognize their DNA sequences in solution also recognize them when they are incorporated into nucleosomes? Following Morse and Simpson,[276] I next describe some considerations pertaining to this question.

The DNA helix is severely deformed by its incorporation into nucleosomes. (1) The substantial bending of the DNA helix in looping around the nucleosomes changes the widths of the grooves. The crystal structure of the nucleosomal core particle[225] shows that in the region about 10 bp from the midpoint, the width of the minor groove varies from 7 Å on the inner face of the helix (facing the histone octoamer) to 13 Å on the outside. Likewise, the major groove varies from 11 to 20 Å. Another severe conformational perturbation that is likely

* **Note added in proof:** Structural studies have recently shown that the two α-helices of the leucine-zipper are arranged in a parallel orientation.

to affect protein-DNA recognition is (2) the presence of sharp kinks positioned one and four helical turns away from each side of the nucleosomal DNA. If the recognition sequence is inside the nucleosome, and in particular within the region near the midpoint, the first type of deformation (1) may imply that the protein recognizes bent DNA. In the nucleosome the DNA is bent toward the histones. In this case the protein factor would probably have to prefer binding to DNA bent away from it. We recall that the crystal structures of proteins complexed with DNA have mostly shown the DNA bending toward the protein. This is the case, for example, for the phage 434 repressor[223] and EcoRI.[131] However, in the DNAase I co-crystal the DNA is bent away from the enzyme.[174] If the recognition sequence embraces the DNA kinks (the second type of perturbation, [2], noted above), its backbone and base-pair stacking interactions would be grossly deformed. This could either hinder recognition by interrupting a sequence, or might serve as specific motifs for interaction between proteins and nucleosomal DNA. Recent studies (summarized in Reference 276) indicate that in comparison with other helical parameters, alterations in DNA twist are of marginal importance in considerations of proteins interacting with nucleosomal DNA.

Morse and Simpson[276] raise additional steric and ionic considerations, pointing against recognition of signals inside the nucleosomal particle. The association of the histone proteins with the DNA, and the proximity of almost two turns of DNA around the nucleosomes with the DNA also "entering" and "exiting" the nucleosome core particle can render the DNA inacessible to many regulatory proteins.

It appears that regulatory proteins may cover a whole spectrum, from recognizing only nucleosome-free DNA (e.g., TFIID, the TATA factor[277]) to recognizing only nucleosomal DNA in a particular orientation.[276] In addition, there may be proteins that recognize the ~50 bp of linker DNA, joining the two nucleosomes. Cross-linking studies[262] have shown that this DNA is also associated with histones. Recognition of DNA by transcription factors may thus require the nucleosomes to be displaced.[117,277,278] For example, in the mouse mammary tumor virus promoter, steroid hormone induction appears to be accompanied by an alternation or loss of nucleosomal structure.[276,279] Other strategies might also be employed by regulatory proteins in dealing with the presence of chromatin in their interactions with eukaryotic DNA.

Is there a preferred nucleotide sequence orientation in nucleoprotein complexes? As noted above, the tight bending of DNA by a protein requires that particular stacking interactions between adjacent base pairs can accommodate the deformations in groove width.[168] In nucleosomal DNA, the variation in groove width is correlated with the preferred occurrence of AAA/TTT with the minor groove and GGG/GCC with the major groove facing the histone octamer.[166] As discussed in Section V describing DNA structure, A/T sequences have a narrower minor groove, whereas G/C sequences have a narrower major groove. In bent DNA, the groove that faces inward is narrower, and, consequently, the groove that faces outward is wider. Thus, when DNA is bent toward the protein, as in nucleosomes, A/T sequences are preferentially positioned with their minor groove facing in. In G/C sequences, the major groove faces in. The preferential occurrence of these sequences in nucleosomal DNA is periodically modulated. If the direction of curvature is constant and the bending is uniform, the sequence periodicity should coincide with the double helical periodicity.[168,173] This means recurrence of A/T or of G/C sequences every 10 bp. The distance between the A/T and G/C nucleotides should preferably be half a helical turn. This periodic recurrence has been noted not only for nucleosomal sequences,[166,168] but for DNA sequences involved in other nucleoprotein complexes as well.[168] The relative contribution of bending to the complex formation is likely to depend both on the magnitude of the curvature and on the length of the curved DNA segment.

A sequence-specific regulatory protein that bends the DNA considerably[280,281] is the *E. coli* CAP protein. From the dimensions of the complex formed and its energetics[282] it has been concluded that this palindromic DNA is bent toward the protein. This is in agreement with predictions based on the periodic G/C and A/T sequence recurrence and their positioning with respect to the palindromic axis of symmetry.[168]

VIII. POTENTIAL UPSTREAM SEQUENCE/STRUCTURE SIGNALS

To motivate the search for signals, let us recall some of the basic generalities discussed above and illustrate with a few additional examples.[283] Within long DNA stretches in eukaryotes, specific sequences and/or structures are present to serve as signals to the eukaryotic transcriptional proteins. Recognition and regulation are not based on just one signal, but rather on the combination of several components. One signal type specifies the site of transcription initiation. Other signals regulate its efficiency. Some of the signals may be gene- or sequence-specific (e.g., the glucocorticoid-repressed and -induced genes,[284-287] the large T-antigen binding sites in SV40,[288] upstream activation elements induced by amino acid starvaion[289-292]). Other signals are of a more general nature[293] and some are present in most genes. Some of the signals are transient. In active genes the chromatin structure[294,295] and the methylation patterns[296,297] are different from those of inactive genes as judged by sites sensitive to nucleases.

Traditionally, searches for signals in upstream regions are often characterized by looking for "consensus" sequences. In deriving the consensus, similar sequences that are presumed to play the same role in regulation are aligned and the most frequent nucleotide(s) in a given position is (are) noted. This procedure has some shortcomings:[298]

1. The number of aligned sequences is often too small to derive a statistically meaningful consensus.
2. The consensus is an average result, and in any single sequence the variations from the consensus may be quite large.

For a relatively short consensus, even a single nucleotide change may result in a somewhat different DNA structure.[299] Two or more variant nucleotides may yield a region with a very different geometry. Our capability to predict the structural changes stemming from such nucleotide variations is limited.[300] Often, in constructing the consensus, thermodynamic considerations have played a major role. This appeared to make A·T and T·A base pairs interchangeable (e.g., ...GC$^{\text{T}}_{\text{A}}$GC . . .). It is worthwhile to note, however, that it is unclear what effect a single base change may have on the thermodynamnics of the opening of double-stranded DNA. Recent calorimetric measurements[301] indicate that single-base composition is insufficient to determine the free energy required to transform the double helical into single-stranded DNA. On the other hand, predictions based on dinucleotide composition are in good agreement with the experimental data.[301] Thus, in our example here, the sequence GC$\underline{\text{T}}$GC thermodynamically differs from GC$\underline{\text{A}}$GC. We have recently used the doublets free energy data for prokaryotic promoters.[302] There, extensive compilation of point mutations indicated that the up and down mutations in the TATAAT −10 region are correlated with the instability at that site. Comparisons of the free energies of each mutant with its wild type, up mutations have higher free energies; down mutations have lower free energies.[302] We have not observed such a correlation in the TTGACA −35 prokaryotic promoter region. In any case, in general it is questionable whether or not the thermodynamic change resulting from a point alteration is more important than a change in the DNA geometry, and in the hydrogen-bonding hydrophobic-interaction scheme. Another assumption is that a transition mutation has a limited effect on the DNA geometry and thus is preferable in the construction of a consensus. However, the A·T base pair geometry is very different from that G·C. Because of these potential difficulties, I shall describe here computer searches for putative signals that are composed of exactly the same nucleotide sequence. In order to have a statistically meaningful sample, oligomers longer than pentamers were not considered in position-specific searches, and hexamers were the maximal size in cumulative upstream searches.

A. Alignment of Nucleotide Sequences by their Transcription Initiation Sites

1. The Method

All sequences having a mRNA start site noted in their GenBank description are pulled out and aligned by that site. The mRNA transcription initiation site is assigned position zero. The analyzed primate sample contains 515 sequences containing 584 start sites; the rodents 607 sequences with 698 starts; other mammals 61 sequences with 66 starts; nonmammalian vertebrates 135 sequences with 141 mRNA starts, the invertebrates 75 sequences with 100 starts, and the viruses 42 sequences with 56 start sites. Altogether, 1641 mRNA start sites were used. We studied the distribution of query oligomer (mononucleotides up to some penta-nucleotides) with respect to the transcription initiation site. A window (mostly of length 25, see caption to Figure 5) is moved from position −500 to +500 in the aligned sequences. At each step, the number of occurrences of the query oligomer is counted. The total number of bases scanned in this window and the percent occurrence of the oligomer in the window are calculated. (Note that the total number of bases in each window decreases toward ±500, since fewer of the available sequences extend all the way to these points [for further details see References 283, 303].) In the following, the main results are listed. The computations were repeated with "cleaned" GenBank data from which cDNAs have been omitted. The analysis has also been performed omitting some highly represented gene families. The basic results presented here have not changed.

2. Upstream Sequences Are Extremely Rich in GC, Accentuating the AT-Rich TATA Box Signal

The more interesting results of these computations are displayed in Figure 5. The mononucleotide distributions are shown in Figure 5A, the doublets in Figure 5B, triplets in Figure 5C, quartets in Figure 5D, and pentamers in Figure 5E. Inspection of these results reveals the following:

1. The distributions (Figure 5A) of adenines and thymines are the opposite of those of cytosines and guanines. The concentrations of Cs and Gs rise from the flanks of the graphs toward upstream of the transcription initiation sites. The sharp peak in the distributions of A and related oligomers, AA, TA (Figure 5B), ATA (Figure 5C), ATAA, TATA (Figure 5D), ATAAA (Figure 5E), correspond to the TATAAAT sequence at −30.
2. G and G + C oligomers are extremely frequent upstream of transcription initiation. This is seen for doublet (GG, CG, GC, CC — Figure 5B), triplets (GGG, CGG, GGC, CCC — Figure 5C), quartets (CCCC, CGGG, GCGG, GGGC, GGCG — Figure 5D) and pentamers (CGCCC, GCGGG, GGGCC, GGGCG — Figure 5E).
3. The two quartets, CAAT and TTGG, are not frequent at position −80 (Figure 5D). However, the pentamers ATTGG and CCAAT (Figure 5E) already stand out considerably in the higher eukaryotes. This is particularly interesting since CCAAT is contained within the GGCCAATCT box of higher eukaryotes. ATTGG is part of the yeast UAS2 element, having the consensus TG$^{\text{G}}_{\text{A}}$TTGGT. This consensus is recognized by the HAP2/HAP3 yeast proteins.

Further, as shown in Figure 3, the pattern of twist angles and of roll angles (computed using Table 1 values) is symmetric, both for the CCAAT box (Figures 3A and 3B) and for its yeast analog (Figure 3C and 3D). The mammalian CCAAT and the yeast UAS2 thus have similar structures and potential hydrogen bonding patterns (Table 2) that can be recognized by their binding proteins. It is not surprising then, that the subunits of the yeast[65] and of the human CCAAT binding protein[64] are functionally interchangeable.[66]

As already discussed in Section III, the Sp1 transcription factor binds to a GC "box" containing the consensus GGGCGG. The GC-rich region observed here covers a wide span upstream of transcription initiation, and peaks around −50, i.e., just prior to the TTAAAT "TATA box" sequence. The GC richness upstream of transcription initiation sites accentuates the AT-rich TATA signal. This variation between the G/C and A/T sequences can affect the DNA structure and flexibility. It is noteworthy, however, that the frequent GC sequences mostly contain oligo d(G)$_n$ runs, whereas frequent d(GC)$_n$ alternations are not observed. We shall come back to this point below.

From the statistical standpoint, the large data set used makes these results highly significant.

3. Advantages and Disadvantages of This Method

Scanning sequences aligned by their mRNA start sites detect only oligomers that recur frequently at a given distance from the transcription initiation site. This method is thus efficient for signals that have preferred positions, such as the TATA and CCAAT. By moving a window we allow the distances to vary to some extent.

There are, however, two major disadvantages:

1. As we have seen throughout this review, the position of upstream/enhancer elements can vary. If this distance is not changed drastically (up to several kilobases), the effect on the efficiency of transcription initiation is marginal. Thus, regulatory elements whose position with respect to transcription initiation varies in the different sequences would not be picked up by this analysis.
2. The second disadvantage is that this method would not pick up tissue- or gene-specific elements. The percentage of such sequences (and elements) in the data base might be too small.

B. Strong Patterns in Homooligomer Tract Occurrences in Noncoding and in Upstream Regions

In the last few years, attention has been focused on oligo(dA)·oligo(dT) tracts. The experiments of Wu and Crothers[280] have indicated that such tracts, when embedded in mixed DNA sequence, have unique conformations. If the oligo$(dA)_{5-6}$·oligo$(dT)_{5-6}$ repeats are phased with the helical pitch (i.e., $(N_5A_{5-6})_n$ with N any base), the sequences are reported to be appreciably bent.[304] This behavior may result from the geometry of the (dA)·(dT) base pair steps per se, or from the junctions of the tracts with the mixed DNA.[136,169,170,280] As noted in Section V, oligo (dA)·(dT) are frequent in upstream regions and in some cases have been shown to play a role in protein recognition and binding.[138,139,141,146,149-152] Physical-chemical studies suggest that oligo (dA)·oligo (dT) can have some unique conformations. The crystal structures obtained by Nelson et al.,[169] Coll et al.,[170] and Aggarwal et al.[224] show highly propeller-twisted, well-stacked A·T base pairs with bifurcated hydrogen bonds. While consecutive A·T base pairs have very small roll and tilt, there are large roll angles at the base pair steps at the ends of the tract. Overall bending could, in part, result from the junctions if the roll angles at the ends of oligo (dA)·oligo (dT) tracts had opposite directions with respect to the minor and major grooves. The overall bending may be increased if the 5′ and 3′ ends of the run are half a helical turn apart, as was the case in the experiments of Wu and Crothers.[280] The hydroxyl radical cutting patterns studied by Burkhoff and Tullius[172] and Tullius and Dombroski[223] also suggest that sequence-directed DNA bending results from the NA and AN (N ≠ A) junctions rather than from the A·A steps per se. If indeed the bends originate from the junctions of the unique $(dA)_n$·$(dT)_n$ conformation and the mixed sequence DNA, then the degree of bending and its detailed geometry might depend on the flanking sequences. We are thus led to analyzing genomic DNA for flanking sequence preferences.[305-308] The limited data of Koo et al.[136] suggested that maximal bending is obtained when C precedes the run and T follows it and that the bend at the 3′ is larger than that at the 5′ junction. Given the 10.5 bp periodicity in DNA, clearly different tract lengths may result in a different spatial orientation of the bend. The minimal length of the A·T tracts resulting in bent DNA is uncertain.[136,309] A question that arises, then, is whether or not tracts of varying lengths differ in the type or strength of the flanking sequence preference.[305-308] Although I have stressed here oligo(dA)·oligo(dT) tracts, oligo(dG)·oligo(dC) tracts apparently possess distinct characteristics as well.[153-155] The occurrence of such runs in eukaryotic upstream regions (see above and Figure 5) implies the possibility of special properties.

Computer analysis of the data base[306,307] indicates that in both prokaryotes and eukaryotes, A/T runs are preferentially flanked on either the 5′ or the 3′ ends by A and/or T. G/C runs are preferentially flanked by G or C (group 1). There is discrimination against A/T runs flanked by G or C and G/C runs flanked by A or T (group 3). The frequencies of A/T runs flanked by T or A and G or C and G/C runs flanked by G or C and T or A is intermediate (group 2). See Figure 6 for a complete list of the group members. These features are more pronounced in the eukaryotic data base and become stronger as the length of the tracts increases from two to three to four base pairs. The difference between prokaryotes and eukaryotes could be a manifestation of the differences in the packaging

Group	Oligomers			
1a	AT AA . . .	CG CC . . .	GC GG . . .	TA TT . . .
	. . . AA TA	. . . CC GC	. . . GG CG	. . . TT AT
1b (YR) (Complementary)	TT AA . . .	CC GG . . .	. . . CC GG	. . . TT AA
1c (RY) (Complementary)	AA TT . . .	GG CC . . .	. . . GG CC	. . . AA TT
1t	Group 1a +	Group 1b		
2	AC AA . . .	AG AA . . .	TC AA . . .	TG AA . . .
	. . . AA CA	. . . AA GA	. . . AA CT	. . . AA GT
	CA CC . . .	CT CC . . .	GA CC . . .	GT CC . . .
	. . . CC AC	. . . CC TC	. . . CC AG	. . . CC TG
	CA GG . . .	CT GG . . .	GA GG . . .	GT GG . . .
	. . . GG AC	. . . GG TC	. . . GG AG	. . . GG TG
	AC TT . . .	AG TT . . .	TC TT . . .	TG TT . . .
	. . . TT CA	. . . TT GA	. . . TT CT	. . . TT GT
	CT AA . . .	GT AA . . .	CA TT . . .	GA TT . . .
	. . . AA TC	. . . AA TG	. . . TT AC	. . . TT AG
	AC GG . . .	TC GG . . .	AG CC . . .	TG CC . . .
	. . . GG CA	. . . GG CT	. . . CC GA	. . . CC GT
3	CC AA . . .	CG AA . . .	GC AA . . .	GG AA . . .
	. . . AA CC	. . . AA GC	. . . AA CG	. . . AA GG
	AA CC . . .	AT CC . . .	TA CC . . .	TT CC . . .
	. . . CC AA	. . . CC TA	. . . CC AT	. . . CC TT
	AA GG . . .	AT GG . . .	TA GG . . .	TT GG . . .
	. . . GG AA	. . . GG TA	. . . GG AT	. . . GG TT
	CC TT . . .	CG TT . . .	GC TT . . .	GG TT . . .
	. . . TT CC	. . . TT GC	. . . TT CG	. . . TT GG

FIGURE 6. A listing of the oligomers included in each group (see Table 3). The larger size letters are the flanking doublets. The smaller letters depict the nucleotides constituting the run. Group 1a is the most frequent. Surprisingly, group 1a is the most frequent. Surprisingly, group 1b is considerably and consistently more frequent than its mirror image, group 1c. In both groups 1b and 1c, the flanking doublet contains nucleotides complementary to the run. However, in group 1b there is a pyrimidine (Y)-purine (R) junction, whereas in group 1c there is a RY junction. The frequency of occurence of group 2 is just about the expected. In group 3 the flanking doublet does not contain any nucleotide complementary to the run. This group is strongly disfavored, i.e., its frequency of occurrence is very low. The actual values in upstream regions and in all the noncoding sequences are given in Table 3.

of the eukaryotic nucleosomal DNA. The wrapping of the DNA around the histone core proteins might impose structure-correlated constraints on the sequence that would differ from those found in prokaryotic DNA.

If the distinct patterns in the frequencies of occurrence of these oligomers are indeed related to protein-DNA interaction, then one might expect the patterns to be more pronounced in protein binding sites, i.e., in regulatory regions. In particular, the group 1 oligomers (i.e., runs flanked by their complementary nucleotide and/or the same nucleotide, e.g., A_4TA, CCG_4, etc.) should be even more prominent upstream and/or downstream of coding regions. We thus examine the oligomer frequencies in consecutively larger sections of the upstream regions. We also examine the behavior of each oligomer within these three groups. Within group 1, ATA_n, CGC_n, GCG_n, TAT_n, A_nTA, C_nGC, G_nCG, and T_nAT (group 1a) are most frequent overall, but particularly so in regulatory regions. TTA_n, C_nGG, CCG_n, and T_nAA, having YR (Y is a pyrimidine and R a purine) junction (group 1b), are frequent overall and as such might conceivably play a role in eukaryotic DNA packaging. All trends increase in strength as the run lengthens. These features may have structural implications since the various oligomers have in general different anisotropic flexibilities (i.e., they bend more easily in different directions).

Interestingly, A_nTT, AAT_n, G_nCC, and GGC_n, having RY junction between the complementary nucleotides (group 1c), are considerably less frequent than their YR symmetric oligomers (group 1b). It is intriguing that one observes consistent trends in the frequencies of occurrence of these groups, group 1a (is more frequent than) > group 1b > group 1c. Also, group 1t (= groups 1a + 1b) > group 2 > group 3. Indeed, every single oligomer of group 1b is more frequent than every oligomer of group 1c, even though they are composed of exactly the same mono- and dinucleotides, except for the junction where the order is reversed. In group 3 the situation is exactly reversed, with RY junctions being more frequent than YR. Of interest also is the finding that within group 1a and particularly 1b, CG junction oligomers are consistently more frequent than TA ones (see also Figure 5). These and other consistent patterns derived from a wealth of data (4, 5 × 10^6 nucleotides in the noncoding eukaryotic and 0.8 × 10^6 in the noncoding prokaryotic data base) might be instructive in elucidating some aspects of protein-DNA interaction. This section is heavily based on Reference 308.

1. Methods

All eukaryotic noncoding regions present in the GenBank data base (about 4.5 × 10^6 nucleotides) have been scanned for occurrences of homooligomer runs. For any given minimal length homooligomer tract the bases 5′ and 3′ to the run have been recorded separately. Thus, for NA_n and A_nN, where $n \geqslant 2$ and $N \neq A$, we count the number of times we have CA_n, GA_n, and TA_n. Similarly, we also note the occurrences of A_nC, A_nG, and A_nT. The 5′ and 3′ results are compiled separately.

Because there are large differences in the mono- and dinucleotide composition of eukaryotic sequences,[310,311] each of these values is normalized by dividing these counts by the counts of the overlapping doublets and multiplying by the counts of the mononucleotides except those at the ends.

Thus,

$$\tilde{P}(NA_n) = \frac{P(NA_n) \cdot [P\ (A)]^{n-1}}{P(NA) \cdot [P\ (AA)]^{n-1}} \qquad (1)$$

$$\tilde{P}(A_nN) \cdot \frac{P(A_nN) \cdot [P\ (A)]^{n-1}}{P(AN) \cdot [P\ (AA)]^{n-1}} \qquad (2)$$

For example, $\tilde{P}(NA_n)$ can be considered to be constructed from the actual occurrences $P(NA_n)$ divided by a $P_2(NA_n)$, a probability constructed for the pairwise conditional probabilities as

$$P_2(NA_n) \cdot P(NA) \qquad \frac{P(AA)^{n-1}}{P(A)}$$

Analogous computations have also been carried out for C, G, and T runs.

In addition to the mononucleotide recurrences, the dinucleotide frequencies at the 5′ and 3′ ends of the tracts were computed as well. For a given homooligomer tract, all doublets preceding and following it are recorded. Thus, for A runs we separately count all MNA_n and A_nNM where $N \neq A$, M = A,C,G,T, and n = 2,3,4,5,6. Again we normalize these counts by the respective counts of the overlapping doublets and by the mononucleotides except those at the ends:

$$\tilde{P}(MNA_n) = \frac{P(MNA_n) \cdot [P\ (A)]^{n-1} \cdot P\ (N)}{P(MN) \cdot P(NA) \cdot [P\ (AA)]^{n-1}} \qquad (3)$$

$$\tilde{P}(A_nNM) = \frac{P(A_nNM) \cdot [P\ (A)]^{n-1} \cdot P(N)}{P(NM) \cdot P(AN) \cdot [P\ (AA)]^{n-1}} \qquad (4)$$

As before, the analysis has also carried out for C_n, G_n, and T_n with n = 2,3,4,5,6.

The deviation of the normalized frequencies from the mean is also computed. Specifically, for an A run and for a given N(N ≠ A) the mean is

$$\begin{array}{lll} 1/4 & \Sigma\tilde{P}(MNA_n) & \text{for the 5' function} \\ & M = A, C, G, T & \end{array}$$

$$\begin{array}{lll} 1/4 & \Sigma\tilde{P}(A_nNM) & \text{for the 3' junction} \\ & M = A, C, G, T & \end{array}$$

Analogous computations are carried out for C_n, G_n, and T_n. For the deviation of the normalized mononucleotide frequencies from the mean, the mean is

$$\begin{array}{lll} 1/3 & \Sigma\tilde{P}(NA_n) & \text{for the 5'} \\ & N = C, G, T & \end{array}$$

$$\begin{array}{lll} 1/3 & \Sigma\tilde{P}(A_nN) & \text{for 3' junctions} \end{array}$$

The above calculations were carried out for each homooligomer separately and jointly for groups 1a, 1b, 1c, 1t (1a + 1b), 2, and 3.

The whole analysis was then repeated 20 times: for the first 200, 300, 400, and 500 nucleotides upstream from the coding regions and for the whole upstream regions in eukaryotes and prokaryotes. (Analogous analysis has also been carried out for the downstream [from coding] regions.)

In "whole upstream" computations the DNA sequences are used from the first nucleotide left (5′) of the coding region to the end of the sequence. If there is a second coding region to the left of the first one, the "whole" upstream region is included up to the 3′ end of that second region. Analogous approach is employed in the 200, 300, 400, and 500 nucleotides upstream (as well as downstream) calculations. This detailed upstream (and dowsntream) analysis has been carried out only for n = 2,3,4, since longer runs are too scarce to yield statistically meaningful results.

2. Homooligomer Tracts Are Preferentially Flanked by Their Complementary Bases with a Pyrimidine-Purine Junction (e.g., CCGGGG, TTTTAA)

The deviations from the mean for the six groups (listed in Figure 6) are tabulated in Tables 3A and 3B. The most frequent group is 1a. Runs X_n of group 1a contain XZ..., ...ZX flanking doublets, with Z complementary to X (e.g., GCG_4,A_4TA). The least frequent group, group 3, contains UV flanking doublets, with U, V ≠ X, Z (e.g., TAG_4; A_4CG). Group 2 contains mixed doublets, ZU, XU and UX, UZ (e.g., CAG_4; A_4CT). It has intermediate frequency, just as expected. Groups 1b and 1c have ZZ flanking doublets with YR junction in the first (e.g., T_4AA, CCG_4) and RY in the second (e.g., A_4TT, GGC_4). Interestingly, Group 1b is consistently more frequent than group 1c.

To further discern the major contributors to the groups 1a, 1b, and 1c behavior, the deviations from the mean of the single oligonucleotides constituting these groups have been inspected (Tables 3C and 3D). Although the numbers are much smaller, trends are observed. In eukaryotes the major contributor to the upstream abundance within group 1a is GCG_4, and to a lesser extent G_4CG, followed by ATA_4 and TAT_4. Table 3D compares the group 1b oligomers with their mirror images of group 1c. These oligomers, whose only difference is that group 1b contains YR junctions (TTA_4, CCG_4, T_4AA, C_4GG), whereas group 1c contains RY (A_4TT, G_4CC, AAT_4, GGC_4), differ strikingly in their behavior. This table also shows the homogeneous behavior of the various 1b oligomers in the overall noncoding and in the regulatory regions.

Table 3a

Group	Tract length 2		3		4		5
1a	22	<	39	<	58	<	81
	V		V		V		V
1b(YR)	4	<	20	<	31	<	36
	V		V		V		V
1c(RY)	1		-3		4		−2
1t [(1a + 1b)/2]	13	<	30	<	45	<	58
	V		V		V		V
2	2		2		3		1
	V		V		V		V
3	−9		−16		−23		−26

Our studies of context preferences of homooligomer tracts unraveled some patterns.[305-308] First, there are no significant differences in behavior between the 5′ and 3′ flanks.[136,155, 172] A(T) is frequent both prior to and following T(A) tracts in eukaryotes. The most infrequent nucleotide preceding an A run is a C. Thus, there is no direct relationship with the Koo et al.[136] bending data. The most frequent nucleotide preceding (and following) G (C) runs are C (G). Also, following an interruption of a run by any base, there is a tendency of the run to continue with the nucleotide identical to the run.

The doublets flanking homooligomer tracts divide them into three groups. The most frequent tracts are those flanked by the complementary bases and the same base, i.e., G/C runs flanked by C and/or G, GCG_n, CCG_n, G_nCG, G_nCC, and CGC_n, GGC_n, C_nGC, C_nGG. Similarly, A/T runs are preferentially flanked by A and/or T. The second group is of a mixed type: A/T runs flanked by C/G and T/A; or G/C runs flanked by A/T and C/G, e.g., GTG_n, A_nTC, C_nAG, etc. The frequency of this group is about the overall mean. There is a clear-cut discrimination against the third group, in which both flanking nucleotides are of a type opposite to the run, i.e., A/T runs flanked only by G and/or C, G/C runs by A and/or T. Although runs of various lengths (n = 2,3,4) behave similarly, the trends increase with run lengths n. Prokaryotes and eukaryotes noncoding regions showed similar trends, although the patterns are stronger in eukaryotes.

If indeed the flanking preferences are related to DNA con-

Table 3b

	Groups 1a		1b		1c	1t		2		3
n = 2										
All noncoding	22	>	4	>	1	13	>	2	>	− 9
300 NT upstream	28	>	5	>	2	17	>	1	>	−10
Whole upstream	24	>	6	>	4	15	>	2	>	−10
n = 3										
All noncoding	39	>	20	>	−3	30	>	2	>	−16
300 NT upstream	42	>	16	>	−4	29	>	2	>	−15
Whole upstream	41	>	22	>	−1	31	>	3	>	−17
n = 4										
All noncoding	58	>	31	>	4	45	>	3	>	−23
300 NT upstream	76	>	21	>	4	49	>	0	>	−22
Whole upstream	66	>	32	>	3	49	>	3	>	−26

Table 3c

	Group 1a							
	ATA_4	A_4TA	TAT_4	T_4AT	CGC_4	C_4GC	GCG_4	G_4CG
All noncoding	57	58	67	48	42	69	73	50
300 NT upstream	80	67	74	32	45	70	145	98
Whole upstream	67	72	73	47	46	68	100	59

Table 3d

Oligomer	TTA_4		A_4TT		T_4AA		AAT_4		CCG_4		G_4CC		C_4GG		GGC_4
Group	1b		1c		1b		1c		1b		1c		1b		1c
Pattern	YR	>	RY	<	YR	>	RY	<	YR	>	RY	<	YR	<	RY
All noncoding	22	>	5	<	24	>	3	<	35	>	−2	<	45	>	8
300 NT upstream	21	>	3	<	17	>	13	<	26	>	−2	<	19	>	0
Whole upstream	28	>	6	<	26	>	6	<	29	>	−5	<	45	>	4

Note: Frequencies of occurrence of homooligomer tracts in upstream and all noncoding regions. (a) The deviations from the mean of the homooligomer tracts of the six groups. (b) The deviations from the mean in 300 nucleotides upstream, "whole upstream", and all noncoding regions. In "whole upstream" the upstream sequence is taken into account until the first coding region to the left is encountered. If there is no such coding sequence, we take the whole upstream noncoding sequence present in the data base. Tract lengths n = 2, 3, and 4 are analyzed here. (c) The deviations from the mean of the oligomers constituting group 1a. (d) The deviations from the mean of group 1b (Y_4RR and YYR_4) vs. their mirror image group 1c (R_4YY and RRY_4), with Y complementary to R. The clear-cut trends are marked in the tables. The oligomers belonging to each group are detailed in Figure 6.

formation and to protein-DNA interactions, then these preferences should be stronger in regulatory regions with a higher concentration of specific protein binding sites. While the numbers are smaller, the large current data base permits such an analysis. We have also attempted to further refine the three groups noted above and to identify within the first preferred group those oligomers contributing most to the effect.

As Table 3B shows, the frequencies of group 1t, 2, and 3 are roughly the same in the regulatory region and in the overall noncoding sequences. The first is very frequent, the second near the mean, and the third is disfavored. Within group 1t, 1a is favored over 1b, particularly in regulatory regions. Possibly group 1a plays a role in protein-DNA interactions or conformation in the regulatory regions. Groups 1a and 1b might have a role in nucleosomal DNA packaging. There is a consistency in this respect that in prokaryotes with different packaging constraints[312-314] the frequencies of group 1b are lower. The oligomers constituting the groups are listed in Figure 6. Group 1a consists of X_nZX and XZX_n, where X and Z are complements of each other. Group 1b consists of X_nZZ and ZZX_n with pyrimidine-purine (YR) junction. Why group 1a should be more frequent in regulatory regions than 1b is unclear. Within 1a, GCG_4 and G_4CG are particularly abundant. These oligomers are similar to the Sp1 binding sequence, shown to be important in the regulation of gene expression.[3] In the overall eukaryotic noncoding regions, groups 1a and 1b are still preferred for homooligomer tracts of lengths 5 and 6. If these preferences are indeed related to groove width as the DNA winds around the protein, one would expect some tract length preference limit, since the major and minor grooves are alternatively facing the protein.

Why is the "other type" group (group 3) so disfavored? While the answer is unknown, it could conceivably be related to the different characteristics displayed by A/T vs. G/C sequences. The consistently different frequencies of occurrence seen in the group 1b vs. 1c (Table 3D) is particularly intriguing. Oligomers composed of complementary nucleotides are preferred only if the flanking doublets (underlined) consist of one complementary and one identical base to the run (1a, e.g., ATA_4) or if it consists of two nucleotides complementary to the run (group 1b, e.g., TTA_4) with YR junction. If there is a RY junction (group 1c, A_4TT) the oligonucleotide is disfavored. While in eukaryotes the frequency of the 1c group is near the mean, in prokaryotes it is among the least favored groups (Table 3b).

Is there a concrete explanation for such a discrimination of these doublets? From Calladine's[142] model, we could argue that sequences with larger YR steric hindrance clashes can be preferred. Indeed, as Table 3d shows, CG junction oligomers (C_4GC, CCG_4) with worse clashes are more frequent than TA ones (TTA_4, T_4AA). Still, group 3 with YR clashes (GCA_4, T_4GG) is disfavored. Zhurkin et al.[315] have suggested that YR (RY) doublets are preferentially positioned in nucleosomal DNA with the compressed major (minor) groove facing the protein. However, the trends in group 3 (not shown) are the opposite of those of Table 3d, namely, RY junction oligomers (A_4CC,

AAC_4, G_4TT, GGT_4) are somewhat preferred over YR (CCA_4, C_4AA, TTG_4, T_4GG). Also, the trends are strengthened with tract length. In particular, these trends are also strong in prokaryotes with nonnucleosomal DNA (although there might be other forms of DNA-protein packaging).

It thus appears that the effect observed here might be the result of both different G/C (vs. A/T) characteristics and the RY/YR geometries. However, the nonspecific nature of the preference would seem to imply a conformation effect rather than a specific recognition of atoms. In agreement with this conclusion are the higher frequencies of occurrence of the group 1a CG junction sequences (GCGGGG, CCCCGC) in the noncoding regions over TA ones (TTTTAT, ATAAAA) (see Table 3C). In particular, group 1b CG junction oligomers (CCG_n, C_nGG) are much more frequent than TA ones (TTA_n, T_nAA). This trend is seen for n = 3,4 in eukaryotes and n = 2,3,4 in prokaryotes. It is not found in the RY junction oligomers of group 1c. The presence of RY doublets (in addition to the YR ones) in the group 1a oligomers might explain why this CG>TA trend is weaker in group 1a vs. 1b.

A CG junction doublet embedded in C/G containing oligomers might then bend more easily into the major groove owing both to its G + C content[166-168,173] as well as its being a YR junction,[315] possibly leading to the observed preferences. Our conformational calculations[24,132] also suggest that a CG junction may assume a large roll angle. In this respect it is also of interest to note that the nucleosomal positionings of CCGG and TTAA are apparently more important than those of GGCC and AATT.[315] That, again, seems to be in accord with the YR group 1b (first two oligomers) higher frequencies than RY group 1c (latter two oligomers). Our data seem to suggest that within these ...TTAA..., ...CCGG... junctions the former might play a lesser role than the latter.

While models and explanations of the trends shown here may differ, the consistent recurring trends are based on a very large data base, making the patterns themselves statistically highly significant.

IX. GIVEN AN UPSTREAM SEQUENCE, CAN A RECOGNITION ELEMENT BE PREDICTED? DNA FLEXIBILITY MAY BE A CRUCIAL FACTOR

Can a recognition element be predicted? Can we take an upstream (or any other) DNA sequence, scan it, and suggest with some degree of certainty which elements (if any) within it are recognized by protein factors? This would be the case if (1) we understand the principles governing protein-DNA interactions, and (2) that these principles can be translated into a simplified set of "rules". We do not fully understand the protein-DNA interactions, although much relevant information has accumulated to date. Most likely a simplified set of universal rules cannot encompass all the complex aspects; we recall some of these below: there are several modes of protein-DNA interaction. Different protein motifs might interact with the major groove, minor groove, and/or backbone in different ways. As can be seen from the co-crystals,[222,224,227] even the helix-turn-helix motif of the repressors is not always aligned in the operators in an identical way. Several side chains sometimes contribute to the recognition of a single base pair.[222,227] The conformation of a given DNA element can change in response to protein binding. Searches for elements by comparison of hydrogen bonding (see Table 2) "consensus" patterns may have to take into account potential phosphate backbone interactions.[176,224] In the major groove, nonpolar contacts may be as important as hydrogen bonds. Also, positively charged amino acid residues near the DNA backbone (lysines in the phage 434 repressor, Reference 224) can contribute to longer-range Coulombic interactions, although they are not anchored by direct hydrogen bonds. In addition, not all base pairs are contacted by the protein (e.g., the T-antigen binding site).[139]

Predicting potential DNA recognition elements may involve scoring of hydrogen bonding donors-acceptors (Table 2), or searches for "consensus" DNA angular parameters, such as twist or roll. Searches of this type have already been carried out for eukaryotes[164,316] and prokaryotes[165,298] upstream sequences. For the reasons outlined above, it appears, however, that such approaches as well as other simplified schemes might be insufficient for predicting protein recognition sequences.

Hydrogen bonding and hydrophobic and electrostatic interactions play a crucial role in the specificity of protein-DNA recognition. Recent protein-DNA interaction studies have also shown the large flexibility of the double-helical DNA. DNA structural parameters are extremely dynamic and may vary constantly in solution. DNA elements may thus distort or "wrinkle" easily. Indeed, data suggest the dependence of DNA helix flexibility on base composition and sequence.[24,317] Perhaps too much emphasis has been given lately to such parameters as DNA bending, easily detected by gel electrophoresis. Regions of the DNA may be distinguished by their intrinsic thermodynamics or flexibility, and not by different local structures, because they might not be trapped as such.

ACKNOWLEDGMENTS

I would like to thank Drs. R. Dhar, J. Hearst, S. Hughes, R. Jernigan, J. Maizel, H. Margalit, and A. Sarai for helpful discussions. I thank Drs. P. Chambon, M. Depamphilis, L. Guarente, W. Herr, J. Nevins, R. Roeder, W. Schaffner, P. Sharp, R. Tjian, and B. Wasylyk for supplying useful information, communicating results prior to publication, and making helpful suggestions. I thank M. Babcock and Drs. W. Olson and A. Srinivasan for the SV40 upstream region structures presented in Plate 1, and Dr. J. Mazur for the structural computations presented in Plate 2. J. Owens has written the pro-

gram for the sequence alignments, and D. Wang and G. Smythers have written the programs for the flanking tracts computations. The computations have been carried out using the Cray at the Advanced Scientific Computing Laboratory at NCI, FCRF.

REFERENCES

1. **Lewin, B. M.,** *Genes III,* John Wiley & Sons, New York, 1987.
2. **Maniatis, T., Goodbourn, S., and Fischer, J. A.,** Regulation of inducible and tissue-specific gene expression, *Science,* 236, 1237, 1987.
3. **McKnight, S. and Tjian, R.,** Transcriptional selectivity of viral genes in mammalian cells, *Cell,* 46, 795, 1986.
4. **Hope, I. A., Mahadevan, S., and Struhl, K.,** Structural and functional characterization of the short acidic transcriptional activation region of yeast GCN4 protein, *Nature,* 333, 635, 1988.
5. **Struhl, K.,** Promoters activator proteins and the mechanism of transcriptional initiation of yeast, *Cell,* 49, 295, 1987.
6. **Sigler, P.,** Acid blobs and negative noodles, *Nature,* 333, 210, 1988.
7. **Brent, R. and Ptashne, M.,** A eukaryotic transcriptional activator bearing the DNA specificity of a prokaryotic repressor, *Cell,* 43, 729, 1985.
8. **Hope, I. and Struhl, K.,** Functional dissection of a eukaryotic transcriptional activator protein, GCN4 of yeast, *Cell,* 46, 885, 1986.
9. **Guarente, L.,** UASs and enhancers: common mechanism of transcriptional activation in yeast and mammals, *Cell,* 52, 303, 1988.
10. **Guarente, L., Yocum, R. R., and Gifford, P.,** A GAL 10-CYC1 hybrid yeast promoter identifies the GAL4 regulatory region as an upstream site, *Proc. Natl. Acad. Sci. U.S.A.,* 79, 7410, 1982.
11. **Banerji, J., Rusconi, S., and Schaffner, W.,** Expression of a β-globin gene is enhanced by remote SV40 sequences, *Cell,* 27, 299, 1981.
12. **Benoist, C. and Chambon, P.,** In vivo sequence requirements of the SV40 early promoter region, *Nature,* 290, 304, 1981.
13. **Fromm, M. and Berg, P.,** Deletion mapping of DNA regions required for SV40 early region promoter function in vivo, *J. Mol. Appl. Genet.,* 1, 457, 1982.
14. **Maki, Y., Bos, T. J., Davis, C., Starbuck, M., and Vogt, P. K.,** Avian sarcoma virus 17 carries the jun oncoprotein, *Proc. Natl. Acad. Sci. U.S.A.,* 84, 2848, 1987.
15. **Struhl, K.,** The jun oncoprotein, a vertebrate transcription factor, activates transcription in yeast, *Nature,* 332, 649, 1988.
16. **Struhl, K.,** The DNA binding domains of jun oncoprotein and the yeast GCN4 transcriptional activator proteins are functionally homologous, *Cell,* 50, 841, 1987.
17. **Bohmann, O., Bos, T. J., Admon, A., Nishimura, T., Vogt, P. K., and Tjian, R.,** Human proto-oncogene c-jun encodes a DNA binding protein with structural and functional properties of transcription factor AP-1, *Science,* 238, 1386, 1987.
18. **Kakidani, H. and Ptashne, M.,** GAL4 activates gene expression in mammalian cells, *Cell,* 52, 161, 1988.
19. **Webster, N., Jin, J. R., Green, S., Hollis, M., and Chambon, P.,** The yeast UAS_G is a transcriptional enhancer in human HeLa cells in the presence of the GAL4 trans-activator, *Cell,* 52, 169, 1988.
20. **Yamamoto, K.,** Steroid receptor regulated transcription of specific gene and gene networks, *Annu. Rev. Genet.,* 19, 209, 1985.
21. **Lech, K., Anderson, K., and Brent, R.,** DNA-bound fos proteins activate transcription in yeast, *Cell,* 52, 179, 1988.
22. **Vogt, P. K., Bos, T. J., and Doolittle, R. F.,** Homology between DNA-binding domain of the GCN4 regulatory protein of yeast and the carboxyl-terminal region of a protein coded for by the oncogene jun, *Proc. Natl. Acad. Sci. U.S.A.,* 84, 3316, 1987.
23. **Rauscher, F. J., III, Sambucetti, L. C., Curran, T., Distel, R. J., and Spiegelman, B. M.,** A common DNA binding site for the fos protein complexes and transcription factor AP-1, *Cell,* 52, 471, 1988.
24. **Sarai, A., Mazur, J., Nussinov, R., and Jernigan, B. L.,** Sequence dependence of DNA conformational flexibility, *Biochemistry,* 28, 707, 1989.
25. **Serfling, E., Jasin, M., and Schaffner, W.,** Enhancers and eukaryotic gene transcription, *Trends Genet.,* 1, 224, 1985.
26. **Wasylyk, B.,** Transcription elements and factors of RNA polymerase B promoters of higher eukaryotes, *CRC Crit. Rev. Biochem.,* 23, 77, 1988.
27. **Ondek, B., Gloss, L., and Herr, W.,** The SV40 enhancer contains two distinct levels of organization, *Nature,* 333, 40, 1988.
28. **McKnight, S. L., Kingsbury, R. C., Spence, A., and Smith, M.,** The distal transcription signals of the herpes virus tk gene share a common hexanucleotide control sequence, *Cell,* 37, 253, 1984.
29. **Santoro, C., Mermod, N., Andrews, P. C., and Tjian, R.,** A family of human CCAAT-box-binding proteins active in transcription and DNA replication: cloning and expression of multiple cDNAs, *Nature,* 334, 218, 1988.
30. **Benoist, C., O'Hare, K., Breathnach, R., and Chambon, P.,** The ovalbumin gene-sequence of putative control regions, *Nucleic Acids Res.,* 8, 127, 1980.
31. **Efstratiadis, A., Posakony, J. W., Maniatis, T., Lawn, R. M., O'Connell, C., Spritz, R. A., DeRiel, J. K., Forget, B. G., Weissmann, S. M., Slightom, J. L., Blechl, A. E., Smithies, O., Baralle, F. E., Shoulders, C. C., and Proudfoot, N. J.,** The structure and evolution of the human β-globin gene family, *Cell,* 21, 653, 1980.
32. Reference deleted.
33. **Siebenlist, V., Hennighausen, L., Battey, J., and Leder, P.,** Chromatin structure and protein binding in the putative regulatory region of the c-*myc* gene in Burkitt lymphoma, *Cell,* 37, 381, 1984.
34. **Hennighausen, L., Siebenlist, V., Danner, D., Leder, P., Rawlins, D., Rosenfeld, P., and Kelly, T. J.,** High affinity binding site for a specific nuclear protein in the human IgM gene, *Nature,* 314, 289, 1985.
35. **Nowock, J., Borgmeyer, U., Puschel, A. W., Rupp, R. A. W., and Sippel, A. E.,** The TGGCA protein binds to the MMTV-LTR, the adenovirus origin of replication and the Bk virus enhancer, *Nucleic Acids Res.,* 13, 2045, 1985.
36. **Bienz, M. and Pelham, H. R. B.,** Heat shock regulatory elements function as an inducible enhancer in the *Xenophus hsp* 70 gene and when linked to a heterologous promoter, *Cell,* 45, 753, 1986.
37. **Graves, B. J., Johnson, P. F., and McKnight, S. L.,** Homologous recognition of a promoter domain common to the MSV LTR and HSV tk gene, *Cell,* 44, 565, 1986.
38. **Theisen, M., Stief, A., and Sippel, A.,** The lysozyme enhancer: cell-specific activation of the chicken lysozyme gene by a far-upstream DNA element, *EMBO J.,* 5, 719, 1986.
39. **Cereghini, S. and Yaniv, M.,** Assembly of transfected DNA into chromatin: structural changes in the origin promoter-enhancer region upon replication, *EMBO J.,* 3, 1243, 1984.
40. **Morgan, W. D., Williams, G. T., Morimoto, R. I., Greene, J., Kingston, R. E., and Tjian, R.,** Two transcriptional activators, CCAAT-box binding transcription factor and heat shock transcription factor, interact with a human hsp 70 gene promoter, *Mol. Cell. Biol.,* 7, 1129, 1987.
41. **Nagata, K., Guggenheimer, R. A., and Hurwitz, J.,** Specific binding of a cellular DNA binding protein to the origin of replication of adenovirus DNA, *Proc. Natl. Acad. Sci. U.S.A.,* 80, 6177, 1983.

42. **Rawlins, D. R., Rosenfeld, P. J., Wides, R. J., Challberg, M. D., and Kelly, T. J., Jr.,** Structure and function of the adenovirus origin of replication, *Cell,* 37, 309, 1984.
43. **Guggenheimer, R. A., Stillman, B. W., Nagata, K., Tamanoi, F., and Hurwitz, J.,** DNA sequence required for the *in vitro* replication of adenovirus DNA, *Proc. Natl. Acad. Sci. U.S.A.,* 81, 3069, 1984.
44. **Leegwater, P. A., vanDriel, W., and van der Vliet, P. C.,** Recognition site of nuclear factor 1, a sequence specific DNA-binding protein from HeLa cells that stimulates adenovirus DNA replication, *EMBO J.,* 4, 1515, 1984.
45. **Takahashi, K., Vigneron, M., Matthes, H., Wilderman, A., Zenke, M., and Chambon, P.,** Requirement of stereospecific alignments for initiation from the simian virus 40 early promoter, *Nature,* 319, 121, 1986.
46. **Wingender, E.,** Compilation of transcription regulating protein, *Nucleic Acids Res.,* 16, 1879, 1988.
47. **Johnson, A. D. and Herskowitz, I.,** A repressor (MATα2 product) and its operator control expression of a set of cell type specific genes in yeast, *Cell,* 42, 237, 1985.
48. **Brent, R.,** Repression of transcription in yeast, *Cell,* 42, 3, 1985.
49. **Guarente, L. and Hoar, E.,** Upstream activation sites of the CYC1 gene of *Saccharomyces cerevisiae* are active when inverted but not when placed upstream of the "TATA box", *Proc. Natl. Acad. Sci. U.S.A.,* 81, 7860, 1984.
50. **Struhl, K.,** Genetic properties and chromatin structure of the yeast GAL regulatory element: an enhancer-like sequence, *Proc. Natl. Acad. Sci. U.S.A.,* 81, 7865, 1984.
51. **Ogden, J., Stanway, C., Kim, S., Mellor, J., Kingsman, A., and Kingsman, S.,** Efficient expression of the Saccharomyces cerevisiae PGK gene depends on an upstream activation sequence, but does not require TATA sequences, *Mol. Cell. Biol.,* 6, 4335, 1986.
52. **Buratowski, S., Hahn, S., Sharp, P. A., and Guarente, L.,** Function of a yeast TATA element-binding protein in a mammalian transcription system, *Nature,* 334, 37, 1988.
53. **Davidson, B. L., Egly, J.-M., Mulvihill, E. R., and Chambon, P.,** Formation of stable preinitiation complexes between eukaryotic class B transcription factors and promoter sequences, *Nature,* 301, 680, 1983.
54. **Fire, A., Sanuels, M., and Sharp, P. A.,** Interactions between RNA polymerase II, factors, and template leading to accurate transcription, *J. Biol. Chem.,* 259, 2509, 1984.
55. **Hawley, D. K. and Roeder, R. G.,** Separation and partial characterization of three functional steps in transcription initiation by human RNA polymerase II, *J. Biol. Chem.,* 260, 8163, 1985.
56. **Hawley, D. K. and Roeder, R. G.,** Functional steps in transcription initiation and reinitiation from the major late promoter in a HeLa nuclear extract, *J. Biol. Chem.,* 262, 3452, 1987.
57. **Reinberg, D., Horikoshi, M., and Roeder, R. G.,** Factors involved in specific transcription by mammalian RNA polymerase II: functional analysis of initiation factors IIA and IID and identification of a new factor operating at sequences downstream of the initiation site, *J. Biol. Chem.,* 262, 3322, 1987.
58. **Matsui, T., Segall, J., Weil, P. A., and Roeder, R. G.,** Multiple factors required for accurate initiation of transcription by purified RNA polymerase II, *J. Biol. Chem.,* 255, 11992, 1980.
59. **Reinberg, D. and Roeder, R. G.,** Factors involved in specific transcription by mammalian RNA polymerase II: purification and functional analysis of initiation factors IIB and IIE, *J. Biol. Chem.,* 262, 3310, 1987.
60. **Sawadogo, M. and Roeder, R. G.,** Factors involved in specific transcription by human RNA polymerase II: analysis by a rapid and quantitative *in vitro* assay, *Proc. Natl. Acad. Sci. U.S.A.,* 82, 4394, 1985.
61. **Horikoshi, M., Hai, T., Lin, Y.-S., Green, M. R., and Roeder, R. G.,** Transcription factor ATF interacts with the TATA factor to facilitate establishment of a preinitiation complex, *Cell,* 54, 1033, 1988.
62. **Sawadogo, M. and Roeder, R. G.,** Interaction of a gene-specific transcription factor with the adenovirus major late promoter upstream of the TATA box region, *Cell,* 43, 165, 1985.
63. **Short, N. J.,** Flexible interpretation, *Nature,* 334, 192, 1988.
64. **Chodosh, L. A., Baldwin, A. S., Carthew, R. W., and Sharp, P. A.,** Human CCAAT-binding proteins have heterologous subunits, *Cell,* 53, 11, 1988.
65. **Hahn, S. and Guarente, L.,** Yeast HAP2 and HAP3: transcriptional activators in a heteromer complex, *Science,* 240, 317, 1988.
66. **Chodosh, L. A., Olesen, J., Hahn, S., Baldwin, A. S., Guarente, L., and Sharp, P. A.,** A yeast and a human CCAAT-binding protein have heterologous subunits that are functionally interchangeable, *Cell,* 53, 25, 1988.
67. **Kingston, R. E., Baldwin, A. S., and Sharp, P. A.,** Transcriptional control by oncogenes, *Cell,* 41, 3, 1985.
68. **DePamphilis, M. L.,** Transcriptional elements as components of eukaryotic origins of DNA replication, *Cell,* 52, 635, 1988.
69. **Chalifour, L. E., Wirak, D. O., Hansen, U., Wassarman, P. M., and Depamphilis, M. L.,** *cis-* and *trans-*acting sequences required for expression of simian virus 40 genes in mouse oocytes, *Genes Dev.,* 1, 1096, 1987.
70. **Danna, K. J. and Nathans, D.,** Bidirectional replication of SV40, *Proc. Natl. Acad. Sci. U.S.A.,* 69, 3097, 1972.
71. **Fareed, G. G., Garon, C. F., and Salzman, N. P.,** Origin and direction of simian virus 40 deoxyribonucleic acid replication, *J. Virol.,* 10, 484, 1972.
72. **Myers, R. M. and Tjian, R.,** Construction and analysis of simian virus 40 origins defective in tumor antigen binding and DNA replication, *Proc. Natl. Acad. Sci. U.S.A.,* 77, 6491, 1980.
73. **Tjian, R.,** The binding site on SV40 DNA for a T-antigen-related protein, *Cell,* 13, 165, 1978.
74. **Tjian, R.,** Protein-DNA interactions at the origin of simian virus 40 DNA replication, *Cold Spring Harbor Symp. Quant. Biol.,* 43, 655, 1978.
75. **Shalloway, D., Kleinberger, T., and Livingston, D. M.,** Mapping of SV40 DNA replication origin region binding sites for the SV40 T antigen by protection against exonuclease III digestion, *Cell,* 20, 411, 1980.
76. **Khoury, G., Martin, M. A., Lee, T. H. N., Danna, K. J., and Nathans, D.,** A map of SV40 transcription sites expressed in productively infected cells, *J. Mol. Biol.,* 78, 377, 1973.
77. **Sambrook, J., Sugden, B., Keller, W., and Sharp, P. A.,** Transcription of simian virus 40. III. Mapping of "early" and "late" species of RNA, *Proc. Natl. Acad. Sci. U.S.A.,* 70, 3711, 1973.
78. **Ghosh, P. K. and Lebowitz, P.,** SV40 early mRNAs contain multiple 5′-termini upstream and downstream from a Hogness-Goldberg sequence. A shift in 5′-termini during the lytic cycle is mediated by large T antigen, *J. Virol.,* 40, 224, 1981.
79. **Benoist, C. and Chambon, P.,** Deletions covering the putative promoter region of early mRNA of SV40 do not abolish T antigen expression, *Proc. Natl. Acad. Sci. U.S.A.,* 77, 3865, 1980.
80. **Rio, D., Robbins, A., Myers, R., and Tjian, R.,** Regulation of simian virus 40 early transcription *in vitro* by a purified tumor antigen, *Proc. Natl. Acad. Sci. U.S.A.,* 77, 5706, 1980.
81. **Ghosh, P. K., Lebowitz, P., Frisque, F. J., and Gluzman, Y.,** Identification of a promoter component involved in positioning the 5′ termini of simian virus 40 early mRNAs, *Proc. Natl. Acad. Sci. U.S.A.,* 78, 100, 1981.
82. **Myers, R. M., Rio, D. C., Robbins, A. K., and Tjian, R.,** SV40 gene expression is modulated by the cooperative binding of T antigen to DNA, *Cell,* 25, 373, 1981.
83. **Fromm, M. and Berg, P.,** Deletion mapping of DNA regions required

for SV40 early region promoter function *in vivo, J. Mol. Appl. Genet.*, 1, 457, 1982.

84. **Everett, R. D., Baty, D., and Chambon, P.,** The repeated GC-rich motifs upstream from the TATA box are important elements of the SV40 early promoter, *Nucleic Acids Res.*, 11, 2447, 1983.

85. **Gidoni, D., Kadonaga, J. T., Barrera-Saldana, H., Takahashi, K., Chambon, P., and Tjian, R.,** Bidirectional SV40 transcription mediated by tandem Sp1 binding interactions, *Science*, 230, 511, 1985.

86. **Dynan, W. S. and Tjian, R.,** Isolation of transcription factors that discriminate between different promoters recognized by RNA polymerase II, *Cell*, 32, 669, 1983.

87. **Dynan, W. S. and Tjian, R.,** The promoter-specific transcription factor Sp1 binds to upstream sequences in the SV40 early promoter, *Cell*, 35, 79, 1983.

88. **Dynan, W. S. and Tjian, R.,** Control of eukaryotic mRNA synthesis by sequence-specific DNA binding proteins, *Nature*, 316, 774, 1985.

89. **Barrera-Saldana, H., Takahashi, K., Vigneron, M., Wildeman, A., Davidson, I., and Chambon, P.,** All six GC-motifs of the SV40 early upstream element contribute to promoter activity *in vivo* and *in vitro, EMBO J.*, 4, 3839, 1985.

90. **Dynan, W. S., Saffer, J. D., Lee, W. S., and Tjian, R.,** Transcription factor Sp1 recognizes promoter sequences from the monkey genome that are similar to the simian virus 40 promoter, *Proc. Natl. Acad. Sci. U.S.A.*, 82, 4915, 1985.

91. **Dynan, W. S., Sazer, S., Tjian, R., and Schimke, R. T.,** The transcription factor Sp1 recognizes a DNA sequence in mouse dihydrofolate reductase promoter, *Nature*, 319, 246, 1986.

92. **Holler, M., Westin, G., Jiricny, J., and Schaffner, W.,** Sp1 transcription factor binds DNA and activates transcription even when the binding site is CpG methylated, *Genes Dev.*, in press.

93. **Herr, W. and Clarke, J.,** The SV40 enhancer is composed of multiple elements that can compensate for one another, *Cell*, 45, 461, 1986.

94. **Herr, W. and Gluzman, Y.,** Duplications of a mutated simian virus 40 enhancer restore its activity, *Nature*, 313, 711, 1985.

95. **Clarke, J. and Herr, W.,** Activation of mutated simian virus 40 enhancers by amplification of wild-type enhancer elements, *J. Virol.*, 61, 3536, 1987.

96. **Ondek, B. A., Shepard, A., and Herr, W.,** Discrete elements within the SV40 enhancer region display different cell-specific enhancer activities, *EMBO J.*, 6, 1017, 1987.

97. **Schirm, S., Jiricny, J., and Schaffner, W.,** The SV40 enhancer can be dissected into multiple segments, each with a different cell type specificity, *Genes Dev.*, 1, 65, 1987.

98. **Weiher, H., Konig, M., and Gruss, P.,** Multiple point mutations affecting the simian virus 40 enhancer, *Science*, 219, 626, 1983.

99. **Zenke, M., Grundstrom, T., Matthes, H., Wintzerith, M., Schatz, C., Wildeman, A., and Chambon, P.,** Multiple sequence motifs are involved in SV40 enhancer function, *EMBO J.*, 5, 387, 1986.

100. **Pinkert, C. A., Ornitz, D. M., Brinster, R. L., and Palmiter, R. D.,** An albumin enhancer located 10 kb upstream functions along with its promoter to direct efficient, liver-specific expression in transgenic mice, *Genes Dev.*, 1, 268, 1987.

101. **Fromental, C., Kanno, M., Nomiyama, H., Ruffenach, F., and Chambon, P.,** Cooperativity and hierarchical levels of functional organization in the SV40 enhancers, *Cell*, 54, 943, 1988.

102. **Davidson, I., Xiao, J. H., Rosales, R., Staub, A., and Chambon, P.,** The HeLa cell protein TEF-1 binds specifically and cooperatively to two SV40 enhancer motifs of unrelated sequence, *Cell*, 54, 931, 1988.

103. **Wildeman, A., Zenke, M., Schatz, C., Wintzerith, M., Grundstrom, T., Matthes, H., Takahashi, K., and Chambon, P.,** Specific protein binding to the simian virus 40 enhancer *in vitro, Mol. Cell. Biol.*, 6, 2098, 1986.

104. **Rosales, R., Vigneron, M., Macchi, M., Davidson, I., Xian, J. H., and Chambon, P.,** *In vitro* binding of cell specific and ubiquitous nuclear proteins to the octamer motif of the SV40 enhancer and related motifs present in other promoters and enhancers, *EMBO J.*, 6, 3015, 1987.

105. **Sen, R. and Baltimore, D.,** Multiple nuclear factors interact with the immunoglobulin enhancer sequences, *Cell*, 46, 705, 1986.

106. **Nomiyama, H., Fromental, C., Xiao, J. H., and Chambon, P.,** Cell specific activity of the constituent elements of the simian virus 40 enhancer, *Proc. Natl. Acad. Sci. U.S.A.*, 84, 7881, 1987.

107. **Schaffner, G., Schirm, S., Muller-Baden, B., Weber, F., and Schaffner, W.,** Redundancy of information in enhancers as a principle of mammalian transcription control, *J. Mol. Biol.*, 201, 81, 1988.

108. **Nevins, J. R.,** Regulation of early adenovirus gene expression, *Microbiol. Rev.*, 51, 419, 1987.

109. **Kovesdi, I., Reichel, R., and Nevins, J. R.,** Role of an adenovirus E2 promoter binding factor in E1A-mediated coordinate gene control, *Proc. Natl. Acad. Sci. U.S.A.*, 84, 2180, 1987.

110. **Raychaudhuri, P., Rooney, R., and Nevins, J. R.,** Identification of an E1A-inducible cellular factor that interacts with regulatory sequences within the adenovirus E4 promoter, *EMBO J.*, 6, 4073, 1987.

111. **Kovesdi, I., Reichel, R., and Nevins, J. R.,** Identification of a cellular transcription factor involved in E1A trans-activation, *Cell*, 45, 219, 1986.

112. **Simon, M. C., Fisch, T. M., Benecke, B. J., Nevins, J. R., and Heintz, N.,** Definition of multiple, functionally distinct TATA elements, one of which is a target in the hsp70 promoter for E1A regulation, *Cell*, 52, 723, 1988.

113. **Kovesdi, I., Satake, M., Furukawa, K., Reichel, R., Ito, Y., and Nevins, J. R.,** A factor discriminating between wild-type and mutant polyomavirus enhancer, *Nature*, 328, 87, 1987.

114. **Wang, J. C. and Giaever, G. N.,** Action at a distance along a DNA, *Science*, 240, 300, 1988.

115. **Williamson, P. and Felsenfeld, G.,** Transcription of histone-covered T7 DNA by *Escherichia coli* RNA polymerase, *Biochemistry*, 17, 5695, 1978.

116. **Wasylyk, B., Thevenin, G., Oudet, P., and Chambon, P.,** Transcription of *in vitro* assembled chromatin by *Escherichia coli* RNA polymerase, *J. Mol. Biol.*, 128, 411, 1979.

117. **Lorch, Y., Lapointe, J. W., and Kornberg, R. D.,** Nucleosomes inhibit the initiation of transcription but allow chain elongation with the displacement of the histones, *Cell*, 49, 203, 1987.

118. **Pflugfelder, G. and Sonnenbichler, J.,** Oligonucleosomes as a model system for chromatin transcription. Transcription with *Escherichia coli* DNA-dependent RNA polymerase, *J. Mol. Biol.*, 158, 365, 1982.

119. **Hochschild, A. and Ptashne, M.,** Homologous interactions of λ cro with λ operator, *Cell*, 44, 925, 1986.

120. **Ptashne, M.,** Gene regulation by proteins acting nearby and at a distance, *Nature*, 322, 697, 1986.

121. **Griffith, J., Hochschild, A., and Ptashne, M.,** DNA loops induced by cooperative binding of λ repressor, *Nature*, 322, 750, 1986.

122. **Mossing, M. C. and Record, M. T.,** Upstream operators enhance repression of the lac promoter, *Science*, 233, 889, 1986.

123. **Bellomy, G. R., Mossing, M. C., and Record, M. T., Jr.,** Physical properties of DNA in vivo as probed by the length dependence of the lac operator looping process, *Biochemistry*, 27, 3900, 1988.

124. **Allison, L., Moyle, M., Shales, M., and Ingles, J.,** Extensive homology among the largest subunits of eukaryotic and prokaryotic RNA polymerase, *Cell*, 42, 599, 1985.

125. **Corden, J., Cadena, D., Ahearn, J., and Dahmus, M.,** A unique structure at the carboxyl terminus of the largest subunit of eukaryotic RNA polymerase II, *Proc. Natl. Acad. Sci. U.S.A.*, 79, 34, 1985.

126. **Schleif, R.,** DNA looping, *Science*, 240, 127, 1988.

127. **Sarai, A., Mazur, J., Nussinov, R., and Jernigan, R. L.,** Sequence dependence of DNA conformations: means and fluctuations, in *Structure and Expression in DNA Bending and Curvature*, Vol. 3, Olson, W. K., Sarma, M. H., Sarma, R., and Sundaralingam, M., Eds., Adenine Press, New York, 1988, 213.

128. **Saenger, W.,** Principles of nucleic acid structure, Springer-Verlag, New York, 1984.

129. **Maroun, R. C. and Olson, W. K.,** Base sequence effects in double helical DNA II. Configurational statistics of rodlike chains, *Biopolymers*, 27, 561, 1988.

130. **Maroun, R. C. and Olson, W. K.,** Base sequence effects in double helical DNA III. Average properties of curved DNA, *Biopolymers*, 27, 585, 1988.

131. **McClarin, J. A., Frederick, C. A., Wang, B.-C., Green, P., Boyer, H. W., Grable, J., and Rosenberg, J. M.,** Structure of the DNA-EcoRI endonuclease recognition complex at 3 Å resolution, *Science*, 234, 1526, 1986.

132. **Sarai, A., Mazur, J., Nussinov, R., and Jernigan, R. L.,** Origin of DNA helical structure and its sequence dependence, *Biochemistry*, 27, 8499, 1988.

133. **Arnott, S., Dover, S. D., and Wonacott, A. J.,** Least-squares refinement of the crystal and molecular structures of DNA and RNA from X-ray data and standard bond lengths and angles, *Acta Crystallogr.*, B25, 2192, 1969.

134. **Dickerson, R. E. and Drew, H. R.,** Structure of a B-DNA dodecamer. II. Influence of base sequence on helix structure, *J. Mol. Biol.*, 149, 761, 1981.

135. **Shakked, Z. and Rabinovich, D.,** The effect of the base sequence on the fine structure of the DNA double helix, *Prog. Biophys. Mol. Biol.*, 41, 159, 1986.

136. **Koo, H.-S., Wu, H.-M., and Crothers, D. M.,** DNA bending at adenine-thymine tracts, *Nature*, 320, 501, 1986.

137. **Hagerman, P. J.,** Sequence-directed curvature of DNA, *Nature*, 321, 449, 1986.

138. **Bossi, L. and Smith, D. M.,** Conformational change in the DNA associated with an unusual promoter mutation in a tRNA operon of Salmonella, *Cell*, 39, 643, 1984.

139. **Ryder, K., Silver, S., DeLucia, A. L., Fanning, E., and Tegtmeyer, P.,** An altered DNA conformation in origin region I is a determinant for the binding of SV40 large T antigen, *Cell*, 44, 719, 1986.

140. **Snyder, M., Buchman, A. R., and Davis, R. W.,** Bent DNA at a yeast autonomously replicating sequence, *Nature*, 324, 87, 1986.

141. **Zahn, K. and Blattner, F. R.,** Direct evidence for DNA bending at the Lambda replication origin, *Science*, 236, 416, 1987.

142. **Calladine, C. R.,** Mechanics of sequence-dependent stacking of bases in B-DNA, *J. Mol. Biol.*, 161, 343, 1982.

143. **Srinivasan, A. R., Torres, R., Clark, W., and Olson, W. K.,** Base sequence effects in double helical DNA I. Potential energy estimates of local base morphology, *J. Biomol. Struct. Dyn.*, 5, 459, 1987.

144. **Dickerson, R. E.,** Base sequence and helix variation in B and A DNA, *J. Mol. Biol.*, 166, 419, 1983.

145. **Olson, W. K. and Srinivasan, A. R.,** The translation of DNA primary base sequence into three-dimensional structure, *Comput. Appl. Biosci.*, 4, 133, 1988.

146. **Milton, D. L. and Gesteland, R. F.,** Bends in SV40 DNA: use of mutagenesis to identify the critical bases in involved, *Nucleic Acids Res.*, 16, 3931, 1988.

147. **Hsieh, C.-H. and Griffith, J. D.,** The terminus of SV40 replication and transcription contains a sharp sequence-directed curve, *Cell*, 52, 535, 1988.

148. **Plaskan, R. R. and Wartell, R. M.,** Sequence distributions associated with DNA curvature are found upstream of strong *E. coli* promoters, *Nucleic Acids Res.*, 15, 785, 1987.

149. **Thompson, R., Taylor, L., Kelly, K., Everett, R., and Willets, N.,** The F plasmid origin of transfer DNA sequence of mutant origins and location of origin-specific nicks, *EMBO J.*, 3, 1175, 1984.

150. **Koepsel, R. R. and Khan, S. A.,** Static and initiator protein-enhanced bending of DNA at a replication origin, *Science*, 233, 1316, 1986.

151. **Inokuchi, K., Nakayoma, A., and Hishinuma, F.,** Sequence-directed bends of DNA helix at the upstream activation sites of α-cell-specific genes in yeast, *Nucleic Acids Res.*, 16, 6693, 1988.

152. **Fujimura, F. K.,** Point mutation in the polyoma enhancer alters local DNA conformation, *Nucleic Acids Res.*, 16, 1987, 1988.

153. **McGhee, J. D., Wood, W. I., Dolan, M., Engel, J. D., and Felsenfeld, G.,** A 200 base pair region at the 5′ end of the chicken adult β-globin gene is accessible to nuclease digestion, *Cell*, 27, 45, 1981.

154. **Larsen, A. and Weintraub, H.,** An altered DNA conformation detected by S1 nuclease occurs at specific regions in active chick globin chromatin, *Cell*, 29, 609, 1982.

155. **Kohwi-Shigematsu, T. and Kohwi, Y.,** Poly(dG)-poly(dC) sequences, under torsional stress induce an altered DNA conformation upon neighboring DNA sequences, *Cell*, 43, 199, 1985.

156. **Anderson, J. N.,** Detection, sequence patterns and function of unusual DNA structures, *Nucleic Acids Res.*, 14, 8513, 1986.

157. **Olson, W. K.,** DNA sequence and structure, in *Structure and Expression in DNA Bending and Curvature*, Vol. 3, Olson, W. K., Sarma, M. H., Sarma, R., and Sundaralingam, M., Eds., Adenine Press, New York, 1988, 1.

158. **Ostrander, E. A., Karty, R. A., and Hallick, L. M.,** High resolution psoralen mapping reveals an altered DNA helical structure in the SV40 regulatory region, *Nucleic Acids Res.*, 16, 2B, 1988.

159. **Zhou, N., Bianucci, A. M., Pattabiraman, N., and James, T. L.,** Solution structure of [d(GGTATACC)]$_2$: wrinkled D structure of the TATA moiety, *Biochemistry*, 26, 7905, 1987.

160. **Arnott, S., Chandrasekaran, R., Hall, I. H., and Puigjaner, L. C.,** Heteronomous DNA, *Nucleic Acids Res.*, 11, 4141, 1983.

161. **Suzuki, E.-I., Pattabiraman, N., Zon, G., and James, T. L.,** Solution structure of [d(A-T)$_5$]$_2$ via complete relaxation matrix analysis of two-dimensional nuclear overhauser effect spectra and molecular mechanics calculations: evidence for a hydration tunnel, *Biochemistry*, 25, 6854, 1986.

162. **Nussinov, R., Shapiro, B., Lipkin, L. E., and Maizel, J. V.,** DNAase I hyper-sensitive sites may be correlated with genomic regions of large structural variation, *J. Mol. Biol.*, 177, 591, 1984.

163. **Guarente, L., Lalonde, B., Gifford, P., and Alani, E.,** Distinctly regulated tandem upstream activation sites mediate catabolite repression of the CYC1 gene of *Saccharomyces cerevisiae*, *Cell*, 36, 503, 1984.

164. **Nussinov, R.,** Structural wrinkles and eukaryotes genomic regulatory sites, *J. Mol. Evol.*, 22, 150, 1986.

165. **Nussinov, R.,** Large helical conformational deviation from ideal B-DNA and prokaryotic regulatory sites, *J. Theor. Biol.*, 115, 179, 1985.

166. **Satchwell, S. C., Drew, H. R., and Travers, A. A.,** Sequence periodicities in chicken nucleosome core DNA, *J. Mol. Biol.*, 191, 659, 1986.

167. **Travers, A.,** DNA bending and nucleosome positioning, *Trends Biochem. Sci.*, 12, 108, 1987.

168. **Travers, A. and Klug, A.,** DNA wrapping and writhing, *Nature*, 327, 280, 1987.

169. **Nelson, H. C. M., Finch, J. T., Luisi, B. F., and Klug, A.,** The structure of an oligo (dA) · oligo (dT) tract and its biological implications, *Nature*, 330, 221, 1987.

170. **Coll, M., Frederick, C. A., Wang, A. H.-J., and Rich, A.,** A bifurcated hydrogen-bonded conformation in the d(A·T) base pairs of the DNA dodecamer d(CGCAAATTTGCG) and its complex with distamycin, *Proc. Natl. Acad. Sci. U.S.A.*, 84, 8385, 1987.

171. **Dickerson, R. E., Goodsell, D. S., Kopka, M. L., and Pjura, P. E.,** The effect of crystal packing on oligonucleotide double helix structure, *J. Biomol. Struct. Dyn.*, 5, 557, 1987.

172. **Burkhoff, A. M. and Tullius, T. D.,** The unusual conformation adopted by the adenine tracts in kinetoplast DNA, *Cell*, 48, 935, 1987.

173. **Drew, H. R. and Travers, A. A.,** DNA bending and its relation to nucleosome positioning, *J. Mol. Biol.*, 186, 773, 1985.

174. **Suck, D., Lahm, A., and Oefner, C.,** Structure refined to 2 Å of a nicked DNA octanucleotide complex with DNAase I, *Nature*, 332, 464, 1988.

175. **Dickerson, R. E., Kopka, M. L., and Pjura, P. E.,** The major and minor grooves of the DNA helix as conduits for information transfer, in *DNA-Ligand Interactions*, Guschlbauer, W. and Saenger, W., Eds., Plenum Press, New York, 1987, 45.

176. **Otwinowski, Z., Schevitz, R. W., Zhang, R.-G., Lawson, C. L., Joachimiak, A., Marmorstein, R. Q., Luisi, B. F., and Sigler, P. B.,** Crystal structure of trp repressor/operator complex at atomic resolution, *Nature*, 335, 321, 1988.

177. **Oleson, J., Hahn, S., and Guarente, L.,** Yeast HAP2 and HAP3 activators both bind to the CYC1 upstream activation site, UAS2, in an interdependent manner, *Cell*, 51, 953, 1987.

178. **Bram, R. J. and Kornberg, R. D.,** Specific protein binding to far upstream activating sequences in polymerase II promoters, *Proc. Natl. Acad. Sci. U.S.A.*, 82, 43, 1985.

179. **Giniger, E., Varnum, S. M., and Ptashne, M.,** Specific DNA binding of GAL4, a positive regulatory protein of yeast, *Cell*, 40, 767, 1985.

180. **Hope, I. and Struhl, K.,** GCN4 protein, synthesized *in vitro*, binds HIS4 regulatory sequences: implications for the general control of amino acid biosynthetic genes in yeast, *Cell*, 43, 177, 1985.

181. **Arndt, K. and Fink, G. R.,** GCN4 protein, a positive transcription factor in yeast, binds general control promoters at all 5′ TGACTC 3′ sequences, *Proc. Natl. Acad. Sci. U.S.A.*, 83, 8516, 1986.

182. **Barberis, A., Superti-Furga, G., and Busslinger, M.,** Mutually exclusive interaction of the CCAAT binding factor and of a displacement protein with overlapping sequences of a histone gene promoter, *Cell*, 50, 347, 1987.

183. **Dorn, A., Bollekens, J., Staub, A., Benoist, C., and Matthis, D.,** A multiplicity of CCAAT box-binding proteins, *Cell*, 50, 863, 1987.

184. **Keegan, L., Gill, G., and Ptashne, M.,** Separation of DNA binding from the transcription-activating function of a eukaryotic regulatory protein, *Science*, 231, 699, 1986.

185. **Hill, D. E., Hope, I. A., Macke, J. P., and Struhl, K.,** Saturation mutagenesis of the yeast his3 regulatory site: requirements for transcriptional induction and for binding by GCN4 activator protein, *Science*, 234, 451, 1986.

186. **Chandler, V. L., Maler, B. A., and Yamamoto, K. R.,** DNA sequences bound specifically by glucocorticoid receptor *in vitro* render a heterologous promoter responsive *in vivo*, *Cell*, 33, 489, 1983.

187. **Ponta, H., Kennedy, N., Skroch, P., Hynes, N. E., and Groner, B.,** Hormonal response region in the mouse mammary tumor virus long terminal repeat can be dissociated from the proviral promoter and has enhancer properties, *Proc. Natl. Acad. Sci. U.S.A.*, 82, 1020, 1985.

188. **Renkawitz, R., Shutz, G., von der Ahe, D., and Beato, M.,** Sequences in the promoter region of the chicken lysozyme gene required for steroid regulation and receptor binding, *Cell*, 37, 503, 1984.

189. **Karin, M., Haslinger, A., Holtgreve, A., Richards, R. I., Krauter, P., Westphal, H. M., and Beato, M.,** Characterization of DNA sequences through which cadmium and glucocorticoid hormones induce human metallothionein-II$_A$ gene, *Nature*, 308, 513, 1984.

190. **Slater, E. P., Rabenau, O., Karin, M., Baxter, J. D., and Beato, M.,** Glucocorticoid receptor binding and activation of a heterologous promoter by dexamethasone by the first intron of the human growth hormone gene, *Mol. Cell. Biol.*, 5, 2984, 1985.

191. **Jantzen, H.-M., Strahle, U., Gloss, B., Stewart, F., Schmid, W., Boshart, M., Miksicek, R., and Schutz, G.,** Cooperativity of glucocorticoid response elements located far upstream of the tyrosine amino-transferase gene, *Cell*, 49, 29, 1987.

192. **Klein-Hitpass, L., Schorpp, M., Wagner, U., and Ryffel, G.,** An estrogen-responsive element derived from the 5′ flanking region of the *Xenopus* vitellogenin A2 gene functions in transfected human cells, *Cell*, 46, 1053, 1986.

193. **Payvar, F., DeFranco, D., Firestone, G. L., Edgar, B., Wrange, O., Okret, S., Gustafsson, J.-A., and Yamamoto, K. R.,** Sequence-specific binding of glucocorticoid receptor to MTV DNA at sites within and upstream of the transcribed region, *Cell*, 35, 381, 1983.

194. **von der Ahe, D., Janich, S., Scheidereit, C., Renkawitz, R., Schutz, G., and Beato, M.,** Glucocorticoid and progesterone receptors bind to the same sites in two hormonally regulated promoters, *Nature*, 313, 706, 1985.

195. **Cato, A. C. B., Miksicek, R., Schutz, G., Arnemann, J., and Beato, M.,** The hormone regulatory element of mouse mammary tumor virus mediates progesterone induction, *EMBO J.*, 5, 2237, 1986.

196. **Jost, J.-P., Sledran, M., and Geiser, M.,** Preferential binding of estrogen-receptor complex of a region containing the estrogen-dependent hypomethylation site preceding the chicken vitellogenin II gene, *Proc. Natl. Acad. Sci. U.S.A.*, 81, 429, 1984.

197. **Godowski, P. J., Picard, D., and Yamamoto, K. R.,** Signal transduction and transcriptional regulation by glucocorticoid receptor-LexA fusion proteins, *Science*, 241, 812, 1988.

198. **Miesfeld, R., Rusconi, S., Godowski, P. J., Maler, B. A., Okret, S., Wikstrom, A.-C., Gustafsson, J.-A., and Yamamoto, K. R.,** Genetic complementation of a glucocorticoid receptor deficiency by expression of cloned receptor cDNA, *Cell*, 46, 389, 1986.

199. **Druege, P. M., Klein-Hitpass, L., Green, S., Stack, G., Chambon, P., and Ryffel, G. V.,** Introduction of estrogen-responsiveness into mammalian cell lines, *Nucleic Acids Res.*, 14, 9329, 1986.

200. **Giguere, S. M., Rosenfeld, M. G., and Evans, R. M.,** Functional domains of the human glucocorticoid receptor, *Cell*, 46, 645, 1986.

201. **Danielsen, M., Northrop, J. P., Jonklaas, J., and Ringold, G. M.,** Domains of the glucocorticoid receptor involved in specific and nonspecific deoxyribonucleic acid binding, hormone activation, and transcriptional enhancement, *Mol. Endocrinol.*, 1, 816, 1987.

202. **Godowski, P. J., Rusconi, S., Miesfeld, R., and Yamamoto, K. R.,** Glucocorticoid receptor mutants that are constitutive activators of transcriptional enhancement, *Nature*, 325, 365, 1987.

203. **Rusconi, S. and Yamamoto, K. R.,** Functional dissection of the hormone and DNA binding activities of the glucocorticoid receptor, *EMBO J.*, 6, 1309, 1987.

204. **Kumar, V., Green, S., Staub, A., and Chambon, P.,** Localization of the oestradiol-binding and putative DNA-binding domains of the human oestrogen receptor, *EMBO J.*, 5, 2231, 1986.

205. **Freedman, L. P., Luisi, B. F., Korszun, Z. R., Basavappa, R., Sigler, P. B., and Yamamoto, K. R.,** The function and structure of the metal coordination sites within the glucocorticoid receptor DNA binding domain, *Nature*, 334, 543, 1988.

206. **Green, S. and Chambon, P.,** Oestradiol induction of a glucocorticoid-responsive gene by a chimaeric receptor, *Nature*, 325, 75, 1987.

207. **Kumar, V., Green, S., Stack, G., Berry, M., Jin, J.-R., and Chambon, P.,** Functional domains of the human estrogen receptor, *Cell*, 51, 941, 1987.

208. **Tsai, S. Y., Sagami, I., Wang, H., Tsai, M. J., and O'Malley, B. W.,** Interactions between a DNA-binding transcription factor (COUP) and a non-DNA binding factor (S300-II), *Cell*, 50, 701, 1987.

209. **Distel, R. J., Ro, H.-S., Rosen, B. S., Groves, D. L., and Spie-**

gelman, B. M., Nucleoprotein complexes that regulate gene expression in adipocyte differentiation: direct participation of *c-fos*, *Cell*, 49, 835, 1987.

210. **Verma, I. M. and Sasson-Corsi, P.**, Proto-oncogene fos: complex but versatile regulation, *Cell*, 51, 513, 1987.

211. **Ma, J. and Ptashne, M.**, Deletion analysis of GAL4 defines two transcriptional activating segments, *Cell*, 48, 847, 1987.

212. **Ma, J. and Ptashne, M.**, A new class of yeast transcriptional activators, *Cell*, 51, 113, 1987.

213. **Gill, G. and Ptashne, M.**, Mutants of GAL4 protein altered in an activation function, *Cell*, 51, 121, 1987.

214. **Sadowski, I., Ma, J., Triezenberg, S., and Ptashne, M.**, GAL4-VP16 is an unusually potent transcriptional activator, *Nature*, 335, 563, 1988.

215. **Zehring, W. A., Lee, J. M., Weeks, J. R., Jokers, R. S., and Greenleaf, A. L.**, The C-terminal repeat domain of RNA polymerase II largest subunit is essential *in vivo* but is not required for accurate transcription initiation *in vitro*, *Proc. Natl. Acad. Sci. U.S.A.*, 85, 3698, 1988.

216. **Bartholomei, M. S., Halden, N. F., Cullen, C. R., and Corden, J. L.**, Genetic analysis of the repetitive carboxyl-terminal domain of the largest subunit of mouse RNA polymerase II, *Mol. Cell. Biol.*, 8, 330, 1988.

217. **Allison, L. A., Wong, J. K.-C., Fitzpatrick, D., Moyle, M., and Ingles, C. J.**, The C-terminal domain of the largest subunit of RNA polymerase II of *Saccharomyces cerevisiae, Drosophila melanogaster* and mammals: a conserved structure with an essential function, *Mol. Cell. Biol.*, 8, 321, 1988.

218. **Courey, A. J. and Tjian, R.**, Analysis of Sp1 *in vivo* reveals multiple transcriptional domains, including a novel glutamine-rich activation motif, *Cell*, 55, 887, 1988.

219. **Klug, A. and Rhodes, D.**, "Zinc fingers": a novel protein motif for nucleic acid recognition, *Trends Biochem. Sci.*, 12, 464, 1987.

220. **Pabo, C. O. and Sauer, R. T.**, Protein-DNA recognition, *Annu. Rev. Biochem.*, 53, 293, 1984.

221. **Landschulz, W. H., Johnson, P. F., and McKnight, S. L.**, The leucine zipper: a hypothetical structure common to a new class of DNA binding proteins, *Science*, 240, 1759, 1988.

222. **Jordan, S. R. and Pabo, C. O.**, Structure of the Lambda complex at 2.5 Å resolution: details of the repressor-operator interactions, *Science*, 242, 893, 1988.

223. **Tullius, T. and Dombroski, B. A.**, Hydroxyl radical footprinting: high resolution information about DNA-protein contacts and application of λ repressor and cro protein, *Proc. Natl. Acad. Sci. U.S.A.*, 83, 5469, 1986.

224. **Aggarwal, A. K., Rodgers, D. W., Drottar, M., Ptashne, M., and Harrison, S. C.**, Recognition of a DNA operator by the repressor of phage 434: a view at high resolution, *Science*, 242, 899, 1988.

225. **Richmond, T. J., Finch, J. T., Rushton, B., Rhodes, D., and Klug, A.**, Structure of the nucleosome core particle at 7Å resolution, *Nature*, 311, 532, 1984.

226. **Anderson, J. E., Ptashne, M., and Harrison, S. C.**, Structure of the repressor-operator complex of bacteriophage 434, *Nature*, 326, 846, 1987.

227. **Matthews, B. W.**, No code for recognition, *Nature*, 335, 294, 1988.

228. **Gehring, W. J.**, The homoeo box: a key to the understanding of development?, *Cell*, 40, 3, 1985.

229. **Levine, M. and Hoey, T.**, Homoeobox proteins as sequence-specific transcription factors, *Cell*, 55, 537, 1988.

230. **Thali, M., Muller, M. M., DeLorenzi, M., Matthias, P., and Bienz, M.**, Drosophila homoeotic genes encode transcriptional activators similar to mammalian OTF-2, *Nature*, 336, 598, 1988.

231. **Sturm, R. A. and Herr, W.**, The POU domain is a bipartite DNA-binding structure, *Nature*, 336, 601, 1988.

232. **Robertson, M.**, Homeo boxes, POU proteins and the limits to promiscuity, *Nature*, 336, 522, 1988.

233. **Mihara, H. and Kaiser, E. T.**, A chemically synthesized Antennapedia homeo domain binds to a specific DNA sequence, *Science*, 242, 925, 1988.

234. **Miller, J., McLachlan, A. D., and Klug, A.**, Repetitive zinc-binding domains in the protein transcription factor IIIA from *Xenopus* oocytes, *EMBO J.*, 4, 1609, 1985.

235. **Diakun, G. P., Fairall, L., and Klug, A.**, EXAFS study of the zinc-binding sites in the protein transcription factor IIIA, *Nature*, 324, 698, 1986.

236. **Frankel, A. D., Berg, J. M., and Pabo, C. O.**, Metal-dependent folding of a single zinc finger from transcription factor IIIA, *Proc. Natl. Acad. Sci. U.S.A.*, 84, 4841, 1987.

237. **Evans, R. M. and Hollenberg, S. M.**, Zinc fingers: gilt by association, *Cell*, 52, 1, 1988.

238. **Berg, J. M.**, Potential metal-binding domain proteins, *Science*, 232, 485, 1986.

239. **Berg, J. M.**, Proposed structure for the zinc-binding domains from transcription factor IIIA and related proteins, *Proc. Natl. Acad. Sci. U.S.A.*, 85, 99, 1988.

240. **Kadonaga, J. T., Carner, K. R., Masiarz, F. R., and Tjian, R.**, Isolation of cDNA encoding transcription factor Sp1 and functional analysis of the DNA binding domain, *Cell*, 51, 1079, 1987.

241. **Westin, G. and Schaffner, W.**, Heavy metal ions in transcription factors from Hela cells: Sp1, but not octamer transcription factor requires zinc for DNA binding and for activator function, *Nucleic Acids Res.*, in press.

242. **Huckaby, C. S., Beattie, O. M., Dobson, A. D. W., Tsai, M.-J., and O'Malley, B. W.**, Structure of the chromosomal chicken progesterone receptor gene, *Proc. Natl. Acad. Sci. U.S.A.*, 84, 8380, 1987.

243. **Severne, Y., Wieland, S., Schaffner, W., and Rusconi, S.**, Metal binding "finger" structures in the glucocorticoid receptor defined by site-directed mutagenesis, *EMBO J.*, in press.

244. **Furey, W. F., Robbins, A. H., Clancy, L. L., Winge, D. R., Wang, B. C., and Stout, C. D.**, Crystal structure of Cd, Zn metallothionein, *Science*, 231, 704, 1986.

245. **Frankel, A. D. and Pabo, C. O.**, Fingering too many proteins, *Cell*, 53, 675, 1988.

246. **Richmond, T. J.**, Another folding problem?, *Nature*, 326, 18, 1987.

247. **Lomonossoff, G. P., Butler, P. J. G., and Klug, A.**, Sequence dependent variations in the conformation of DNA, *J. Mol. Biol.*, 149, 745, 1981.

248. **Drew, H. R. and Travers, A. A.**, DNA structural variations in the *E. coli* tyrT promoter, *Cell*, 37, 491, 1984.

249. **Jernigan, R. L. and Nash, H.**, personal communication.

250. **Anderson, J. E., Ptashne, M., and Harrison, S. C.**, The structure of a phage repressor-operator complex at 7 Å resolution, *Nature*, 316, 596, 1985.

251. **Steitz, T., Ohlendorf, D. H., McKay, D. B., Anderson, W. F., and Matthews, B. W.**, Structural similarity in DNA-binding domains of catabolite gene activator and cro repressor proteins, *Proc. Natl. Acad. Sci. U.S.A.*, 79, 3097, 1982.

252. **Ohlendorf, D. H., Anderson, W. F., Fisher, R. G., Takeda, Y., and Matthews, B. W.**, The molecular basis of DNA-protein recognition inferred from the structure of the cro repressor, *Nature*, 298, 718, 1982.

253. **Nakabeppu, Y., Ryder, K., and Nathans, D.**, DNA binding activities of three murine jun proteins: stimulation by fos, *Cell*, 52, 907, 1988.

254. **Halazonetis, T. D., Georgopoulos, K., Greenberg, M. E., and Leder, P.**, *c-jun* dimerizes with itself and with *c-fos*, forming complexes of different DNA binding affinities, *Cell*, 52, 917, 1988.

255. **Marx, J. L.**, jun is bustin' out all over, *Science*, 242, 1377, 1988.

256. **Kouzarides, T. and Ziff, E.,** The role of the leucine zipper in the fos-jun interaction, *Nature,* 336, 646, 1988.

257. **Sassone-Corsi, P., Ransone, L. J., Lamph, W. W., and Verma, I. M.,** Direct interaction between the fos and jun nuclear oncoproteins: role of "leucine zipper" domain, *Nature,* 336, 962, 1988.

258. **Hogan, M. E. and Austin, R. H.,** Importance of DNA stiffness in protein-DNA binding specificity, *Nature,* 329, 263, 1987.

259. **McGhee, J. D. and Felsenfeld, G.,** Nucleosome structure, *Annu. Rev. Biochem.,* 49, 1115, 1980.

260. **Weintraub, H.,** Assembly and propagation of repressed and derepressed chromosomal states, *Cell,* 42, 705, 1985.

261. **Pederson, D. S., Thoma, F., and Simpson, R. T.,** Core particle, fiber, and transcriptionally active chromatin structure, *Annu. Rev. Cell. Biol.,* 2, 117, 1986.

262. **Solomon, M. J., Larsen, P. L., and Varshavsky, A.,** Mapping protein-DNA interactions *in vivo* with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene, *Cell,* 53, 937, 1988.

263. **Wu, C., Wong, Y.-C., and Elgin, S. C. R.,** The chromatin structure of specific genes. II. Disruption of chromatin structure during gene activity, *Cell,* 16, 807, 1979.

264. **Levy, A. and Noll, M.,** Chromatin fine structure of active and repressed genes, *Nature,* 289, 198, 1981.

265. **Levinger, L. and Varshavsky, A.,** Selective arrangement of ubiquitinated and D1 protein-containing nucleosomes within the *Drosophila* genome, *Cell,* 28, 375, 1982.

266. **Wu, C.,** The 5′ ends of *Drosophila* heat shock genes in chromatin are hypersensitive to DNAase I, *Nature,* 286, 854, 1980.

267. **Wu, C.,** Two protein-binding sites in chromatin implicated in the activation of heat shock genes, *Nature,* 309, 229, 1984.

268. **Costlow, N. and Lis, J. T.,** High resolution mapping of DNAase I-hypersensitive sites in Drosophila heat shock genes in *Drosophila melanogaster* and *Saccharomyces cerevisiae, Mol. Cell. Biol.,* 4, 1853, 1984.

269. **Brown, M. E., Amin, J., Schiller, P., Voellmy, R., and Scott, W. A.,** Determinants for DNAase I-hypersensitive chromatin structure 5′ to a human hsp70 gene, *J. Mol. Biol.,* 203, 107, 1988.

270. **Scott, W. A. and Wigmore, D. J.,** Sites in SV40 chromatin which are preferentially cleaved by endonucleases, *Cell,* 15, 1511, 1978.

271. **Choder, M., Bratosin, S., and Aloni, Y.,** A direct analysis of transcribed minichromosomes: all transcribed SV40 minichromosomes have a nuclease-free domain, *EMBO J.,* 3, 2929, 1984.

272. **Varshavsky, A. J., Sundin, O. H., and Bohn, M. J.,** A stretch of late SV40 viral DNA about 400 bp long which includes the origin of replication is specifically exposed in SV40 minichromosomes, *Cell,* 16, 453, 1979.

273. **Saragosti, S., Moyne, G., and Yaniv, M.,** Absence of nucleosomes in a fraction of SV40 chromatin between the origin of replication and the region coding the late leader DNA, *Cell,* 20, 65, 1980.

274. **Singh, H., Sen, R., Baltimore, D., and Sharp, P. A.,** A nuclear factor that binds to a conserved sequence motif in transcriptional control elements of immunoglobulin genes, *Nature,* 319, 154, 1986.

275. **Bergman, Y., Rice, D., Grosschedl, R., and Baltimore, D.,** Two regulatory elements for immunoglobulin k light chain gene expression, *Proc. Natl. Acad. Sci. U.S.A.,* 81, 7041, 1984.

276. **Morse, R. H. and Simpson, R. T.,** DNA in the nucleosome, *Cell,* 54, 285, 1988.

277. **Workman, J. L. and Roeder, R. G.,** Binding of transcription factor TFIID to the major late promoter during *in vitro* nucleosome assembly potentiates subsequent initiation by RNA polymerase II, *Cell,* 51, 613, 1987.

278. **Losa, R. and Brown, D. D.,** A bacteriophage RNA polymerase transcribes *in vitro* through a nucleosome core without displacement of histones, *Cell,* 50, 801, 1987.

279. **Richard-Foy, H. and Hager, G. L.,** Sequence-specific positioning of nucleosomes over the steroid-inducible MMTV promoter, *EMBO J.,* 6, 2321, 1987.

280. **Wu, H.-M. and Crothers, D. M.,** The locus of sequence-directed and protein-induced DNA bending, *Nature,* 308, 509, 1984.

281. **Liu-Johnson, H.-N., Gartenberg, M. R., and Crothers, D. M.,** The DNA binding domain and bending angle of *E. coli* CAP protein, *Cell,* 47, 995, 1986.

282. **Weber, I. T. and Steitz, T. A.,** Model of specific complex between catabolite gene activator protein and B-DNA suggested by electrostatic complementarity, *Proc. Natl. Acad. Sci. U.S.A.,* 81, 3973, 1984.

283. **Nussinov, R.,** Nucleotide quartets in the vicinity of eukaryotic transcriptional initiation sites: some DNA and chromatin structural implications, *DNA,* 6, 13, 1987.

284. **Donehower, L. A., Huang, A. L., and Hager, G. L.,** Regulatory and coding potential of the mouse mammary tumor virus long terminal redundancy, *J. Virol.,* 37, 226, 1981.

285. **Barta, A., Richards, R. I., Baxter, J. D., and Shine, J.,** Primary structure and evolution of rat growth hormone, *Proc. Natl. Acad. Sci. U.S.A.,* 78, 4867, 1981.

286. **Cochet, M., Chang, A. C. Y., and Cohen, S. N.,** Characterization of the structural gene and putative 5′-regulatory sequence for human proopiomelanocortin, *Nature,* 297, 335, 1982.

287. **Matsumoto, Y. and Korn, L. J.,** Upstream sequences required for transcription of the TFIIIA gene in *Xenopus* oocytes, *Nucleic Acids Res.,* 16, 3801, 1988.

288. **Jones, K. A. and Tjian, R.,** Essential contact residues within SV40 large T antigen binding sites I and II identified by alkylation interference, *Cell,* 36, 155, 1984.

289. **Struhl, K.,** Regulatory sites for His 3 expression in yeast, *Nature,* 300, 284, 1982.

290. **Zalkin, H. and Yanofsky, C.,** Yeast gene TRP5: structure, function, regulation, *J. Biol. Chem.,* 257, 1491, 1982.

291. **Donahue, T. F., Daves, R. S., Lucchini, G., and Fink, G. R.,** A short nucleotide sequence required for the regulation of His 1 by the general control system of yeast, *Cell,* 32, 89, 1983.

292. **Hinnebusch, A. G., Lucchini, G., and Fink, G. R.,** A synthetic His 4 regulatory element confers general amino acid control on the cytochrome c gene (CYC 1) of yeast, *Proc. Natl. Acad. Sci. U.S.A.,* 82, 498, 1985.

293. **Chodosh, L. A., Buratowski, S., and Sharp, P. A.,** A yeast protein possesses the DNA binding properties of the adenovirus major late transcription factor, *Mol. Cell. Biol.,* submitted.

294. **Weisbord, S.,** Active chromatin, *Nature,* 297, 289, 1982.

295. **Weintraub, H.,** DNA methylation patterns, formation and function, *Cell,* 32, 1191, 1983.

296. **Razin, A. and Szyf, M.,** DNA methylation pattern, formation and function, *Biochim. Biophys. Acta,* 782, 331, 1984.

297. **Wolf, S. F. and Migeon, B. R.,** Clusters of C_pG dinucleotides implicated by nuclease hypersensitivity as control elements of house keeping genes, *Nature,* 314, 467, 1985.

298. **Nussinov, R.,** Promoter helical structure variation at the *E. coli* polymerase interaction sites, *J. Biol. Chem.,* 259, 6798, 1984.

299. **Jernigan, R., Sarai, A., Ting, K. L., and Nussinov, R.,** Hydrophobic interactions in the major groove can influence DNA local structure, *J. Biomol. Struct. Dyn.,* 4, 41, 1986.

300. **Lennon, G. G. and Nussinov, R.,** Homonyms, synonyms and mutations of the sequence-structure vocabulary, *J. Mol. Biol.,* 175, 425, 1984.

301. **Breslauer, K. J., Frank, R., Blocker, H., and Marky, L. A.,** Predicting DNA duplex stability from base sequence, *Proc. Natl. Acad. Sci. U.S.A.,* 83, 3746, 1986.

302. **Margalit, H., Shapiro, B. A., Nussinov, R., Owens, J., and Jernigan, R. L.,** Helix stability in prokaryotic promoter regions, *Bio-*

chemistry, 27, 5179, 1988.

303. **Nussinov, R.**, Putative elements in the vicinity of viral transcription initiation sites, *Int. J. Biochem.*, 20, 721, 1988.
304. **Marini, J. C., Levene, S. D., Crothers, D. M., and Englund, P. T.**, Bent helical structure in kinetoplast DNA, *Proc. Natl. Acad. Sci. U.S.A.*, 79, 7664, 1982.
305. **Nussinov, R., Sarai, A., Wang, D., and Jernigan, R. L.**, Sequence context of homooligomer tracts in eukaryotic genomes: some DNA conformational implications, in *Structure and Expression*, DNA Bending and Curvature, Vol. 3, Olson, W. K., Sarma, M. K., Sarma, R. H., and Sundaralingam, M., Eds., Adenine Press, New York, 1988, 129.
306. **Nussinov, R., Sarai, A., Smythers, G. W., and Jernigan, R. L.**, Sequence context of oligomer tracts in eukaryotic DNA: biological and conformational implications, *J. Biomol. Struct. Dyn.*, 6, 543, 1988.
307. **Nussinov, R., Sarai, A., Smythers, G. W., and Jernigan, R. L.**, Distinct patterns in homooligomer tract sequence context in prokaryotic and eukaryotic DNA, *Biochem. Biophys. Acta*, 1008, 329, 1989.
308. **Nussinov, R., Sarai, A., Smythers, G. W., and Jernigan, R. L.**, Strong patterns in homooligomer tracts occurrences in potential regulatory sites in eukaryotic genomes, *J. Biomolec. Str. Dyn.*, 7, 707, 1989.
309. **Hagerman, P. J.**, Sequence dependence of the curvature of DNA: a test of the phasing hypothesis, *Biochemistry*, 24, 7033, 1985.
310. **Nussinov, R.**, Some rules in the ordering of nucleotides in the DNA, *Nucleic Acids Res.*, 8, 4545, 1980.
311. **Nussinov, R.**, Doublet frequencies in evolutionary distinct groups, *Nucleic Acids Res.*, 12, 1749, 1984.
312. **Broyles, S. S. and Pettijohn, D. E.**, Interaction of the *Escherichia coli* HU protein with DNA: evidence for formation of nucleosome-like structures with altered DNA helical pitch, *J. Mol. Biol.*, 187, 47, 1986.
313. **Drlica, K.**, The nucleoid, in *Escherichia coli* and *Salmonella typhimurium*, cellular and molecular biology, Neidhardt, F. C., Ed., American Society for Microbiology, Washington, D.C., 1987, 91.
314. **Schmid, M. B.**, Structure and function of the bacterial chromosome, *Trends Biochem. Sci.*, 13, 131, 1988.
315. **Zhurkin, V. B., Ulyanov, N. B., and Ivanov, V. I.**, Mechanisms of DNA bending in the free state and in the nucleosome, in *Structure and Expression*, DNA Bending and Curvature, Olson, W. K., Sarma, M. H., Sarma, R. H., and Sundaralingam, M., Eds., Vol. 3, Adenine Press, New York, 1988, 169.
316. **Nussinov, R.**, Presence of large helical structure variation in yeast his4 upstream region is correlated with general amino acid control on CYC1 gene, *J. Biomol. Struct. Dyn.*, 3, 349, 1986.
317. **Hogan, M., LeGrange, J., and Austin, R.**, Dependence of DNA helix flexibility on base composition, *Nature*, 304, 752, 1983.
318. **Liu, L. F. and Wang, J. C.**, Supercoiling of the DNA template during RNA transcription, *Proc. Natl. Acad. Sci. U.S.A.*, 84, 7024, 1987.
319. **Wu, H.-Y., Shy, S., Wang, J. C., and Liu, L. F.**, Transcription generates positively and negatively supercoiled domains in the template, *Cell*, 53, 433, 1988.
320. **Giaever, G. N. and Wang, J. C.**, Supercoiling of intracellular DNA can occur in eukaryotic cells, *Cell*, 55, 849, 1988.
321. **Tsao, Y.-P., Wu, H.-Y., and Liu, L. F.**, Transcription-driven supercoiling of DNA: direct biochemical evidence from *in vitro* studies, *Cell*, 56, 111, 1989.
322. **Babcock, M. and Olson, W.**, personal communication.