Nucleic Acids Research, 1993, Vol. 21, No. 21 ***4915–4922***

Isolation and characterization of target sequences of the chicken *CdxA* homeobox gene

Yael Margalit, Sinai Yarus, Eli Shapira, Yosef Gruenbaum[1] and Abraham Fainsod*
Department of Cellular Biochemistry, Hebrew University–Hadassah Medical School, Jerusalem 91010 and [1]Department of Genetics, Hebrew University, Jerusalem 91904, Israel

Received July 27, 1993; Revised and Accepted September 23, 1993 EMBL accession nos+

ABSTRACT

The DNA binding specificity of the chicken homeodomain protein CDXA was studied. Using a CDXA-glutathione-S-tranferase fusion protein, DNA fragments containing the binding site for this protein were isolated. The sources of DNA were oligonucleotides with random sequence and chicken genomic DNA. The DNA fragments isolated were sequenced and tested in DNA binding assays. Sequencing revealed that most DNA fragments are AT rich which is a common feature of homeodomain binding sites. By electrophoretic mobility shift assays it was shown that the different target sequences isolated bind to the CDXA protein with different affinities. The specific sequences bound by the CDXA protein in the genomic fragments isolated, were determined by DNase I footprinting. From the footprinted sequences, the CDXA consensus binding site was determined. The CDXA protein binds the consensus sequence A, A/T, T, A/T, A, T, A/G. The CAUDAL binding site in the *ftz* promoter is also included in this consensus sequence. When tested, some of the genomic target sequences were capable of enhancing the transcriptional activity of reporter plasmids when introduced into CDXA expressing cells. This study determined the DNA sequence specificity of the CDXA protein and it also shows that this protein can further activate transcription in cells in culture.

INTRODUCTION

Homeobox genes are a family of genes initially identified in the *Drosophila* genome as regulators of developmental decisions during embryogenesis (1). Subsequently, homeobox genes were identified and isolated from numerous organisms including higher vertebrates such as chickens, mice and man (2). A common feature of these genes is a 183 bp long sequence conserved during evolution, the homeobox (3, 4). It has been shown that the homeobox is part of a larger open reading frame which in turn gives rise to a 61 amino acid domain of a larger protein, a homeodomain protein. Early computer analysis of the homeodomain sequence revealed its potential ability to form a helix-turn-helix motif. Helix-turn-helix motifs had been shown to play a central role in the DNA binding activity of a number of bacterial and yeast proteins (5, 6). This observation suggested that the homeodomain is the DNA binding domain of these proteins, and that these proteins exert their regulatory effect through direct interaction with their target genes. The structure of a number of homeodomains has been studied in detail and they were found to have some differences when compared to the prokaryotic regulatory proteins.

The DNA binding properties of homeodomain proteins have been studied in *in vitro* and *in vivo* systems. One of the key components in these studies has been the identification of DNA fragments to which homeodomain proteins bind. In *Drosophila*, target genes have been identified by a combination of genetic and expression studies. Once a relationship between a homeodomain protein and a gene was established, the DNA element to which it binds can be isolated and studied. In vertebrates, the lack of mutants precludes such an approach. On the other hand, the studies in *Drosophila* identified autoregulatory interactions in homeobox genes and regulatory interactions between various homeobox genes (7). Assuming evolutionary conservation of regulatory networks, similar kinds of interactions between homeobox genes were successfully sought after in vertebrates (8, 9). It has been shown that homeodomain proteins can bind DNA fragments in their own promoters as well as in promoters of other homeobox genes.

For the identification of target sequences and eventually target genes, a number of alternative approaches have been developed (10–16). These target sequences can be either of genomic origin (10, 14) or oligonucleotides of random sequence (11–13, 15). In most cases the binding reaction between a bacterially-produced DNA binding protein and its target DNA is carried out *in vitro* (10–15). The protein–DNA complexes are then isolated via one of various different techniques: they are either immunoprecipitated (10), precipitated due to insolubility (17), precipitated by binding an affinity matrix (14), or isolated from gels after electrophoretic mobility shift assays (EMSA; 13). Usually, multiple rounds of DNA-binding-selection are needed to enrich

* To whom correspondence should be addressed

+ X73511–X73517 (incl.) and X73678

the DNA fragment population for a yield of sequences that bind with relative high affinity the protein of interest.

In *Drosophila*, in addition to the homeobox genes identified from genetic analysis, a number of other genes of this type were isolated based on homeobox sequence cross-hybridization. One of the fly genes first cloned this way was the *caudal* (*cad*) gene (18). This gene was found to be expressed in the posterior regions of the fly embryo (18). Subsequently, mutations in this gene were identified and shown to alter posterior segment specification (19). The same *cad* mutations also affect the pattern of expression of the *fushi tarazu* (*ftz*) gene (19). Analysis of the *ftz* zebra-stripe element identified a region which directs posterior *ftz* expression in a pattern similar to the CAD protein during early embryogenesis (20). Further analysis showed that the CAD protein has in this region, two CDRE's (CAD-protein DNA recognition elements; 20). The CDRE's are composed of two copies of the consensus sequence TTTATG, the one as an inverted repeat separated by 4 bp and the second one as a direct repeat separated by 2 bp (20).

Numerous members of the *caudal* family of genes have been isolated and described. At present there are 9 known *caudal*-type genes from vertebrates; *Cdx1* and *Cdx2* from mice (21, 22), *Xcad1* and *Xcad2* from *Xenopus* (23), *CdxA* and *CHox-cad2* from chicken (24, 25), *Cdx3* from Syrian hamster (26), *Cdx* from rat (27) and *cdx*[Zf-cad1] from zebrafish (28). In the cases where it has been studied it has been shown that these genes overlap to a large extent in their temporal and spatial patterns of expression. The vertebrate pattern of expression is to some extent similar to that of the fly. Only in one case of the vertebrate genes of this family has the DNA-binding property of the protein product been studied. The *Cdx3* gene was isolated due to its binding to the FLAT sequence element of the rat insulin I gene (26). Binding of the CDX3 protein to the FLAT sequence in a reporter gene assay resulted in transcriptional enhancement (26).

In the present work we study the DNA binding properties of the chicken homeodomain containing CDXA protein. The DNA binding sequence of CDXA was initially studied using random sequence oligonucleotides. This study was repeated and expanded with chicken genomic DNA fragments which were studied by competition and DNase I footprinting assays. The consensus binding site for the CDXA protein was identified. In addition some of the genomic fragments are shown to mediate a transcriptional enhancing activity dependent on the CDXA protein.

MATERIALS AND METHODS

Preparation of the CDXA protein

The CDXA protein was prepared as a fusion protein with glutathione-S-transferase (GST). A 1545 bp long Afl III fragment from the *CdxA* cDNA was subcloned into the Sma I site of the pGEX-2T vector (29). Expression of the CDXA-GST fusion protein in *E.coli* was performed according to Smith and Johnson (29). Three different modes of preparation of the fusion protein were utilized. Most EMSA assays were performed with crude bacterial protein extracts only centrifuged after sonication. Isolation of target sequences was performed with fusion protein bound to glutathione-agarose beads resulting from the affinity purification (14, 29). For the DNase I footprinting analysis the CDXA-GST fusion protein was partially purified from the bacterial proteins by ammonium sulfate precipitation according to Pognonec et al. (30).

Isolation of target sequences

Target sequences were isolated from an oligonucleotide mixture with random sequence and from chicken genomic DNA according to the procedure of Kinzler and Vogelstein (10, 17) as modified by Fainsod et al. (14). Briefly, chicken genomic DNA was reduced in size to fragments of about 100−300 bp in length either by sonication or by restriction enzyme digest. To the genomic DNA, linkers were ligated. Either synthetic oligonucleotides with an internal span of 20 random nucleotides or the linkered genomic DNA were mixed with 500 ng of CDXA-GST fusion protein bound to glutathione-agarose beads under DNA-binding conditions. The protein−DNA complexes were spun down, the DNA was released and amplified by PCR. The PCR products were utilized for a second round of DNA-binding purification. After four cycles of DNA-binding enrichment the PCR products were cloned into plasmids.

DNA binding assays

Electrophoretic mobility shift assays (EMSA) were performed according to the procedure of Fried and Crothers (31). The DNA binding reactions contained 0−15 μg of crude bacterial protein extract (about 0−3 μg of CDXA-GST protein), and 10^4 cpm of end labelled DNA fragment. All reactions were carried out in the presence of 2 μg of [poly (dI-dC): poly (dI-dC)] to reduce non-specific binding. DNase I footprinting was performed according to the procedure of Lichsteiner et al. (32). For footprinting assays, 5×10^4 cpm of end labelled fragment were mixed with 0−12 μl (0−2 μg) of CDXA fusion proteinpartially purified by ammonium sulfate precipitation.

Cell transfection and CAT assay

LTK^- cells were transfected with N-[1[(2,3-Dioleoyloxy) propyl]-N,N,N-trimethylammonium methylsulfate (DOTAP; Boehringer Mannheim) according to the recommendations of the manufacturer. Stable transfectants were obtained by co-transfection with a Neomycin resistance plasmid and selected for with Genticin-418. Chloramphenicol acetyltransferase (CAT) assays were performed according to Gorman *et al.* (33).

RESULTS

The main objective of the present work is to study the DNA-binding properties of the CDXA protein, its binding site and the initial characterization of its transcriptional effect. For this purpose, the CDXA protein was expressed in *E.coli* as a fusion protein with glutathione-S-transferase (GST). The CDXA-GST fusion protein contains the CDXA open reading frame from the second methionine (24, 34) and was subcloned in the pGEX-2T vector (29). This fusion protein was used to isolate CDXA target sequences by the procedure of Kinzler and Vogelstein (10). In the original protocol, after an *in vitro* binding reaction, the protein−DNA complex was immunoprecipitated (10). In a modification of this approach, the same authors showed that a precipitated protein due to bacterial overexpression could also be used (17). During the initial stages of this DNA-binding study a good antibody against the CDXA protein was unavailable and therefore an alternative approach for the precipitation of protein−DNA complexes was devised. Taking advantage of the GST component of the CDXA-GST fusion protein, the protein−DNA complexes were precipitated with glutathione-agarose beads (14). The DNA used contained known sequences

on both ends to permit amplification by the polymerase chain reaction (PCR) after the precipitation of protein–DNA complexes. The DNA amplified by PCR served for the next enrichment cycle (10).

Random sequence oligonucleotides

As mentioned earlier, the known binding sites of the *Drosophila* CAD protein include either direct or inverted repeats of the sequence TTTATG separated by 2 and 4 bp resulting in a 14–16 bp long site (20). Assuming the existence of a similar binding site for the CDXA protein, an oligonucleotide was designed with 20 bp of random sequence flanked on both sides by unique sequences for which PCR primers were prepared. The oligonucleotides were subjected to three successive cycles of the isolation procedure to enrich for CDXA target sequences. After cloning, the different oligonucleotides termed CTO (CDXA target oligonucleotides), were tested for binding to CDXA and sequenced. The sequence of 12 such CTO clones is shown in Fig. 1. Seven out of the 12 clones contained two oligonucleotides joined together. The sequences were searched for the CAD consensus binding site (Fig. 1; shaded region). All 12 clones contained a CAD-type binding site and in the cases where two oligonucleotides were joined together, both contained CAD-type binding sites (Fig. 1). In order to prevent biases in the analysis of the oligonucleotide sequences, the nucleotide sequence of the target clones was analyzed with the consensus program (35). This program compares a given set of sequences to determine a consensus sequence of a specified length. The search for a consensus sequence of 7 bp in length among the oligonucleotides isolated gave as a result a sequence almost identical to the CAD binding site (Table 1). The oligonucleotide clones were also analyzed by EMSA and they all exhibited binding to the CDXA protein to different extents (data not shown).

Isolation of CDXA genomic DNA target sequences

The possibility arises that the oligonucleotides as designed could not provide the optimal CDXA binding site due to length or sequence restrictions. This observation was supported by the fact that all the oligonucleotide target sequences contain only one CAD-type binding site and the solution for two such sites appears to be the joining of two such oligonucleotides. To overcome this limitation and to test whether indeed CDXA recognizes a binding site composed of one CAD-type consensus sequence, chicken genomic DNA was utilized as the source of target sequences. The genomic DNA was subjected to four cycles of target sequence enrichment. After enrichment, a library of putative target sequences was prepared. In order to isolate and study the most abundant sequences after four enrichment cycles, the putative target sequence library was screened with the PCR products of the fourth cycle. Seven clones were isolated and termed CTS (CDXA target sequences). The CTS clones were sequenced and characterized as to their binding by CDXA.

CDXA binding to the CTS clones

In order to study the binding of the CDXA protein to the different CTS clones, each clone was studied by EMSA, competition assays and super-shift assays with the addition of CDXA specific antibodies. Binding of the CDXA protein to the different CTS

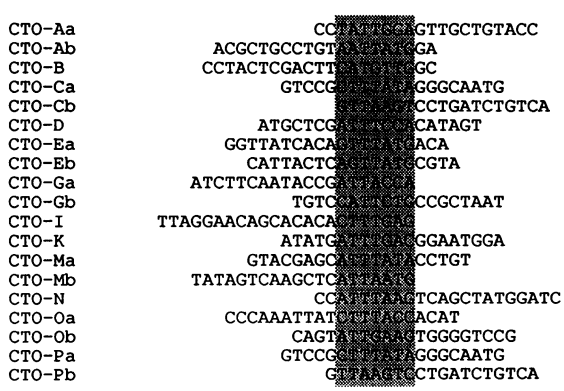

Figure 1. Sequence comparisons of the CTO clones. Specific protein–DNA complexes were precipitated after incubation of the CDXA-GST fusion protein with a population of random sequence oligonucleotides. When a clone contained two oligonucleotides joined together they were termed a and b. The sequences are aligned according to their putative CDXA binding sites. The putative CDXA binding sites are shaded.

Table 1. *CdxA* binding site consensus matrix for CTO clones

nucleotide	position number						
	1	2	3	4	5	6	7
A	9	4	1	3	11	4	6
C	6	0	0	0	3	3	3
G	2	1	0	2	4	2	9
T	2	14	18	13	1	10	1
CDXA consensus	A/C	T	T	T	A	T	A/G
CAD binding site		T	T	T	A	T	G
CDX3 / Flat-E	A	A	T	T	A	G	A
CDX3 / Flat-F	A	T	T	A	A	C	A

Analysis of the sequence of the CTO clones shown in Fig. 1 by the consensus program (35). Computer generated matrix of the representation of each base in each position of the consensus. Comparison to the CAD binding site and the CDX3 binding sites is also shown.

Table 2. *CdxA* binding site consensus matrix for CTO and CTS clones

nucleotide	position number						
	1	2	3	4	5	6	7
A	20	16	9	17	22	2	17
C	3	0	0	0	14	7	0
G	3	2	0	2	0	3	16
T	10	18	27	17	0	24	3
CDXA consensus	A	A/T	T	A/T	A	T	A/G
CAD binding site		T	T	T	A	T	G
CDX3 / Flat-E	A	A	T	T	A	G	A
CDX3 / Flat-F	A	T	T	A	A	C	A

Computer generated matrix of the analysis of all the CTS clones and the CTO clones which bind strongest to the CDXA protein. The number of times each base is found in each position is shown. Comparison to the binding sites of other *caudal*-type proteins is shown.

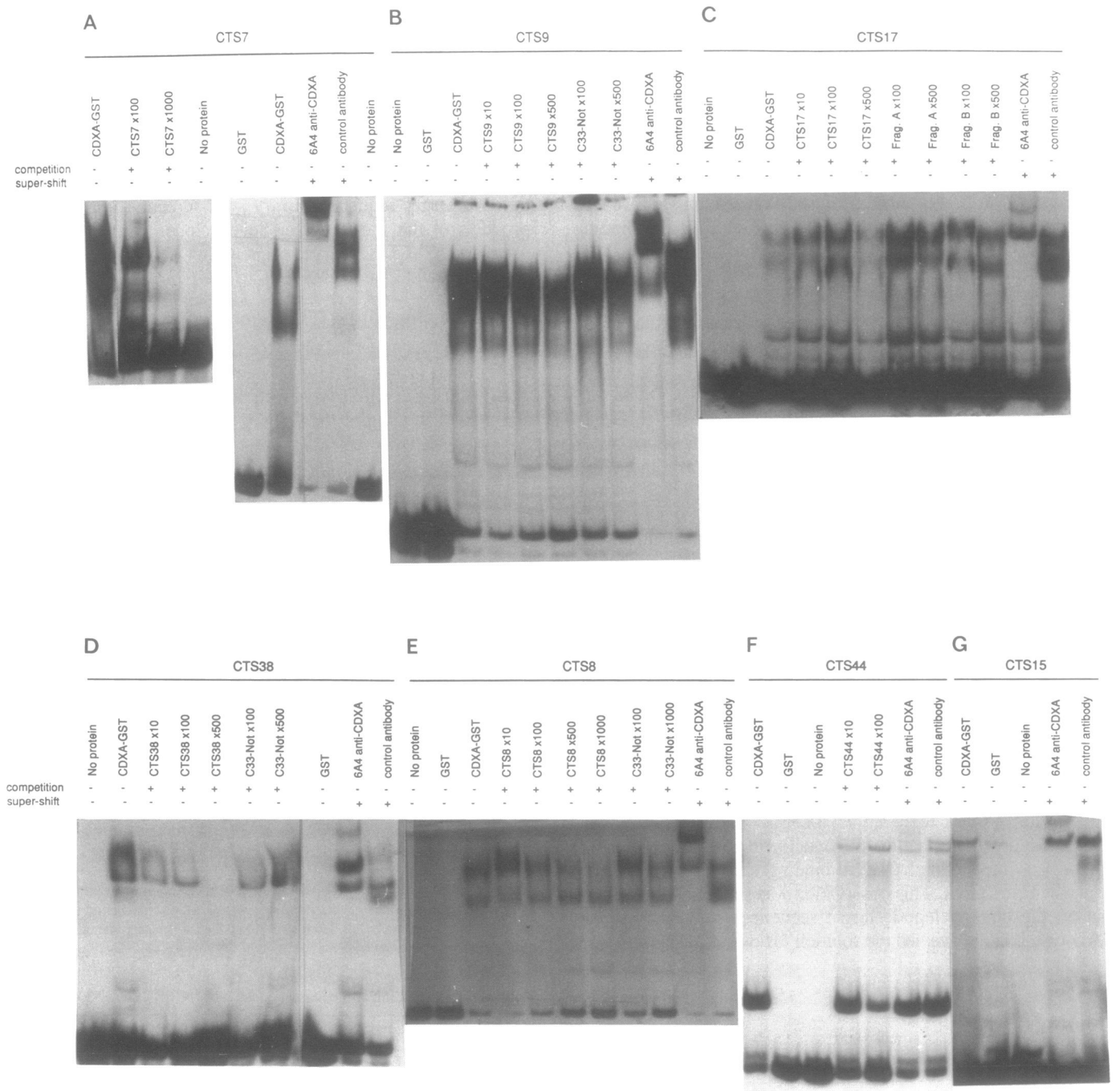

Figure 2. Electrophoretic mobility shift assays (EMSA) of the different CTS clones. Genomic fragments isolated as target sequences for the CDXA protein were subjected to analysis to determine the specificity of the protein–DNA interactions. The different CTS were analyzed by binding, competition and super-shift. The competition or super shift experiments are indicated with + signs. Analysis of **(A)** CTS7, **(B)** CTS9, **(C)** CTS17, **(D)** CTS38, **(E)** CTS8, **(F)** CTS44 and **(G)** CTS15.

clones is shown in Fig. 2. As can be seen, the different CTS clones are bound by CDXA with different strengths, suggesting variations in affinity (Fig. 2A–G). The different affinities could result from slightly different binding sites or a different number of sites in the various clones. All the binding reactions were performed in the presence of large excess of [poly (dI-dC): poly (dI-dC)] to compete non-specific protein–DNA interactions. Under these conditions, addition of unlabeled CTS fragment was capable of competing the binding activity (Figs. 2A–F). This result supports the notion that the interaction between the CDXA protein and the CTS clones is specific. In addition, parallel protein extracts prepared form bacteria producing the GST protein and not the fusion protein were utilized as negative controls. As determined from the sequencing results, the CTS clones have a high AT content (see below). Therefore, as a control for the specificity of binding we also tried competition assays with a fragment rich in GC. This fragment contains the first 135 bp of the *CdxA* cDNA clone, C33 (24) and is 60% GC rich. The

Figure 3. Sequence of the genomic CTS clones. The sequence of the eight CTS clones is shown. The regions protected by the CDXA in the DNase I footprinting assay are shown as shaded areas. The sequences identified as matching the CAD consensus binding sequence are underlined.

addition of the this fragment, C33Not, over a large range of molar ratios did not efficiently compete for the binding of the CDXA protein to the CTS clones (Fig. 2B, 2D, 2E).

In order to ascertain that the CDXA protein is the one responsible for the mobility shift in the binding assays, the protein−DNA complexes were tested for the presence of this protein. The addition of the 6A4 anti-CDXA antibody (34) to the binding reactions resulted in super-shifting of the protein−DNA complexes during EMSA (Fig. 2A−G). This further retardation is the result of the formation of a tripartite complex of antibody−protein−DNA demonstrating that indeed the CDXA protein is part of the binding complex. Addition of a non-specifc antibodies against *E.coli* proteins did not affect the mobility of the protein−DNA complex (data not shown).

Sequence of the CTS clones

The seven CTS clones were subjected to sequence analysis by the dideoxi chain termination method (36). The sequence of the CTS clones is shown in Fig. 3. As can be seen in this figure, the length of the clones ranges between 58 to 160 bp. One of the most striking features of these clones is that on the average they are 61% AT rich, as opposed to the chicken genome which is about 60% GC rich (37). The sequence of the CTS clones was screened for sequences resembling the CAD binding site core sequence 5′TTTATG. This kind of analysis revealed that between the seven clones there are over 20 sites that are similar to the CAD-type binding sites (Fig. 3, underlined). Eventough the CTS clones are not large in themselves, in many instances several CAD-type sites are found in close proximity (Fig. 3).

Determination of the CDXA binding site

A main focus of this study is to determine the consensus binding site of the CDXA protein. As a result of the size of the genomic fragments isolated, it is of importance to determine the specific sequences with which the CDXA protein interacts. The amount of CDXA protein needed for complete binding of the CTS clones was determined by titration in mobility shift assays (data not shown). From this kind of analysis it became evident that the CTS44, CTS15 and CTS 17 clones bind weakly. Thus, the analysis of the sites bound in these clones was precluded (Fig. 2C, 2F, 2G). The clones CTS7, 8, 9 and 38 were subjected to DNase I footprinting analysis to determine the CDXA binding site consensus sequence. As shown in Fig. 4, apart from CTS38, the other three clones have multiple footprinted regions (Fig. 4). Also, in many instances, binding of the CDXA-GST fusion protein to the DNA resulted in the formation of DNase I hypersensitive sites (Fig. 4). All DNase I footprinting reactions were analyzed next to sequencing reactions to determine the sequence of the footprinted region. From the analysis of the footprinted regions it could be seen that the length of the regions protected by the CDXA-GST protein is about 15−20 bp. When smaller footprints were observed, they were usually separated by a hypersensitive site and together they add up to about 20 bp.

The sequence of the footprinted regions was analyzed to determine the CDXA consensus binding site. The sequence of the CDXA protected sites is shaded in Fig. 3. In several instances the footprinted regions included some of the previously identified CAD-type binding sites. In order to determine the CDXA binding site from the protected regions, the sequence of the different footprinted sites with the addition of 1 or 2 flanking nucleotides were utilized to run the consensus program (35). The program was applied to find a consensus sequence of 7 bp in length from the footprinted regions (Table 2). The consensus CDXA binding site identified 5′A, A/T, T, A/T, A, T, A/G is very similar to the sequence deduced from the analysis of the random oligonucleotides.

CDXA-mediated transcriptional activation

All the analysis of the CTS clones as CDXA target sequences was based on *in vitro* studies of protein−DNA interactions. Therefore, it is of interest to test whether the CTS clones can function as transcriptional regulators in cells in culture. In order to test the interaction between the CDXA protein and the CTS clones in cells in culture, an CDXA expression plasmid and a CTS-CAT reporter plasmid were constructed. To express the CDXA protein in cells in culture, two vector plasmids were used: pCMVβ and pSV40β, where expression is driven by the human cytomegalovirus promoter or the SV40 promoter respectively (38). The LacZ coding region was removed from these plasmids and replaced by the *CdxA* cDNA clone, C33 (24). LTK$^-$ cells were stably transfected with either plasmid and independent clones were isolated. The different clones were tested for the expression of the CDXA protein first by ELISA using the 6A4 anti-CDXA antibody (34) and then they were subsequently analyzed by western blotting (data not shown). Two cell lines were chosen for further study: B2, where expression is driven by the CMV promoter, and SV40FLG4, generated with the SV40 promoter. From the ELISA analysis it was concluded that SV40FLG4 has higher CDXA levels than B2 (data not shown).

The reporter plasmids were built in the pTE2 plasmid where expression of a chloramphenicol acetyltransferase gene (CAT) flanked by SV40 splice junctions and polyadenylation signals is

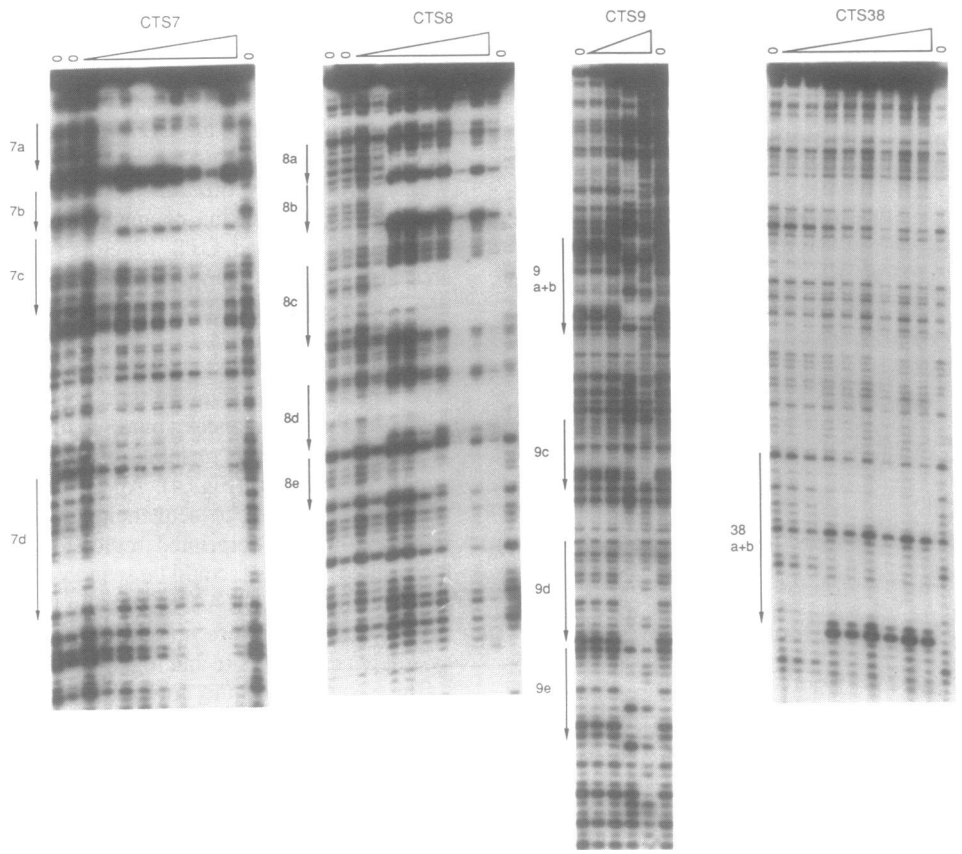

Figure 4. Footprinting analysis of the CTS clones. The four clones that bind to the CDXA protein with highest affinity, CTS7, CTS8, CTS9 and CTS38 were studied by DNase I footprinting. Results of these experiments localized the sequences with which the CDXA protein interacts. Footprinting reactions with increasing amounts of CDXA-GST fusion protein are shown. The footprinted regions are shown with arrows. The sequence of the footprinted regions was determined by running sequencing reactions in the same gel with the footprinting reactions.

driven by a thymidine kinase minimal promoter (39). Upstream to the TK promoter CTS clones were subcloned to generate the pTE2/7, pTE2/8 and pTE2/38 plasmids for CTS7, CTS8 and CTS38 respectively. The different plasmids were transfected into the two CDXA expressing cell lines and into LTK$^-$ cells. After 48 hours the cells were harvested and the CAT assay was performed. The transfections were performed several times and the results of one such experiment are shown in Fig. 5. As shown in Fig. 5 the background of transfection of the reporter plasmids into LTK$^-$ cells is very low giving only about 1−1.5% conversion. Parallel transfections of the same responder plasmids into the two CDXA expressing cell lines resulted in CAT activities ranging from 16 to 60% conversion (Fig. 5). In accordance with the results from the ELISA analysis of the amounts of CDXA protein in the two cell lines shown, the SV40FLG4 line which has higher CDXA levels also gave the highest CAT activity than the B2 line (Fig. 5). Expression of the CDXA protein in LTK$^-$ cells results in about a 10 to 40 fold increase in CAT activity when mediated through the isolated target sequences CTS7, CTS8 or CTS38. Transfection of these three cell lines with the pTE2 vector only resulted in basal levels of CAT activity even in the prescence of the CDXA protein. These results suggest that the CDXA protein is mediating an enhancing activity by binding to its target sequences.

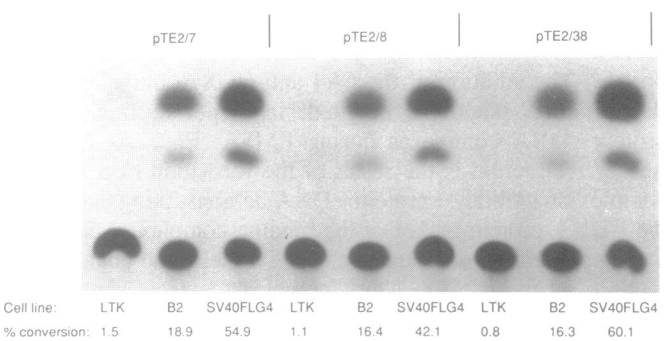

Figure 5. Transcriptional enhancement by the CDXA protein. CAT reporter plasmids where the CTS7, CTS8 or CTS38 were subcloned upstream to the cytomegalovirus promoter were utilized to test the effect of the CDXA protein on transcription. Three cell line were used, LTK$^-$ cells and two stable transformants which constitutively express the CDXA protein, B2 and SV40FLG4.Results of the CAT assay and their quantitation with the different plasmids and in the different cell lines is shown.

DISCUSSION

Many putative gene products have been identified which contain domains known to provide DNA binding capabilities to the proteins such as homeodomains, zinc fingers, helix-loop-helix

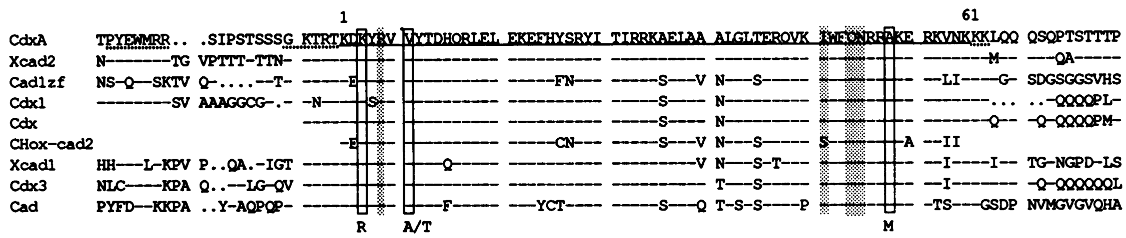

Figure 6. The *caudal*-type homeodomains. Sequence comparison of the *caudal*-type homeodomains and flanking regions, sequences are shown as variations from the CDXA sequence. The hexapeptide motif (YEWMRR) and the five amino acids just upstream from the homeodomain and specific to this family (GKTRT) are shown as dotted underlined. The amino acids shaded or boxed are those identified as responsible for interactions with the DNA. The shaded residues are identical with the ones found in proteins which bind sequences with the TAAT core motif. The boxed amino acids are common to the members of the *caudal* family which bind the TTTATG sequence.

and others. In many of these cases the DNA binding function is inferred from amino acid sequence similarity and the specific DNA sequence to which they bind is unknown. In order to study the protein–DNA interactions of proteins of this kind the DNA sequence to which they bind has to be identified. A number of different approaches have been described that address the problem of target sequence isolation and identification. In most cases an *in vitro* DNA-binding reaction is performed and then the methods vary depending on the approach for the isolation of the protein–DNA complexes (10–15). In order to isolate and study the binding site of the CDXA protein, a CDXA-GST fusion protein was utilized that can be readily precipitated with glutathione-agarose beads (14). This approach overcomes the necessity of antibodies for the precipitation of the protein–DNA complex. In all the above mentioned variations of the *in vitro* binding approach the target DNA is devoid of proteins raising the possibility that the target sequences isolated match the consensus binding site but in the case of genomic DNA they are random sequences and not be a part of target genes. In addition, the binding conditions used in all these assays are probably different from the *in vivo* conditions.

The CTO clones and more extensively the CTS clones were studied by EMSA to study the specificity of the reaction and the protein bound in the complexes. From the binding assays it became evident that some of the clones bind with high affinity while others bind weakly. To determine the specificity of the reaction, competition assays were performed with unlabeled DNA from CTS clones or with DNA fragments chosen for their high content of G/C. Competitions with CTS clones affected the amount of protein–DNA complex detected by EMSA while competitions with irrelevant DNA fragments required higher concentrations to affect binding. Furthermore, the identity of the protein in the protein–DNA complexes was determined by EMSA with the addition of specific anti-CDXA antibodies (34). The results of these experiments showed that indeed the CDXA protein is the one present in the protein–DNA complexes. Together these results show that the binding of the CDXA protein to the target sequences isolated is specific.

Further characterization of the CTS clones was performed by DNase I footprinting analysis of the CTS clones that bound strongly to the CDXA protein. The genomic fragments isolated with the CDXA protein showed by this kind of analysis several protected regions suggesting that more than one CDXA molecule can bind to them. Also, the footprinted regions appear in fragments relative small in size, further supporting the clustering of CDXA binding sites as the preferred configuration. In the case of the CTO clones in contrast to the CAD binding site which is composed of two repeats of the consensus, the oligonucleotides contained only one copy of the consensus sequence. At least in half of the clones this situation was 'corrected' by the joining of two oligonucleotides containing a CAD-type binding site. From our analysis we have identified the sequence A, A/T, T, A/T, A, T, A/G as the consensus binding site for the CDXA protein. This consensus binding site includes as one of its variations the CAD binding site in the *ftz* promoter (20) and is almost identical to the binding site of the hamster CDX3 protein in the FLAT-E and FLAT-F elements of the insulin I gene (26). The two FLAT elements bound by CDX3 are not identical and they represent two of the variations allowed by the CDXA consensus binding site. The difference between the CDXA consensus and the FLAT sites is at position #6 where in the CDXA consensus there is a strong tendency for a T but in the FLAT sites in one there is a G while in the second there is a C. In the case of the CAD binding sites position #6 is always taken up by a T but position #1 in two cases is a T while the other two one is C and the other G. In the CDXA consensus position #1 is usually an A. In any case, the three *caudal*-type proteins bind almost identical sites with differences that may be attributed to small differences in the homeodomain and flanking sequences. The small differences in the sequence of the homeodomain and the known target sequences within the *caudal* family may be of great importance. These small differences may allow the different members of the family to be present in the same cells and still exhibit differential target specificity resulting in differential regulation.

The binding of homeobox proteins has been studied for a number of vertebrate and fly genes. For many of the homeodomain containing proteins it was found that the binding sequence preferred by many of these proteins includes the core 5′TAAT sequence (7, 40–46). Analysis of three fly homeodomain proteins that preferentially bind the TAAT sequence revealed that the interactions with the DNA take place with amino acids from the carboxy and the amino termini of the homeodomain (40, 41, 43–46). From the amino terminus R3 and R5 are important for the interaction with the DNA, and probably the amino acid at position 7 of the homeodomain (44). At the carboxy terminus in helix 3, two important positions are N51 and I47. M54 may also be involved in the recognition outside the core (45). Genetic and biochemical evidence has pointed out that the amino acid at position 50 plays a central role in determining sequence preference in the positions 3′ to the core (41, 47).

The vertebrate *caudal* family of homeobox genes is one of the sub-families outside the *HOX* clusters more extensively studied (1). Nine genes of this family have been cloned and described in vertebrates such as zebrafish, *Xenopus*, rats, hamsters, mice and chickens (21–28). Comparisons of the homeodomains in this family revealed that the homologies range between 80 to 100% (Fig. 6). Most of the differences between the members of the family are at the end of helix 1 and some variation is seen at the end of helix 2 and in the turn (45, 46). In addition to the high homology between the homeodomains in this family upstream to the homeodomain there is high similarity up to the hexapeptide which is characteristic to members of this family. Detailed analysis of the residues shown to play an important role in the sequence specific binding in other genes and the members of this family show some interesting family-specific changes (Fig. 6). Positions R3 and A7 or T7 from *Antp, en, Ubx, Dfd* or *ftz*, are changed to K3 and V7 respectively. Furthermore, in helix 3 M54 in all members of this family is changed to A. The changes in this positions which have been identified as important for the protein–DNA interactions may explain why for three members of the *caudal*-family studied, the consensus is TTTATG and not just the core TAAT sequence. Interestingly, two other homeobox proteins have been shown to bind the TTTATG sequence (48). The two proteins are the products of the human *HoxD9* and *HoxD10* genes. Both proteins contain the I47, Q50, N51 and M54 in helix 3 (49). On the other hand both proteins have K3, R5 and P7 at the amino terminus of the homeodomain. This configuration in the amino acid sequence of the amino terminus of these two homeodomain proteins resembles that of the *caudal* family raising the possibility that this region affects strongly the sequence specificity of the DNA-binding interaction.

Finally, the interaction between the CDXA protein and the CTS clones was tested by transfection into tissue culture cells. From these experiments it was obtained that the CDXA protein can increase the transcriptional activity of a reporter gene through interaction with the CTS clones. The CAD protein has been shown to be important for *ftz* expression in the posterior part of the fly embryo. Also, reporter plasmids containing the CDRE's in transgenic flies were shown to be dependent on the integrity of the TTTATG sequence (20). Similar results in tissue culture were obtained with the *Cdx3* gene and sequences from the rat insulin I gene (26). Together these results again show that as a family, the different *caudal*-type proteins appear to function as positive regulators of gene expression. In addition, the sequences identified for the vertebrate genes as binding sites appear to represent functionally relevant sequences that the proteins can recognize *in vivo*. As it regards to the *Drosophila cad* gene, the biochemical evidence is supported by genetic evidence and transgenic flies. Together these results strengthen the observations demonstrating the positive regulation obtained with the hamster and chicken proteins.

ACKNOWLEDGEMENTS

We wish to thank G.Pillemer for her comments on the manuscript. This work was supported by a grant from the Council for Tobacco Research, USA to A.Fainsod.

REFERENCES

1. McGinnis,W. and Krumlauf,R. (1992). *Cell* **68**, 283–302.
2. Scott,M.P., Tamkun,J.W. and Hartzell,G.W. III (1989). *Biochim. Biophys. Acta* **989**, 25–48.
3. McGinnis,W., Levine,M.S., Hafen,E., Kuroiwa,A. and Gehring,W.J. (1984a). *Nature* **308**, 428–432.
4. Scott,M.P. and Weiner,A.J. (1984). *Proc. Natl. Acad. Sci. USA* **81**, 4115–4119.
5. Laughon,A. and Scott,M.P. (1984). *Nature* **310**, 25–31.
6. Shepherd,J.C.W., McGinnis,W., Carrasco,A.E., De Robertis,E.M. and Gehring,W.J. (1984). *Nature* **310**, 70–71.
7. Hayashi,S. and Scott,M.P. (1990). *Cell* **63**, 883–894.
8. Fainsod,A., Bogarad,L.D., Ruusala,T., Lubin,M. Crothers,D.M. and Ruddle,F.H. (1986). *Proc. Natl. Acad. Sci. USA* **83**, 9532–9536.
9. Cho,K.W.Y., Goetz,J., Wright,C.V.E., Fritz,A., Hardwicke,J. and De Robertis,E.M. (1988). *EMBO J.* **7**, 2139–2149.
10. Kinzler,K.W. and Vogelstein,B. (1989). *Nucl. Acids. Res.* **17**, 3645–3653.
11. Thiesen,H.-J. and Bach,C. (1990). *Nucl. Acids Res.* **18**, 3205–3209.
12. Kirch,H.-C., Krüger,H. and Schulte Holthausen,H. (1991). *Nucl. Acids Res.* **19**, 3156.
13. Blackwell,T.K. and Weintraub,H. (1990). *Science* **250**, 1104–1110.
14. Fainsod,A., Margalit,Y., Haffner,R. and Gruenbaum,Y. (1991). *Nucl. Acids Res.* **19**, 4005.
15. Nøorby,P.S., Pallisgaard,N., Pedersen,F.S. and Jøorgensen,P. (1992). *Nucl. Acids Res.* **20**, 6317–6321.
16. Gould,A.P., Brookman,J.J., Strutt,D.I. and White,R.A.H. (1990). *Nature* **348**, 308–312.
17. Kinzler,K.W. and Vogelstein,B. (1990). *Molec. Cell. Biol.* **10**, 634–642.
18. Mlodzik,M., Fjose,A. and Gehring,W.J. (1985). *EMBO J.* **4**,2961–2969.
19. Macdonald,P.M. and Struhl,G. (1986). *Nature* **324**, 537–545.
20. Dearolf,C.R., Topol,J. and Parker,C.S. (1989). *Nature* **341**, 340–343.
21. Duprey,P., Chowdhury,K., Dressler,G.R., Balling,R., Simon,LD., Guenet,J. and Gruss,P. (1988). *Genes Dev.* **2**, 1647–1654.
22. James,R. and Kazenwadel,J. (1991). *J. Biol. Chem.* **266**, 3246–3251.
23. Blumberg,B.,Wright,C.V.E., De Robertis,E.M.and Cho,K.W.Y.(1991). *Science* **253**, 194–196.
24. Frumkin,A., Rangini,Z., Ben-Yehuda,A., Gruenbaum,Y. and Fainsod,A. (1991). *Development* **112**, 207–219.
25. Serrano,J., Scavo,L., Roth,J., de la Rosa,E.J. and de Pablo,F. (1993). *Biochem. Biophys. Res. Comm.* **190**, 270–276.
26. German,M.S., Wang,J., Chadwick,R.B. and Rutter,W.J. (1992). *Genes Dev.* **6**, 2165–2176.
27. Freund,J.-N., Boukamel,R. and Benazzouz A. (1992). *FEBS lett.* **314**, 163–166.
28. Joly,J.S., Maury,M., Joly,C., Duprey,P., Boulekbache,H. and Condamine,H. (1992) *Differentiation* **50**, 75–87.
29. Smith,D.B. and Johnson,K.S. (1988). *Gene* **67**, 31–40.
30. Pognonec,P., Kato,H., Sumimoto,H., Kretzschmar,M. and Roeder,R.G. (1991). *Nucl. Acids Res.* **19**, 6650
31. Fried,M. and Crothers,D.M. (1981). *Nucl. Acids Res.* **9**, 6505–6525.
32. Lichtsteiner,S., Wuarin,J. and Schibler,U. (1987). *Cell* **51**, 963–973.
33. Gorman,C.M., Moffat,L.F. and Howard,B.H. (1982). *Molec. Cell. Biol.* **2**, 1044–1051.
34. Frumkin,A., Haffner,R., Shapira,E., Tarcic,N., Gruenbaum,Y. and Fainsod,A. (1993). *Development* **118**, 553–562.
35. Hertz,G.Z., Hartzell,G.W. and Stormo,G.D. (1990). *CABIOS* **6**, 81–92.
36. Sanger,F., Nicklen,S. and Coulson,A.R. (1977). *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
37. Eden,F.C. and Hendrick,J.P. (1978). *Biochemistry* **17**, 5838–5844.
38. MacGregor,G.R. and Caskey,C.T. (1989). *Nucl Acids Res.* **17**, 2365.
39. Edlund,T., Walker,M.D., Barr,P.J. and Rutter,W.J. (1985) *Science* **230**, 912–916.
40. Hanes,S.D. and Brent,R. (1989). *Cell* **57**, 1275–1283.
41. Hanes,S.D. and Brent,R. (1991). *Science* **251**, 426–430.
42. Samson,M.L., Jackson-Grusby,L. and Brent,R. (1989). *Cell* **57**, 1045–1052.
43. Florence,B., Handrow,R. and Laghon,A. (1991). *Molec. Cell. Biol.* **11**, 3613–3623.
44. Ekker,S.C., von Kessler,D.P. and Beachy,P.A. (1992). *EMBO J.* **11**, 4059–4072.
45. Otting,G., Qian,Y.Q., Biletter,M., Muller,M., Affolter,M., Gehring,W.J. and Wuthrich,K.(1990). *EMBO J.* **9**, 3085–3092.
46. Kissinger,C.R., Liu,B., Martin-Blanco,E., Kornberg,T.B. and Pabo,C.O. (1990). *Cell* **63**, 579–590.
47. Percival-Smith,A., Müller,M., Affolter,M. and Gehring,W. (1990). *EMBO J.* **9**, 3967–3974.
48. Arcioni,L., Simeone,A., Guazzi,S., Zappavigna,V., Boncinelli,E. and Mavilio,F. (1992). *EMBO J.* **11**, 265–277.
49. Zappavigna,V., Renucci,A., Izpisua-Belmonte,J.C., Urier,G., Peschle,C. and Duboule,D. (1991). *EMBO J* **10**, 4177–4187.