Haematology and Blood Transfusion Vol. 29
Modern Trends in Human Leukemia VI
Edited by Neth, Gallo, Greaves, Janka
© Springer-Verlag Berlin Heidelberg 1985

Modulation of Gene Expression Through DNA-Binding Proteins: Is There a Regulatory Code?

M. Beato[1]

Introduction

The information stored in the DNA of a fertilized egg can be divided into two different classes: structural information, required for the synthesis of all macromolecules that build up the organism, and regulatory information, needed to modulate the expression of the structural information in time and space, that means during the development of the different tissues. The connection between the two types of information is provided by regulatory macromolecules, that are of course encoded in the structural information and regulate its expression through interaction with regulatory elements of the DNA, thus closing the information cycle (Fig. 1).

The structural information is stored in the DNA in the form of the genetic code that was unraveled in the 1960s. Part of the structural information are the signals for initiation and termination of transcription and translation, as well as the signals for RNA modification and splicing. On the other hand, little is known about the molecular mechanisms by which regulatory information is stored in the DNA. The general idea, however, is that recognition of specific features of the DNA molecule by regulatory DNA-binding macromolecules is essential for regulation. What exactly is recognized on the DNA and how the interaction modulates gene expression are the questions to be answered.

During the past decade, several DNA-binding regulatory proteins from prokaryotes have been purified to homogeneity, and their structure as well as their interaction with DNA have been studied in great detail. A comparison of the amino acid sequence of 13 DNA-binding regulatory proteins reveals two regions of homology overlapping the known DNA-binding domains (Fig. 2; [1, 2]). Interestingly, mutants that disturb the binding of the *lac*-repressor to the operator are clustered around these two regions [1].

The secondary, tertiary, and quaternary structure of several DNA-binding regulatory proteins from bacteria and bacteriophages exhibit striking similarities in their DNA-binding domains [2]. Not only are these proteins symmetric dimers or tetramers, but they contain a pair of twofold related α-helices connected by a β-turn that are responsible for most of the contacts with the B-form of the DNA double helix.

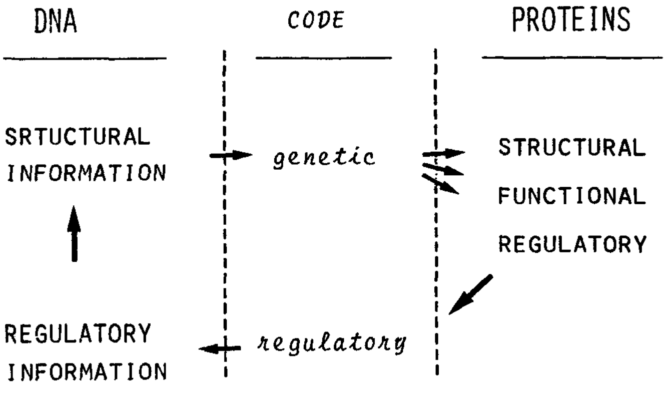


Fig. 1. The information cycle

[1] Physiologisch-Chemisches Institut I der Philipps-Universität, Emil-Mannkopff-Str. 1–2, 3550 Marburg, FRG

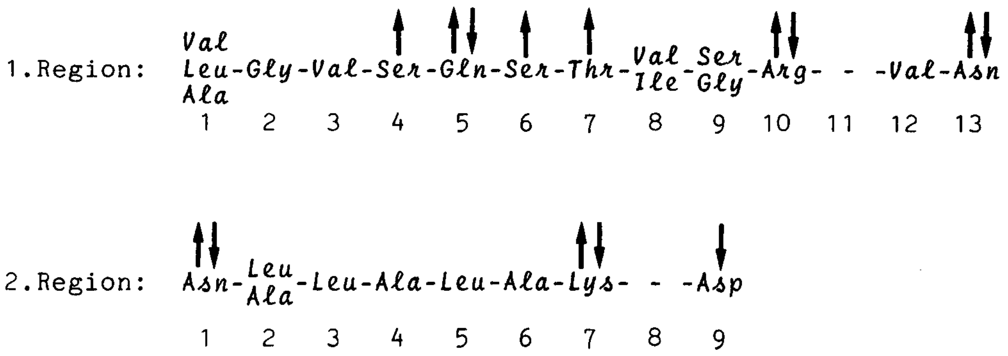

1.Region:
Val
Leu-Gly-Val-Ser-Gln-Ser-Thr-Val Ile-Ser Gly-Arg- - -Val-Asn
Ala
1 2 3 4 5 6 7 8 9 10 11 12 13
2.Region:
Asn-Leu Ala-Leu-Ala-Leu-Ala-Lys- - -Asp
1 2 3 4 5 6 7 8 9

Fig. 2. Regions of homology among 13 prokaryotic DNA-binding regulatory proteins

One α-helix fits into the major groove of DNA while the other lies across it, holding it in position. If one looks at the relevant α-helices along their longitudinal axis, one observes that the orientation of the amino acid side chains exhibits a clear polarity. That means that the nonpolar amino acid side chains are oriented toward one side of the α-helix, whereas the polar and charged amino acid side chains are oriented toward the other side of the helix. This would be the site that contacts the DNA major groove.

This brief summary on the structure of prokaryotic regulatory proteins suggests that a basic protein structure has originated in evolution that can fulfill the requirements for DNA recognition. The actual function of a particular regulatory protein may depend on other domains of the protein that mediate the interaction with different modulator molecules.

As for the DNA sequences that are recognized by the regulatory proteins, they also show considerable homology. Two types of conserved sequences can be derived from a comparison of 23 sites recognized by 13 regulatory proteins [1].

I. TGTGT N_{6-10} ACACA
II. CAC N_{5-10} GTG

Both consensus sequences show a twofold rotational symmetry as expected from DNA sites recognized by dimeric or tetrameric proteins. Both conserved sequences are also similar in that they are composed of two short blocks (3–5 base pairs) of well-conserved nucleotides separated by a more variable region (7–8 base pairs on average). This structure of the binding sites is compatible with a model according to which the regulatory proteins contact the DNA only from one side and interact with two consecutive turns of the double helix (see later discussion). Most of the mutations that prevent binding of a particular regulatory protein to its binding site are located within the strictly conserved regions. It is striking that the homology between different binding sites for the same regulatory protein is not necessarily better than the homology between sites for different proteins, independent of whether they function as positive or negative modulators of transcription. In fact, the cyclic AMP receptor protein (CAP) of *Escherichia coli* can bind not only to its own sites in the regulated promoters, but also to the *lac* and *ara* operators [3, 4]. Thus, it appears that the mechanism by which regulatory proteins recognize their binding sites on DNA is similar regardless of the functional consequences of the interaction.

In higher organisms, several DNA-binding regulatory proteins have been described. The best characterized are probably the T antigens of DNA tumor viruses such as SV40 and polyoma. The behavior of these proteins is reminiscent of that found in the repressor systems of λ bacteriophages. By binding to three adjacent sites on the DNA, they can act as inhibitors of transcription from the early promoter or as activators of the late promoter [5].

I will concentrate on another group of regulatory proteins that have been extensively studied in our and other laboratories during the past 20 years, namely the receptors for steroid hormones. It is now well established that steroid hormones exert their effects on gene expres-

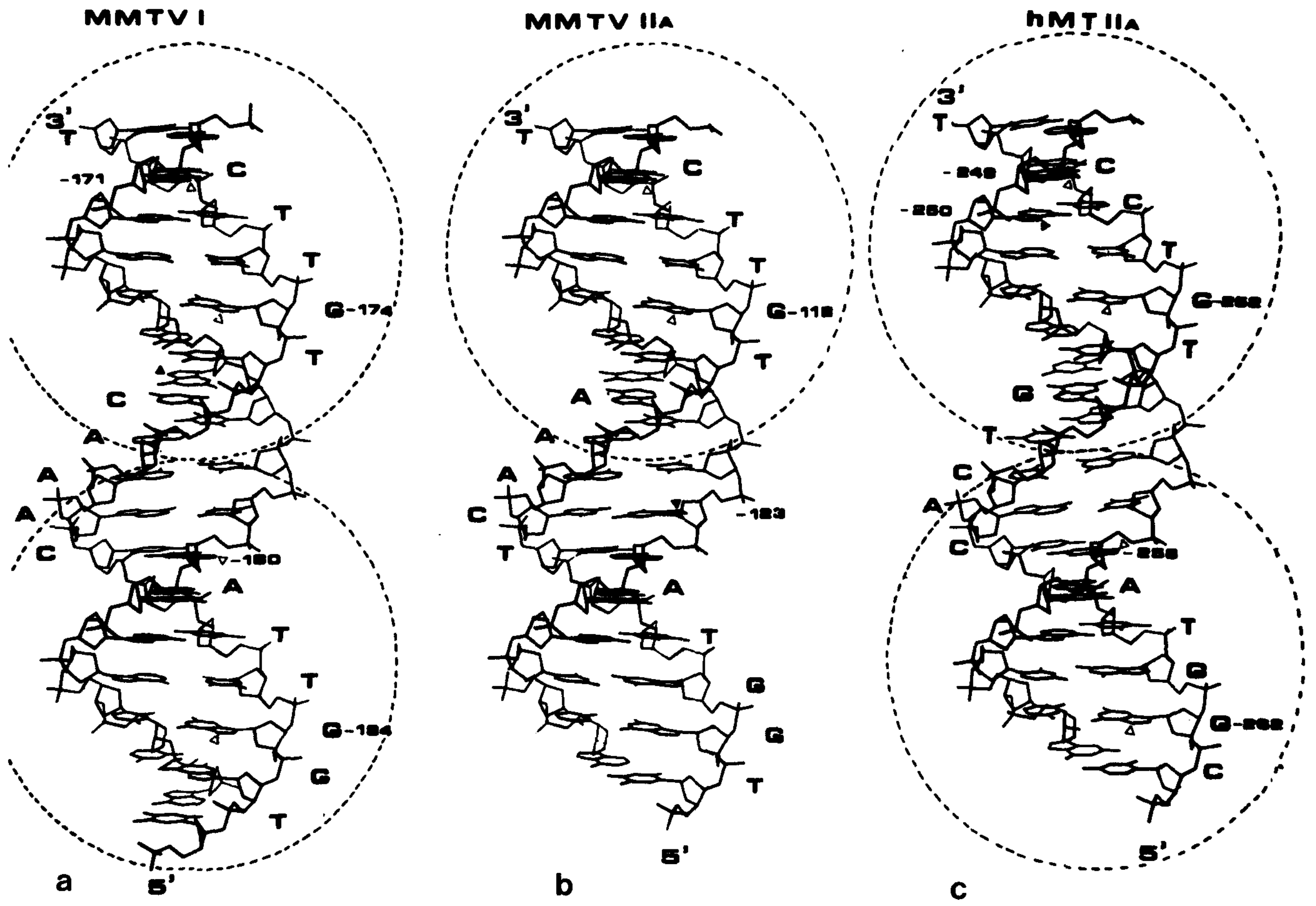


Fig. 3a–c. Structure of the glucocorticoid-binding sites of MMTV and hMTIIA. Computer graphic representation of the DNA double helix containing the nucleotide sequences of **a** MMTVI; **b** MMTVIIA; and **c** hMTIIA (shown in Fig. 5). The sites of contact with the receptor are indicated by *open triangles.* Those positions hypermethylated in the presence of receptor are marked by *full triangles.* The receptor molecules are represented as *broken circles.* Numbers refer to the distance from the "cap" site

sion through interaction with intracellular receptors, that in their turn recognize regulatory elements in the neighborhood of the regulated promoters. Regulatory elements are defined as DNA sequences that in addition to being required for receptor binding, are needed for the hormonal regulation of transcription in gene transfer experiments. They were first reported in the long terminal repeat region (LTR) of mouse mammary tumor virus (MMTV), that contains the main promoter for proviral transcription [6–8]. Glucocorticoids were known to induce viral transcription in different cell lines [9], and gene transfer experiments with deletion mutants in the LTR region showed that the sequences relevant for hormonal regulation are located between 50 and 400 base pairs upstream of the initation of transcription [7, 10–12]. Within this region, several binding sites for the glucocorticoid receptor of rat liver have been described [8, 13]. Using a cloned proviral DNA from GR mice [8], we found four binding sites that share the hexanucleotide

5′-TGTTCT-3′
3′-ACAAGA-5′

Methylation protection studies have shown that both G residues in the hexanucleotides are in direct contact with the receptor [14]. In the binding site with the highest affinity for the receptor, further contacts are located in both strands 9–10 base pairs upstream of the hexanucleotide. These findings suggest an interaction of a dimer of the receptor with one side of the double-stranded DNA involving the major groove in two subsequent turns of the helix [14]. Such a model (Fig. 3a) is very similar to

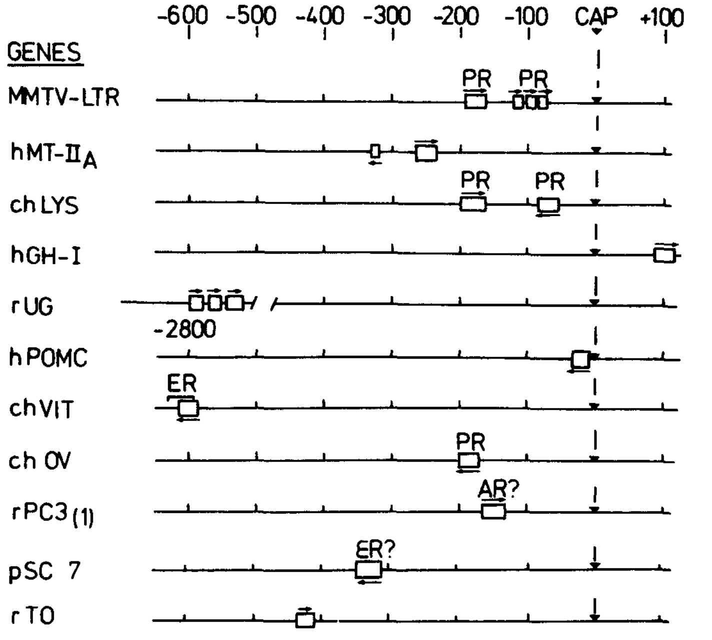


Fig. 4. Position and orientation of the binding sites for the glucocorticoid receptor in hormonally regulated genes. The binding sites are indicated by *open boxes*. The *horizontal arrows* show the orientation: to the right for the upper strand; to the left for the lower strand. The abbreviations are as follows: MMTV-LTD long terminal repeat region of mouse mammary tumor virus; hMT-II$_A$, human metallothioneine II$_A$; chLYS chicken lysozyme; hGH-I human growth hormone; rUG rabbit uteroglobin; hPOMC human proopiomelanocortin; chVIT chicken vitellogenin; chOV chicken ovalbumin; rPC3 [1] rat prostatic protein C3 [3]; pSC 7 *Drosophila* inducible gene at locus 74F; rTO rat tryptophanoxygenase; CAP initiation of transcription; PR progesterone receptor; ER estrogen receptor; AR androgen receptor; ER ecdysone receptor

that already mentioned for prokaryotic DNA-binding regulatory proteins.

An analysis of other glucocorticoid-regulated genes showed that the presence of a regulatory element is not an exclusive property of the retroviral genome. The human metallothioneine II$_A$ gene (hMTII$_A$), that has been shown to be induced by glucocorticoids in many different cell lines, contains a glucocorticoid regulatory element about 250 base pairs upstream of the initiation of transcription [15]. This element is very similar to the strong binding site found in the LTR region of MMTV (compare a and c in Fig. 3). In addition, there is a weak binding site in the hMTII$_A$ promoter located at around 320 base pairs upstream of the initiation of transcription [15]. Similarly to the weak binding site in the LTR region of MMTV (Fig. 3b), the shorter footprint and methylation protection pattern in the weak binding site of hMTII$_A$ suggests binding of a receptor monomer. Interestingly, this weak site at – 320 can be deleted without influencing the hormonal inducibility of hMTII$_A$ [15]. Thus, it could be that a functional interaction requires binding of a receptor dimer to a strong site on the DNA. In the meantime, we have identified binding sites for the glucocorticoid receptor in several hormonally regulated genes. A summary of these results along with data from the literature is shown in Fig. 4.

The promoter for the chicken lysozyme gene (chLYS), contains two binding sites for the glucocorticoid receptor, located at around 180 and 60 base pairs upstream of the initiation of transcription [16]. The upper binding site, that has a lower affinity for the glucocorticoid receptor, coincides with sequences required for hormone-dependent expression of the gene in oviduct cells [16]. In fact, these sequences mediate not only glucocorticoid regulation, but also induction by progesterone in microinjection experiments [16]. Interestingly, the partially purified progesterone receptor from rabbit uterus binds to the same sites as the glucocorticoid receptor, although with different affinity. Thus, it appears that the binding sites for the receptors of two different steroid hormones may be identical or at least share common sequences. That these similarities may not be limited to the progesterone and glucocorticoid receptors is suggested by studies with genes regulated by other steroid hormones (Fig. 4). The chicken vitellogenin II gene that is induced by estrogens in the liver, contains a binding site for the estrogen receptor around 600 nucleotides upstream of the transcription initiation site [17]. An analysis of the nucleotide sequences in this region reveals an element almost identical to the binding sites for the glucocorticoid receptor (Fig. 5).

A review of the literature showed that a rat gene for a prostatic protein, that is

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
MMTV I	−186	T	G	G	T	T	A	C	A	A	A	C	T	G	T	T	C	T
MMTV IIa	−129	T	G	G	T	A	T	C	A	A	A	.	T	G	T	T	C	T
MrDNA	+2	C	T	G	A	C	A	C	A	C	G	C	T	G	T	C	C	T
rPC3(1)	−150	A	T	G	A	A	A	C	.	C	A	G	T	G	T	T	C	T
hMTIIA	−263	C	G	G	T	.	A	C	A	C	T	G	T	G	T	C	C	T
hGH-1	+92	G	G	G	C	.	A	C	A	A	T	G	T	G	T	C	C	T
hPOMC	−6	C	G	C	G	C	T	C	.	C	T	C	T	G	T	C	C	T
chLYS1	−50	T	T	G	A	T	T	C	.	C	T	C	T	G	T	T	C	T
chLYS2	−191	A	A	A	A	T	T	C	.	C	T	C	T	G	T	G	G	C
chVIT	−587	C	G	G	C	A	T	C	A	A	T	G	T	G	T	T	C	T
chOV	−177	G	G	G	C	.	A	C	A	A	T	G	T	G	T	C	C	T
RUG1	−2800	C	T	G	T	.	T	C	A	C	T	C	T	G	T	T	C	T
RUG2	−2820	C	C	G	G	.	A	C	A	C	G	G	A	G	T	C	C	T
RUG3	−2840	G	T	G	T	.	C	A	G	T	C	T	T	G	T	T	C	T
dSC7(2)	−315	T	C	G	A	T	T	T	G	A	T	C	T	G	T	T	C	T
rTO	−435	A	T	G	C	.	A	C	A	G	C	G	A	G	T	T	C	T
CONSENSUS.		C	G	G	T	A	A	C	A	C	T	G	T	G	T	T	C	T
		t	t		A	T	T			A	a	c				c		
YEAST Aktin		A	A	G	A	.	A	C	A	C	C	C	T	G	T	T	C	T

Fig. 5. Consensus sequence for the glucocorticoid regulatory element. The nucleotide sequences of the main binding sites for the glucocorticoid receptor are aligned to yield maximal homology. Abbreviations are as in Fig. 4

known to be induced by androgens, rPC 3(1), also contains a sequence homologous to the binding site for the glucocorticoid receptor some 140 nucleotides upstream of the initiation of transcription ([18]; Figs. 4 and 5). Finally, an ecdysone-inducible gene of *Drosophila* (pSC7) also contains a binding site for the glucocorticoid receptor some 330 nucleotides upstream of the transcription initiation site ([19]; Figs. 4 and 5). These findings, taken together, suggest that the regulatory elements for different steroid hormone receptors may be similar or at least overlap.

The rabbit uteroglobin gene is induced by glucocorticoids in the lung and by estrogen and progesterone in the endometrium [20]. We have looked for binding sites for the glucocorticoid receptor and found none in the neighborhood of the promoter. The closest binding region detected is located 2700 nucleotides upstream of the cap site, and is composed of three binding sites showing sequence homology to other glucocorticoid regulatory elements (Figs. 4 and 5). That this site may be relevant for regulation in vivo is suggested by the finding of a DNase I hypersensitive site in this region only in chromatin of hormonally stimulated endometrium (unpublished results).

The human growth hormone gene (hGHI) is induced by glucocorticoids in several cell lines [21]. In fact, gene transfer experiments with a chimeric gene suggested that a fragment of DNA containing 500 base pairs upstream of the initiation of transcription is sufficient for hormonal regulation [22]. In binding experiments with the glucocorticoid receptor, however, we found a main binding site located around position + 100, within the first intron (Figs. 4 and 5). If this site is involved in transcriptional regulation in vivo, it would mean that the regulatory element can act even when located downstream of the regulated promoter.

Taken together, the data shown in Fig. 4 show that the regulatory elements for steroid hormones share some of the properties of the so-called enhancer elements [23]. They can act at variable distance from the regulated promoters, both upstream and downstream, and in both orientations. There is in fact direct experimental evidence for an enhancer function of the glucocorticoid regulatory element in the LTR region of MMTV [7].

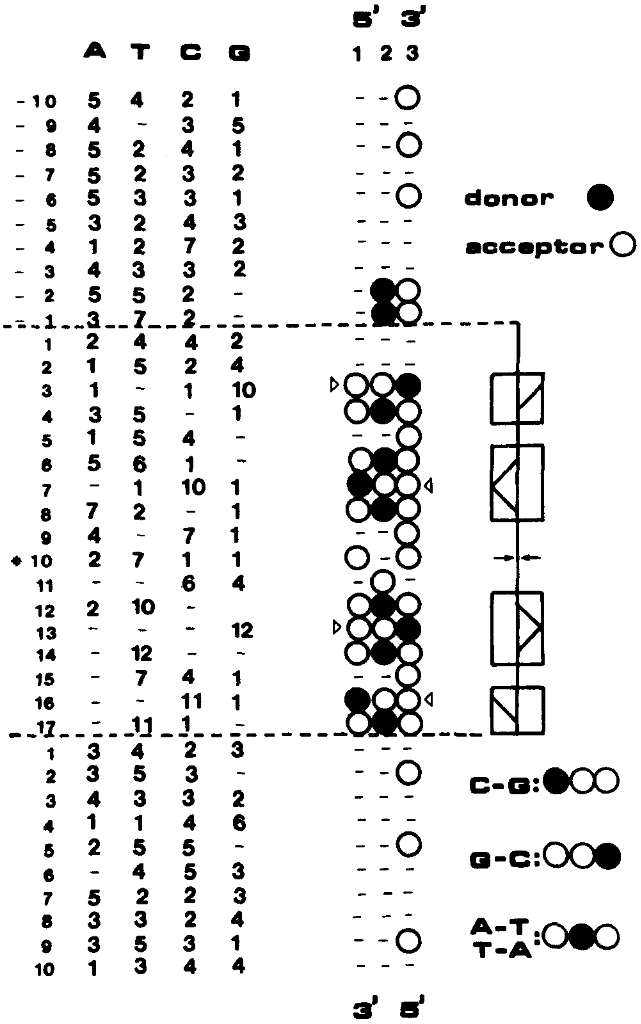
A comparison of the nucleotide sequences of ten different binding sites for the glucocorticoid receptor yields the consensus sequence shown in Fig. 5. Therefore, the glucocorticoid regulatory elements have been conserved in evolution between chicken, rodents, and humans. The best-conserved regions include all those sites that are involved in direct contacts with the receptor [14]. The symmetry in the element

$$\begin{array}{cccccccccc} 4 & 5 & 6 & 7 & & 12 & 13 & 14 & 15 & \\ 5' \; A & A & C & A & N_{8-11} & T & G & T & T & 3' \end{array}$$

is reminiscent of the binding sites for prokaryotic regulatory proteins, suggesting that molecular mechanisms similar to those operating in bacteria may be responsible for DNA recognition in higher organisms.

What could this mechanism be? And, how can a regulatory protein accommodate so much sequence variation in the central part of the recognition site? Of course, a model like the one shown in Fig. 3 will only require the binding sequence to be preserved in the two nucleotide blocks that are the sites of contact between the relevant α-helices and the major groove of double helix. This would explain the tolerance in the central part of the element, but what kind of interactions take place in the conserved regions? Certainly most of the overall energy of binding is sequence independence, and originates from ionic interactions with the phosphate backbone of the helix [2, 24]. This explains why all DNA-binding regulatory proteins also interact nonspecifically with DNA. In addition, specific base recognition is based on a complementary network of hydrogen bonds between amino acid side chains in the relevant α-helices and DNA base pair atoms exposed in the major groove of the double helix [24]. In fact several amino acid side chains such as Arg, Lys, Gln, and Asn, can form multiple hydrogen bonds with paired bases on the DNA [25]. It has been proposed that if the regulatory protein moves a few Ångstroms away from the DNA, most of these hydrogen bonds would be broken or would not be formed, but many of the ionic interactions will be preserved. This mechanism may be utilized by the proteins for sliding along the DNA in search of their target sites [26].

If we consider the base pairs in the major groove in terms of their ability to form hydrogen bonds, we realize that an AT



	A	T	C	G
-10	5	4	2	1
-9	4	-	3	5
-8	5	2	4	1
-7	5	2	3	2
-6	5	3	3	1
-5	3	2	4	3
-4	1	2	7	2
-3	4	3	3	2
-2	5	5	2	-
-1	3	7	2	-
1	2	4	4	2
2	1	5	2	4
3	1	-	1	10
4	3	5	-	1
5	1	5	4	-
6	5	6	1	-
7	-	1	10	1
8	7	2	-	1
9	4	-	7	1
◆ 10	2	7	1	1
11	-	-	6	4
12	2	10	-	
13	-	-	-	12
14	-	12	-	-
15	-	7	4	1
16	-	-	11	1
17	-	11	1	-
1	3	4	2	3
2	3	5	3	-
3	4	3	3	2
4	1	1	4	6
5	2	5	5	-
6	-	4	5	3
7	5	2	2	3
8	3	3	2	4
9	3	5	3	1
10	1	3	4	4

Fig. 6. Pattern of hydrogen bond donor–acceptor sites in the major groove of the DNA double helix in and around the receptor-binding sites. The conserved nucleotide sequences of twelve binding sites for the glucocorticoid receptor with the flanking base pairs on each side, have been analyzed for the pattern of hydrogen bond donor–acceptor sites in the major groove. Only those positions showing more than 90% conservation are shown (*open circles* acceptor sites; *full circles* donor sites). *Arrows* point to the conserved N-7 positions of guanines that represent sites of contact with the receptor [14]

base pair has the structure acceptor–donor –acceptor and is therefore symmetric, whereas a GC base pair has the structure acceptor–acceptor–donor (Fig. 6). If one now compares the ten glucocorticoid receptor binding sites with their flanking sequences in terms of this hydrogen bond pattern, one observes a very good preservation of the donor–acceptor structure around the contact sites, with very little agreement outside the binding region (Fig. 6). A certain symmetry can be detected centered at position 10: two well-preserved blocks, 3 to 8 and 12 to 17, separated by less-preserved positions, and interrupted in symmetric positions at 5 and 15. Of course, other interactions are probably implicated in recognition, but the network of hydrogen bonds seems to be an essential part of the code in which regulatory information is stored in DNA. A precise understanding of the molecular mechanisms by which the regulatory code is read could derive from the fine structural analysis of cocrystals containing the DNA-binding domains of regulatory proteins bound to the corresponding nucleotide sequences [27, 28]. Only then will it be possible to decide whether there is a general rule underlying the mechanism of sequence-specific recognition by regulatory proteins.

Acknowledgments. The experimental work reviewed here has been supported by grants from the Deutsche Forschungsgemeinschaft and the Fond der Chemischen Industrie.

References

1. Gicquel-Sanzey B, Cossart P (1982) EMBO J 1:591–595
2. Takeda Y, Ohlendorf DH, Anderson WF, Matthews BW (1983) Science 221:1020–1026
3. Ogden S, Haggerty D, Stoner CM, Kolodrubetz D, Schleif R (1980) Proc Natl Acad Sci USA 77:3346–3350
4. Schmitz A (1981) Nucleic Acid Res 9:277 –291
5. Tjian R (1978) Cell 13:165–179
6. Geisse S, Scheidereit C, Westphal HM, Hynes NE, Groner B, Beato M (1982) EMBO J 1:1613–1619
7. Chandler VL, Maler BA, Yamamoto KR (1983) Cell 33:489–499
8. Scheidereit C, Geisse S, Westphal HM, Beato M (1983) Nature 304:749–752
9. Ringold GM (1979) Biochim Biophys Acta 560:487–508
10. Hynes NH, van Ooyen AJJ, Kennedy N, Herrlich P, Ponta H, Groner B (1983) Proc Natl Acad Sci USA 80:3637–3641
11. Majors J, Varmus HE (1983) Proc Natl Acad Sci USA 80:5866–5870
12. Buetti E, Diggelmann H (1983) EMBO J 2:1423–1429
13. Payvar F, deFranco DF, Firestone GL, Edgar B, Wrange Ö, Okret S, Gustafsson JA, Yamamoto KR (1983) Cell 35:381–392
14. Scheidereit C, Beato M (1984) Proc Natl Acad Sci USA 81:3029–3033
15. Karin M, Haslinger A, Holtgreve H, Richards RI, Krauter P, Westphal HM, Beato M (1984) Nature 308:513–519
16. Renkawitz R, Schütz G, von der Ahe D, Beato M (1984) Cell 37:503–510
17. Jost JP, Seldran M, Geiser M (1984) Proc Natl Acad Sci USA 8:429–433
18. Parker M, Hurst H, Page M (1984) J Steroid Biochem 20:67–71
19. Moritz T, Edström JE, Pongs O (1984) EMBO J 3:289–295
20. Beato M, Arnemann J, Menne C, Müller H, Suske G, Wenz M (1983) In: McKerns KW (ed) Regulation of gene expression by hormones. Plenum, New York, pp 151–175
21. Martial JA, Baxter JD, Goodman HM, Seeburg PH (1977) Proc Natl Acad Sci USA 74: 1816–1820
22. Robins DM, Paek I, Seeburg P, Axel R (1982) Cell 29:623–631
23. Banerji J, Rusconi S, Schaffner W (1981) Cell 27:299–308
24. Seeman NC, Rosenburg JM, Rich A (1979) Proc Natl Acad Sci USA 72:804–808
25. Rein R, Kieber-Emmons T, Haydock K, Garduno-Juarez R, Shibata M (1983) J Biomol Struct Dyn 1:1051–1079
26. Berg OG, Winter RB, von Hippel PH (1981) Biochemistry 20:6929–6948
27. Anderson J, Ptashne M, Harrison SC (1984) Proc Natl Acad Sci USA 81:1307–1311
28. Frederick CA, Grable J, Melia M, Samudzi C, Jen-Jacobson L, Wang BC, Greene P, Boyer HW, Rosenberg JM (1984) Nature 309:327–331