MOLECULAR AND CELLULAR BIOLOGY, Dec. 1993, p. 7961–7970
0270-7306/93/127961-10$02.00/0

Vol. 13, No. 12

Isolation of cDNAs Encoding the *Drosophila* GAGA Transcription Factor

WALTER C. SOELLER,† CHARLES EUK OH,‡ AND THOMAS B. KORNBERG*

Department of Biochemistry and Biophysics, University of California, San Francisco, California 94143

Received 22 July 1993/Returned for modification 2 September 1993/Accepted 20 September 1993

To investigate the mechanisms involved in expression of the *Drosophila melanogaster engrailed* gene, we purified GAGA protein, one of several putative transcriptional activator proteins that binds to the proximal region of the *engrailed* promoter. Antibodies raised against GAGA protein were used to demonstrate that the protein is present in all nuclei of young embryos. We isolated cDNA clones encoding GAGA protein in which a putative 519-codon open reading frame contains general sequence motifs characteristic of other transcription factors. These include stretches of polyglutamine, a 60-amino-acid region with 18 (30%) lysine or arginine residues, and a single putative zinc finger motif. In addition, a 120-residue N-terminal region shares significant sequence homology with several other known *Drosophila* transcription factors, including those encoded by *Broad Complex* and *tramtrack*. Up to 35-fold GAGA protein-dependent stimulation of transcription in Schneider line 2 tissue culture cells was observed after transfection of GAGA protein-encoding sequences. The GAGA gene is present in one copy in the *Drosophila* genome, at cytological location 70EF, and it encodes RNAs which vary in size between 2.4 and 4.4 kb.

Many of the gene products that organize the *Drosophila melanogaster* embryo are expressed in restricted domains, and their accurate deployment is critical to normal development. Among these is engrailed, a homeodomain-containing protein that is expressed in a subset of the ectoderm and neuroectoderm. In gastrulating embryos, cells containing engrailed protein are in 15 concentric, single-cell-wide rings (15, 19, 29), and the presence of functional engrailed protein in these cells and in the posterior compartments that are populated by their descendants contributes to processes that form and maintain the segments and compartments.

Genetic analysis has implicated the pair-rule segmentation genes *even-skipped* (*eve*) and *fushi tarazu* (*ftz*) in activating *engrailed* transcription in the gastrulating embryo (reviewed by Lawrence [31]). Genes that have been implicated in its subsequent regulation include *engrailed* itself (16), *Polycomb* (9, 39), and *wingless* (3, 16, 38). It is not known whether the action of the protein products of any of these genes on *engrailed* regulation is direct, although logic might suggest that eve and ftz proteins, both of which are homeodomain-containing transcription factors, directly activate *engrailed* transcription in the gastrulating embryo. However, there is little direct evidence to suggest that these or other transcription factors act directly on the *engrailed* promoter. In vitro studies have identified sequences in the *engrailed* 5′ upstream region that bind eve and ftz proteins (23), and *engrailed* is activated in a pattern that corresponds precisely with the anterior borders of the eve and ftz stripes (32). However, functional studies of the *engrailed* promoter failed to identify the eve and ftz binding sites in the *engrailed* upstream region as important sequences for regulation (10), and the response of *engrailed* to ectopic activation of *eve* suggests that *eve* regulation of *engrailed* is indirect and is mediated through the *runt* and *odd-skipped* pair-rule genes (37). Furthermore, *engrailed* autoregulation, at least in imaginal discs, appears to involve the extracellular hedgehog protein (50) and so might be indirect as well.

To identify proteins that directly regulate *engrailed*, we have developed an in vitro transcription system that is prepared from extracts of *Drosophila* embryos and which accurately initiates *engrailed* transcription (48). Despite the large size of the *engrailed* regulatory region (>35 kb), a high density of functional sequence elements within the first 400 bp upstream of the sites of transcription initiation was observed. Eight distinct regions in this 400-bp upstream region were recognized by sequence-specific DNA-binding activities. Competition studies of DNA binding suggested that seven of the eight binding sites were recognized by the same protein, implying that several copies of this protein are involved in initiating *engrailed* transcription. Common to six of the seven binding sites is the sequence GAGAG, and no other obvious homologies are present. The seventh site contains the related sequence GAGTG. Given that the competition studies suggested that these sites are recognized by the same binding protein, the GAGAG pentamer may form a core recognition sequence, and the binding protein has been designated GAGA (16). Analysis of deletion constructs transfected into Schneider line 2 (S2) tissue culture cells indicates that GAGA binding sites are important to achieve maximal levels of expression (48).

GAGA sequences have been noted previously in the proximal promoter regions of several *Drosophila* genes, including *Ultrabithorax* (*Ubx*) (6), *hsp70* (33), *hsp26* (35), *E74* (51), *Krüppel* (28), *eve* (45), *ftz* (52), and genes encoding alcohol dehydrogenase (Adh) (4), the laminin B2 chain (11), and histones H3 and H4 (20). Two related models for GAGA protein action have been proposed. In studies of the sequences important for activation of *hsp26*, GAGA protein binding sites were found to be necessary for optimal activation and nuclease accessibility of *hsp26-lacZ* fusion transgenes, prompting the suggestion that the function of GAGA protein is to mold its target regulatory sequences into an open, nucleosome-free state and, in so doing, to prepare target genes for activation (35). GAGA protein was also

* Corresponding author.
† Present address: Department of Molecular Genetics, Pfizer Central Research, Groton, CT 06340.
‡ Present address: Department of Entomology and Parasitology, University of California, Berkeley, CA 94720.

found to counteract an inhibitory activity of histone H1 that was observed when *Krüppel* transcription was studied in cell extracts of *Drosophila* embryos (13). To account for these observations, it was suggested that GAGA protein is an antirepressor that prevents nonspecific binding of histone H1 in critical regions of target promoters (13).

In this present study, we purified and characterized GAGA protein from *Drosophila* embryos and isolated the GAGA gene and GAGA cDNAs. We found that GAGA protein exists in multiple forms, and the form represented in the cDNAs that we isolated activated transcription in transfected tissue culture cells in a binding site-dependent manner. The level of activation was dependent both on the amount of GAGA protein and on the number of binding sites. GAGA protein was localized to nuclei in embryos but was not restricted to specific spatial domains. These observations indicate that GAGA protein is not likely to be one of the factors responsible for the posterior compartment specificity of *engrailed* expression. Rather, these observations are consistent with the previously proposed models which suggested that GAGA protein prepares a promoter for activation by competing for promoter sequences to which nonspecific repressors would otherwise bind and that its role as an activator is permissive rather than instructional.

MATERIALS AND METHODS

Purification of GAGA and peptide sequencing. GAGA protein was purified by using an affinity chromatography procedure similar to that described by Biggin and Tjian (6). Nuclear extracts were prepared as previously described (48) and fractionated on heparin-Sepharose (Pharmacia). Typically, 20 ml of nuclear extract (10 mg of protein per ml) prepared from 80 g of 0- to 12-h-old Oregon R embryos was applied to 20 ml of resin equilibrated with HEMGK (25 mM *N'*-2-hydroxyethylpiperazine-*N'*-2-ethanesulfonic acid [HEPES; pH 7.6], 0.5 mM EDTA, 12.5 mM $MgCl_2$, 10% glycerol, 1 mM dithiothreitol, 125 mM KCl) plus 0.1% Nonidet P-40 (NP-40). The GAGA binding activity was eluted with 400 mM KCl in HEMG and applied to a Sephacryl S300 column (10 by 80 cm) equilibrated in 125 mM KCl–HEMG. Fractions containing GAGA binding activity (as assessed by DNase I footprinting on an *engrailed* promoter fragment) were pooled (24 ml) and applied to 1 ml of GAGA affinity resin after equilibration of the pool with nonspecific DNA competitor [poly(dI-dC); 5 μg/ml; Sigma). Sepharose CL2B (Pharmacia) affinity resin was prepared as described by Kadonaga and Tjian (25) except that DNA used for coupling was the insert from plasmid pCBA69. pCBA69 has multiple copies of the proximal *engrailed* promoter sequence (−39 to −108), each of which contains three GAGA binding sites. The protein eluted from this column with 1.0 M KCl–HEMG plus 0.1% NP-40 was dialyzed, reapplied to fresh affinity resin (0.5 ml) in the presence of 10-fold-less poly(dI-dC), and eluted with 1.0 M KCl–HEMG without NP-40. Aliquots of protein fractions were denatured in sodium dodecyl sulfate (SDS) plus β-mercaptoethanol and subjected to SDS-polyacrylamide gel electrophoresis. Yields of GAGA protein (molecular size range of 67 to 93 kDa) were typically 5 to 10 μg from 80 g of embryos, and its estimated purity was greater than 90%. An estimated 10 μg of twice-affinity-purified GAGA protein was methanol-chloroform precipitated, resuspended in 25 μl of formic acid, and added to 25 μl of CNBr (120 mg/ml; dissolved in 70% formic acid). The chemical cleavage reaction was allowed to proceed for 22 h at room temperature, after which time the sample was diluted with 0.5 ml of distilled H_2O and lyophilized. The resulting fragments were subjected to reverse-phase high-pressure liquid chromatography (HPLC) in trifluoroacetic acid, eluted with a linear gradient of acetonitrile, and subjected to gas-phase protein sequencing.

Isolation and analysis of GAGA cDNA clones with oligodeoxynucleotides. Two degenerate oligodeoxynucleotides were synthesized on the basis of the amino acid sequences of two CNBr-generated peptides, peptide 1 (NTTAEGDNTVGSD) and peptide 2 (QNVIHIVGDQVFIPQQQQPQP). The oligonucleotide sequence based on peptide 1 (oligo 1) (5′-ACNG TRTTRTCNCCCTC-3′ [R = G+A; N = T+A+G+C; I = inosine) was 64-fold degenerate. The sequence based on peptide 2 (oligo 2) (5′-GGGATGAAIACCTGRTCICCIAC GATRTGGATIACGTTCTGCAT-3′) was fourfold degenerate. These oligonucleotide pools were labeled at the 5′ end with [γ-^{32}P]ATP by using T4 polynucleotide kinase and were used to screen a λgt10 *Drosophila* cDNA library derived from 6- to 9-h embryonic RNA (41). The probe from oligo 2 (44-mer) was hybridized to duplicate lifts in 50% formamide at 42°C overnight and then washed twice in 2× SSC (1× SSC is 0.15 M NaCl plus 0.015 M sodium citrate)–0.1% SDS for 30 min at 50°C. The oligo 1 probe (17-mer) was hybridized at 37°C overnight to a set of duplicate lifts from the secondary screen and then washed in tetramethylammonium chloride (53). Every plaque that hybridized to oligo 2 also hybridized to oligo 1. Of six clones carried through two more rounds of plaque purification, all displayed the same restriction pattern upon digestion with *Eco*RI. One recombinant phage (clone 7B) was selected for subcloning.

Sequencing of GAGA cDNAs. The *Eco*RI fragments of GAGA clone 7B, 1.6 and 0.4 kb in length, were subcloned into the *Eco*RI site of pEMBL18. The convenient spacing of several useful restriction sites was utilized to generate shorter fragments for sequencing. Single-stranded DNA was prepared as described previously (36). Sequencing of both strands was accomplished by the dideoxy method of Sanger et al. (47), using a T7 kit from U.S. Biochemical and the M13 universal and reverse sequencing primers.

Preparation of anti-GAGA antiserum. Mice initially were injected intraperitoneally with 10 μg of GAGA protein that had been purified two times over an affinity column and suspended in complete Freund's adjuvant. Antiserum used for expression cloning and whole-mount embryo immunostaining was obtained 1 week after administration of a second booster injection with antigen suspended in incomplete Freund's adjuvant.

Isolation and analysis of GAGA cDNA clones by expression cloning. A 1:1,000 dilution of the crude mouse anti-GAGA antiserum was used to screen a λgt11 *Drosophila* cDNA library. Secondary screens were carried out with the 1.6-kb *Eco*RI fragment of the GAGA cDNA. A single clone (clone 5A) was isolated and characterized. Clone 5A contained an insert of 2.2 kb whose sequence was identical to that of the 2.0-kb insert in clone 7B except for the 5′ noncoding region, where they diverge at a position 235 bases upstream of the first ATG codon.

Southern and Northern (RNA) analysis. *Drosophila* Oregon R genomic DNA was prepared as described by Levis et al. (34). Ten micrograms of genomic DNA was digested with restriction endonuclease *Eco*RI, *Bgl*II, or *Hin*dIII and electrophoresed through a 0.8% agarose gel (10 by 6 cm) for 400 V-h. The DNA was denatured, blotted, and hybridized to the 1.6-kb *Eco*RI GAGA cDNA fragment (36). The probe was labeled with [α-^{32}P]dATP by random priming. Total *Drosophila* RNA was prepared by the method of Chomczynski

and Sacchi (12). Embryos were staged by collection for the specified times on yeast-agar plates, using twice-precleared population cages. RNA was electrophoresed on 1.0% HGT agarose in 6.5% formaldehyde and blotted by capillary action onto a Nytran membrane. To avoid including polyglutamine sequences in the probe, the randomly primed probe for Northern hybridization was generated from a 684-bp *Pst*I-*Ava*I GAGA cDNA subclone to eliminate the polyglutamine-encoding sequences.

Chromosome in situ hybridization. The 1.6-kb *Eco*RI GAGA cDNA fragment was labeled with biotin-16-dUTP (ENZO Biochemicals) by random priming. Chromosome squashes, pretreatment, hybridization, and washes were performed as described by Hafen (21). Signal detection was accomplished with the Detek-I-hrp kit (ENZO).

Whole-mount immunohistochemical staining of GAGA during embryogenesis. Zero- to 14-h-old embryos were dechorionated and fixed as described previously (27), incubated with a 1:500 dilution of mouse antiserum, and stained with 3,3'-diaminobenzidine.

Plasmid constructions. pD-33CAT (18) was linearized with restriction endonuclease *Sal*I, phosphatase treated, filled in, and blunt-end ligated to single and multiple copies of a double-stranded 19-mer oligonucleotide (5'-CTCGCCCTC TCGCTCCCGC-3') encompassing the GAGA binding site in region C of the *engrailed* promoter (48). pCBA69 was generated by initially ligating the *Hpa*I-*Xho*II *engrailed* promoter fragment (−38 to −107, containing the A, B, and C GAGA binding sites) to *Bam*HI-*Bgl*II-digested pUC18B. pUC18B is a derivative of pUC18 containing a *Bgl*II linker inserted into the *Hin*cII site of the pUC18 polylinker. Derivatives (clones 1.16, 1.3, and 1.4) of this plasmid containing 10 to 14 copies of the insert were derived by digesting pCBA69 with *Xho*II, isolating the insert, and ligating it to itself in the presence of restriction endonucleases *Bam*HI and *Bgl*II in order to select for head-to-tail concatemers. Concatemers of 700 to 1,000 bp in size were religated to pUC18B.

Full-length GAGA cDNA was prepared by ligating the 1.6- and 0.6-kb fragments from λgt11-6B. A ligation product containing the 0.6-kb fragment bound to both ends of the 1.6-kb fragment was isolated and digested with *Spe*I. This fragment was ligated to *Xba*I-cut pUC18 and reisolated by a *Hin*cII-*Bgl*II digestion. The resulting isolate was ligated to *Bam*HI-digested pPac (3' *Bam* end filled in) to generate pPac-GAGA. The presence of the inverted repeat (0.3 kb on both sides of the 1.6-kb GAGA *Eco*RI fragment) necessitated growth in a recombination-deficient host strain.

To construct an *Escherichia coli* expression vector, the *Eco*RI ends of the 1.6-kb fragment of the GAGA cDNA were filled in and blunt-end ligated to *Bam*HI-cut and filled-in pAR3038 (46). This produced a plasmid (pAR-GAGA) that expresses a fusion product substituting 13 amino acid residues of phage T7 gene 10 (MASNTGGQQMGRI) for the first 5 amino-terminal residues of GAGA (MSLPM). *E. coli* BL21 was transformed with pAR-GAGA and induced to express the fusion product as described by Studier and Moffat (49). Bacterial protein extracts were prepared as described previously (1), using Brij 58 as the lysis agent. Crude extracts (500 ml of cells) from both control and pAR-GAGA-transformed cells were loaded onto heparin-Sepharose columns (2.0 ml) equilibrated in 0.125 M KCl–HEMG, and GAGA protein was eluted with 0.4 M KCl–HEMG.

Transfection of *Drosophila* tissue culture cells and CAT assays. S2 cells were grown in 75-cm^2 T flasks, using Schneider's *Drosophila* medium (GIBCO) plus 10% heat-inactivated fetal calf serum. For transfections, 4×10^6 cells were plated onto 6-cm-diameter tissue culture dishes in a volume of 5 ml of medium 24 h prior to the start of the transfection procedure. Calcium phosphate precipitation of the plasmid DNA and addition to the cells were carried out as described previously (30). Typically, the total amount of plasmid added per 6-cm-diameter plate was 21 μg. This included 1 μg of pcop-βGal to allow correction for variations in transfection efficiencies, 1 to 10 μg of reporter plasmid, and 10 μg of expression plasmid. pEMBL18 DNA was used as a carrier when necessary to bring the total DNA amount to 21 μg. Chloramphenicol acetyltransferase (CAT) and β-galactosidase activity assays were performed as described previously (48).

TABLE 1. Purification of GAGA protein from 120 g of 1- to 12-h embryos

Step	Protein (mg)	Activity (10^3 FP units)[a]	Sp act (10^3 U/mg)	% Yield
Crude nuclear extract	640			
Heparin-Sepharose	224	64	0.29	100
S300	35.6	20	0.56	31.2
DNA affinity 1	0.077	14.8	192	23.0
DNA affinity 2	0.020	6.7	335	10.5

[a] One FP unit is equal to the minimum amount of protein required to completely footprint the GAGA binding site on 10 fmol of probe DNA.

Electrophoretic mobility shift assays. Mobility shift assays were carried out essentially as described previously (2). Typically, 3 to 10 ng of ^{32}P-end-labeled DNA fragments was incubated with either purified *Drosophila* GAGA protein or heparin-Sepharose protein fractions of *E. coli* BL21(DE3) transformed with either expression plasmid pAR-3038 or

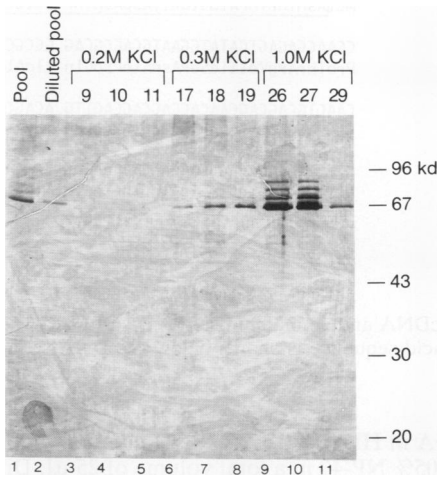

FIG. 1. Affinity purification of GAGA protein from *Drosophila* embryos. An SDS–10% polyacrylamide gel of protein fractions applied and eluted off the GAGA affinity resin was silver stained. For the second passage, fractions 40 to 44 from the first passage were pooled, diluted sixfold in HEMG without KCl, and applied to a second column containing fresh GAGA-resin. Lanes: 1 and 2, pool from first column; 3 to 5, fractions 9 to 11 eluted with 0.2 M KCl–HEMG; 6 to 8, fractions 17 to 19 eluted with 0.3 M KCl–HEMG; 9 to 11, fractions 26, 27, and 29 eluted with 1.0 KCl–HEMG. Molecular size markers are indicated at the right. Fraction volumes were 0.5 ml; all buffers except the 0.3 M and 1.0 M KCl–HEMG washes of the second column contained 0.1% NP-40.

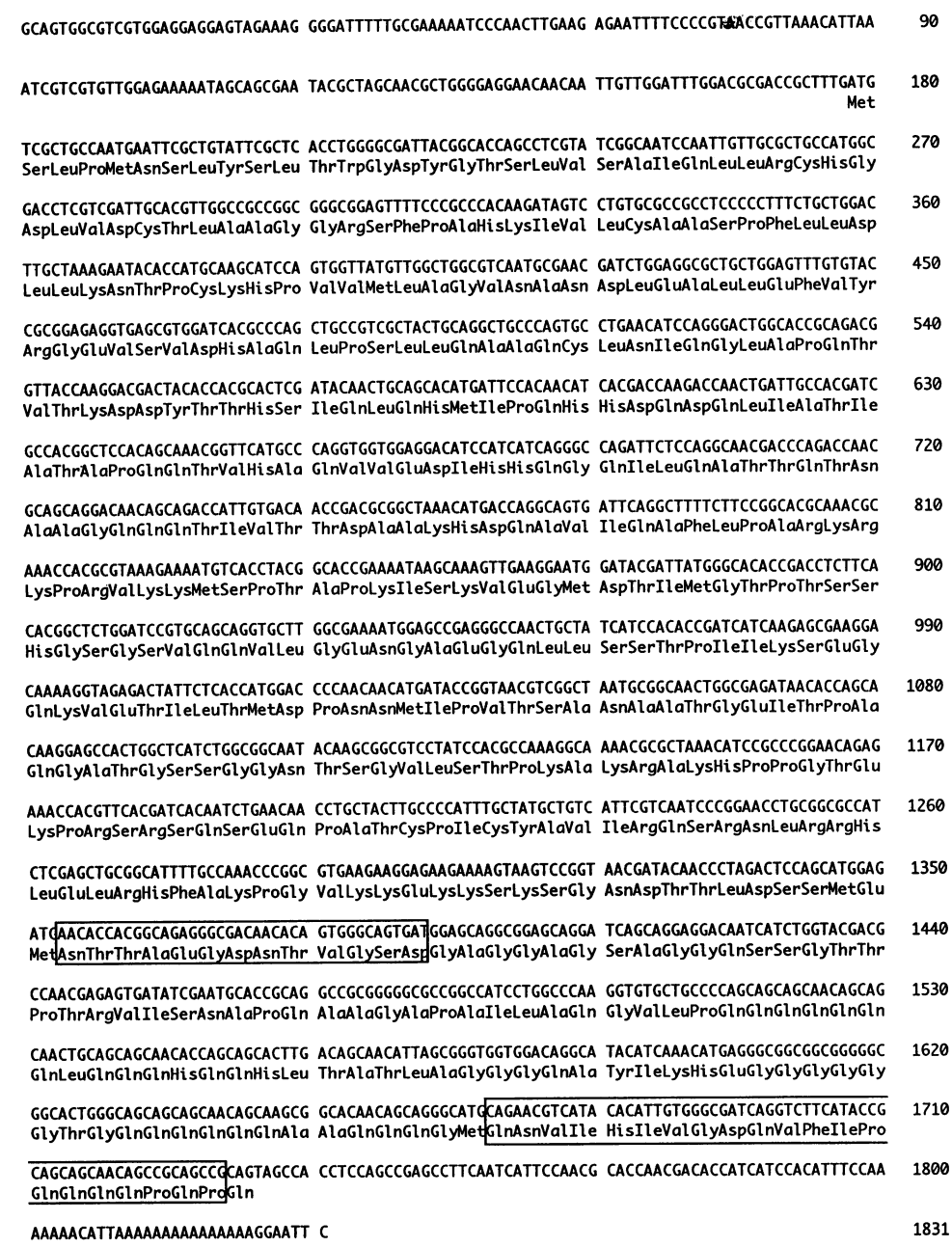

FIG. 2. cDNA and amino acid sequences of GAGA. All sequence data were confirmed by sequencing both DNA strands. The deduced 519-amino-acid sequence is shown below the DNA sequence. The two peptides identified by amino acid sequencing are boxed.

pAR-GAGA in HEMG–0.125 M KCl plus 200 ng of poly(dI-dC) and 0.05% NP-40 in a total volume of 25 μl. Dilutions of protein fractions were carried out in HEMG plus 100 μg of bovine serum albumin (Pentax fraction V) per ml. Protein fractions (50 to 100 ng) were preincubated on ice for 20 min with nonspecific or specific competitor before addition of probe. Incubations were continued on ice for 15 min, at which time an equal volume of 50% glycerol was added and the reactions were loaded onto a 4.5% (29:1 acrylamide/bisacrylamide) gel prerun overnight. The samples were electrophoresed at 4°C for 1 to 3 h at 200 V, depending on the size of the labeled fragment.

Nucleotide sequence accession number. The nucleotide sequence accession number of the sequence determined in this study is L22205.

RESULTS

Isolation of GAGA protein and DNA. Extracts of *Drosophila* embryos that activate specific transcription contain proteins that bind the *engrailed* gene in a sequence-specific manner (48). To isolate and characterize the proteins responsible for binding the GAGA sequence elements, extracts from *Drosophila* embryos were chromatographed sequentially over heparin-Sepharose and an S300 gel permeation column and twice over a sequence-specific DNA affinity

```
GAGA        MSLPMNSLYSLTWGDYGTSLVSAIQLLRCHGDLVDCTLAAGGRSFPAHKIVLCAASPFLLDLLKNTPCKHPVVML
BR-C        --MDDTQHFCLRWNNYQSSITSAFENLRDDEAFVDVTLACEGRSIKAHRVVLSACSPYFRELLKSTPCKHPVILL
TRAMTRACK   ---MASQRFCLRWNNHQSNLLSVFDQLLHAETFTDVTLAVEGQHLKAHKMVLSACSPYFNTLFVSHPEKHPIVIL

GAGA        AGVNANDLEALLEFVYRGEVSVDHAQLPSLLQAAQCLNIQGLAPQTVTKDDYTTHSIQLQHMIPQC
BR-C        QDVNFMDLHALVEFIYHGEVNVHQKSLQSFLKTAEVLRVSGLTQQQAEDTHSHLAQIQNLANSGG-
TRAMTRACK   KDVPYSDMKSLLDFMYRGEVSVDQERLTAFLRVAESLRIKGLTEVNDDKPSPAAAAAGAGATGSEV
```

FIG. 3. Comparison of the first 118 amino acids of the GAGA, tramtrack, and BR-C proteins. Identical residues or residues representing conservative changes are shaded.

column (Table 1). Protein that eluted from the DNA affinity column at greater than 200 mM KCl was distributed among at least four different species, ranging in mass from approximately 67 to 90 kDa (Fig. 1). The 67-kDa species was the predominant species. The chromatographic profile did not change after repeated applications to the affinity column (not shown). This profile of proteins was similar to that reported for GAGA protein by Biggin and Tjian (6).

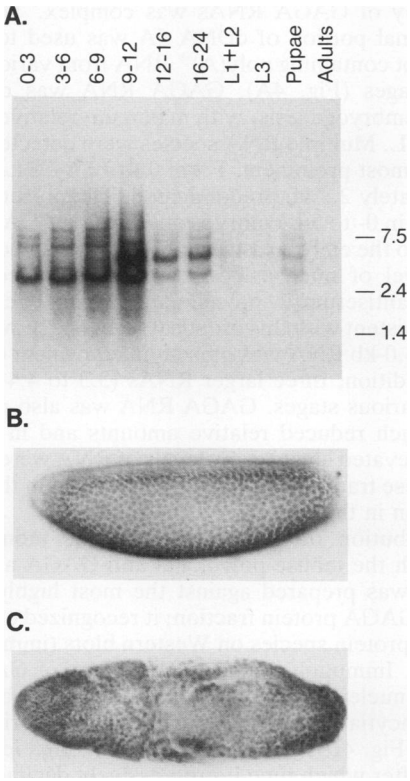

FIG. 4. GAGA gene expression and GAGA protein distribution. (A) Autoradiogram of a Northern blot hybridized with a 684-bp *Pst*I-*Ava*I subclone of the GAGA cDNA. Lanes contained 5 μg of 0- to 3-h, 3- to 6-h, 6- to 9-h, 9- to 12-h, 12- to 16-h, and 16- to 24-h embryo, larval instar 1+2 (L1+L2), larval instar 3 (L3), pupa, and adult poly(A)$^+$ RNAs. Positions of RNA molecular size markers (Bethesda Research Laboratories no. 5620SA) that were in an adjacent lane are indicated at the right in kilobases. (B and C) Photomicrographs taken under bright-field optics of whole mounts of embryos stained with mouse anti-GAGA antiserum. (B) Syncytial blastoderm embryo; (C) gastrulating, germ band-extended embryo.

To isolate GAGA cDNAs, two approaches were taken. In the first approach, purified GAGA protein was subjected to gas-phase protein sequencing. Since the N terminus was apparently blocked, GAGA was cleaved into 13 fragments with CNBr, and the peptides were separated by reverse-phase HPLC. Two CNBr-generated peptides were sequenced, and two degenerate oligonucleotides capable of encoding them were synthesized (Materials and Methods). These oligonucleotides were used to screen a λgt10 cDNA library generated from RNA isolated from embryos 6 to 9 h AEL (after egg laying). In the second approach, purified protein was used to produce anti-GAGA antibody in mice, and a λgt11 expression cDNA library was screened with the antiserum that was produced. Isolates were obtained from both cDNA libraries. All isolated phage had similar patterns of restriction fragments (not shown); one from each library was sequenced. The sequences were identical with the exception that the phage isolated from the λgt10 library (cDNA 5A), with a total insert of 1,835 nucleotides, included several hundred additional residues at the presumptive 5′

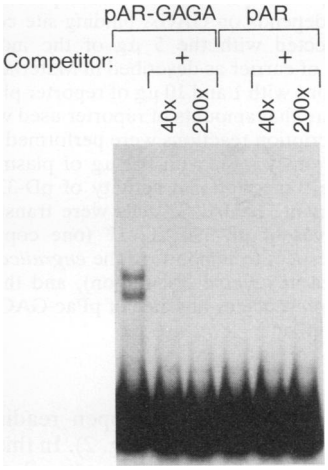

FIG. 5. DNA binding activity of GAGA fusion protein expressed in *E. coli*. Mobility shift assays reveal the DNA sequence-specific binding of recombinant GAGA protein. Lane 1, 3 ng (37 fmol) of pCBA69 (^{32}P-5′-end-labeled) probe plus 40 ng of heparin-Sepharose fraction of *E. coli* BL21(DE3) transformed with pAR-GAGA and induced with isopropyl thiogalactopyranoside (IPTG); lanes 2 and 3, same as lane 1 plus 40- and 200-fold molar excess of competitor GAGA binding sites, respectively; lane 4, 3 ng of probe plus 54 ng of heparin-Sepharose fraction of *E. coli* BL21(DE3) transformed with the parent vector pAR3038; lanes 5 and 6, same as lane 4 plus 40- and 200-fold molar excess of binding sites, respectively.

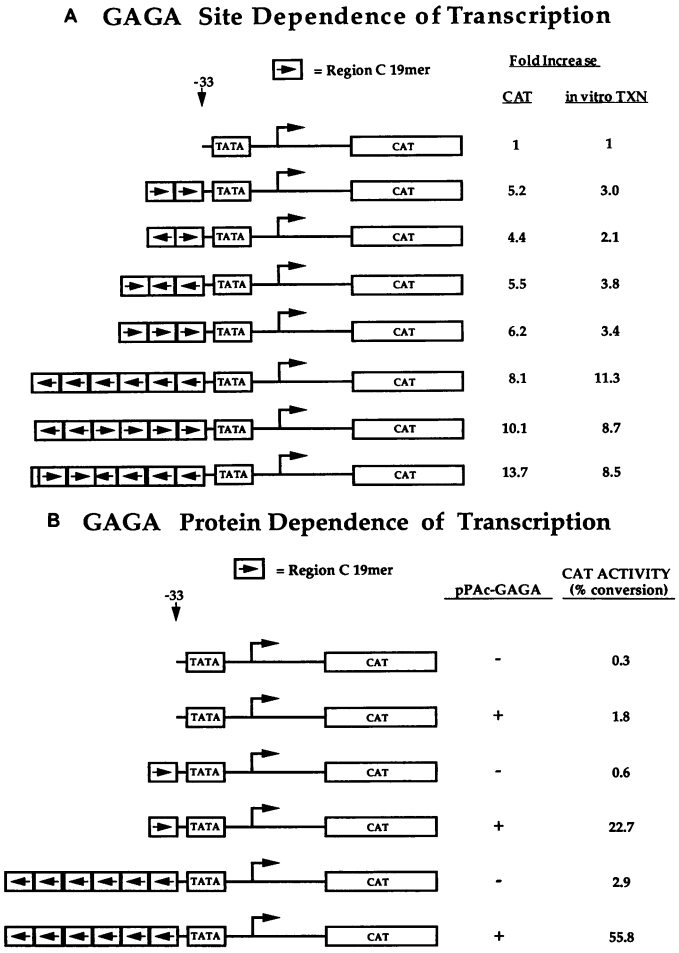

FIG. 6. GAGA-dependent transcriptional activity. (A) The transcriptional activity of the pD-33CAT minimal promoter construct in vivo and in vitro depends on GAGA binding site copy number. S2 cells were transfected with the 5 μg of the indicated reporter plasmids and 5 μg of carrier as described in Materials and Methods. Parallel transfections with 1 and 10 μg of reporter plasmid were also performed to ensure that amounts of reporter used were nonsaturating. In vitro transcription reactions were performed and quantitated as described previously (48), with 0.2 μg of plasmid DNA as the template. (B) The transcriptional activity of pD-33CAT was activated by recombinant GAGA. S2 cells were transfected with two region C derivatives of pD-33CAT, 1F (one copy in the native orientation with respect to region C in the *engrailed* promoter) (48) and 6R (six copies in reverse orientation), and the parent vector pD-33CAT in the presence or absence of pPac-GAGA or the parent expression vector pPac.

end. Both cDNAs contained an open reading frame that includes the two CNBr peptides (Fig. 2). In this open reading frame, use of the first methionine codon downstream of a stop codon (TAA, nucleotides 87 to 90) in the same reading frame putatively generates a protein containing 519 residues. A second potential initiation codon five codons downstream is an alternative translation start.

The presumptive GAGA protein has a calculated molecular weight of 54,826 and is basic (pI = 8.5). Notable features include stretches of polyglutamine in the C-terminal region, a 60-amino-acid region (amino acids 320 to 379) with 18 (30%) arginine and lysine residues, and a single putative zinc finger domain (343-ATCPICYAVIRQSRNLRRHLELRH-366). Such sequences have been observed frequently in other transcription factors. The high proportion of glutamine (12.3%) is unusual among known sequences of *Drosophila* proteins: fewer than 1% have more (26). GAGA protein also has unusually few charged residues and few acidic residues. Most of its positively charged residues are clustered in two regions, residues 209 to 228 and 320 to 379, a characteristic common to less than 3% of known proteins (26). In addition (Fig. 3), the N-terminal 120 residues share significant homology with the N-terminal regions of the proteins encoded by the *Broad Complex* (*BR-C*; 47% identity and 51% similarity [14]) and *tramtrack* (38% identity and 45% similarity [22, 44]) genes. The protein products of both *BR-C* and *tramtrack* are believed to be transcription factors (7, 8, 14, 22, 43).

To characterize the genomic organization of the GAGA gene, clone 5A cDNA was used to probe Southern blots of restriction digests of *Drosophila* DNA and to hybridize to salivary gland polytene chromosomes. Hybridization results indicated that the GAGA gene is present in a single copy (not shown) at site 70EF (not shown). 70EF is the site of the gene *62*, whose known mutant allele dominantly enhances position effect variegation of *wrn4H* (17).

Expression of GAGA. Expression of the GAGA gene and synthesis of GAGA protein were monitored in *Drosophila* embryos and larvae. In general, the levels of GAGA products correlated with the known levels of *engrailed* products, and the array of GAGA RNAs was complex. The unique non-N-terminal portion of cDNA 5A was used to probe a Northern blot containing poly(A)$^+$ RNA from various developmental stages (Fig. 4A). GAGA RNA was detectable throughout embryogenesis, with maximum relative levels at 9 to 12 h AEL. Multiple RNA species were detected, two of which were most prominent. From 0 to 12 h AEL, an RNA of approximately 2.4 kb predominated. The presence of the 2.4-kb RNA in 0- to 3-h embryos suggests that GAGA RNA is provided to the embryo as part of the maternal dowry, and the high level of nuclear-localized antigen detected with anti-GAGA antiserum in precellular, nuclear cycle 10 embryos is consistent with this proposal (data not shown). After 9 h AEL, a 3.0-kb RNA was present in equivalent or greater levels. In addition, three larger RNAs (3.3 to 4.4 kb) were present at various stages. GAGA RNA was also present in larvae in much reduced relative amounts and in pupae in somewhat elevated levels. Low levels of RNA were detected in adults; these transcripts may be synthesized in the ovaries for deposition in the eggs.

The distribution of GAGA protein was monitored in embryos with the mouse polyclonal anti-GAGA antiserum. This serum was prepared against the most highly purified *Drosophila* GAGA protein fraction; it recognized each of the four GAGA protein species on Western blots (immunoblots) (not shown). Immunohistochemical staining of GAGA was restricted to nuclei. Prominent staining was first observed in nuclei of syncytial blastoderm stage embryos, prior to cellularization (Fig. 4B). Staining persisted at high levels until 12 h AEL, after which time it progressively declined. There was no apparent spatial regulation of its synthesis at any time during embryogenesis, indicating that in embryos, GAGA is a ubiquitous factor (Fig. 4C).

GAGA protein is a potent transcriptional activator. The protein encoded by the GAGA 5A cDNA was expressed in *E. coli* as a fusion protein (EC-GAGA). EC-GAGA lacked the N-terminal 5 amino acids and had substituted 13 amino acids from T7 gene 10 in the expression vector pAR-3038 (pET-3C). Western analysis with the mouse anti-GAGA antiserum detected a 67-kDa band in extracts prepared from

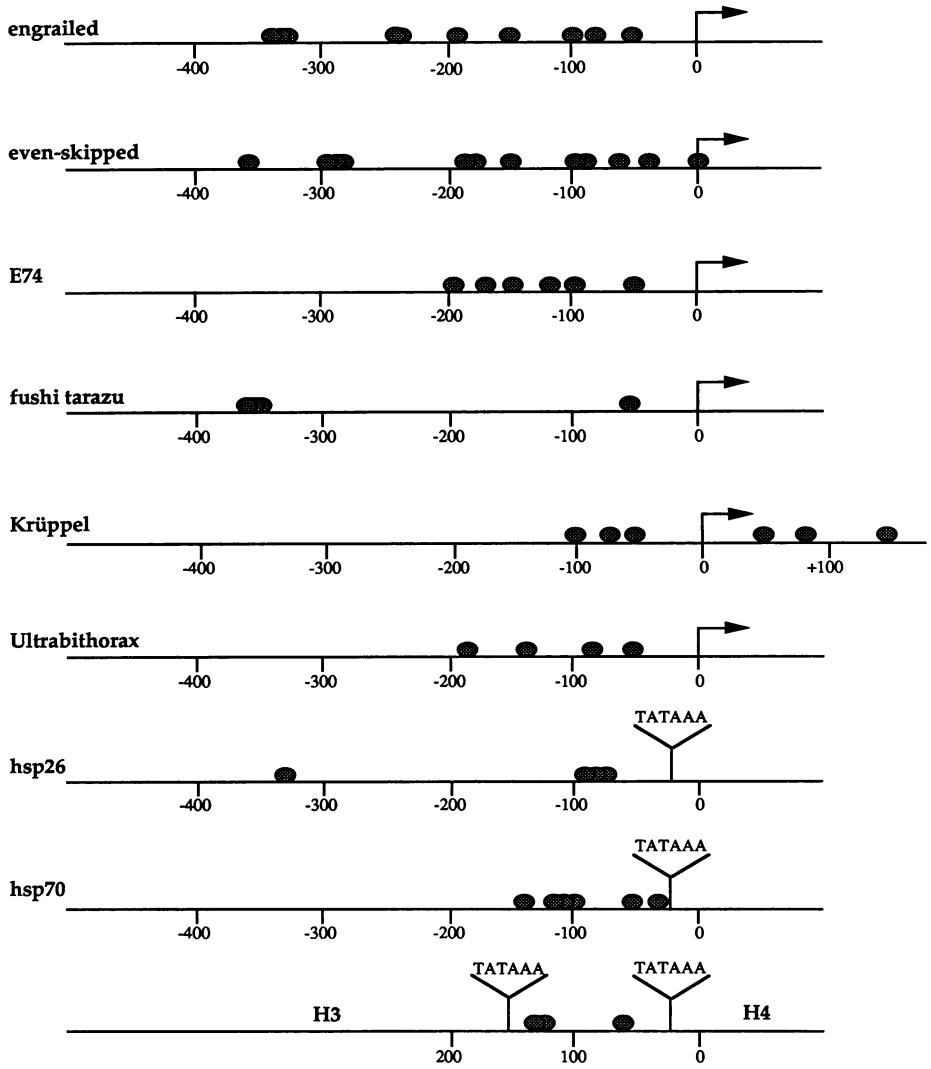

FIG. 7. Distribution of GAGA protein binding sites. The approximate locations of GAGA protein binding sites are indicated with the shaded ovals. These binding sites are defined by regions protected from digestion by DNase I. Data are from the following sources: *engrailed* (48); *even-skipped* (45); *E74* (51); *fushi-tarazu* (52); *Krüppel* (28); *Ultrabithorax* (6); *hsp26* (35); *hsp70* (33); and histone H3 and H4 genes (20).

GAGA-transformed bacteria (data not shown). Mobility shift assays indicate that these extracts possessed GAGA site-specific binding activity (Fig. 5). In addition, identical DNase I footprint patterns were observed with *Drosophila* GAGA protein and EC-GAGA (35).

To assess the capacity of the 5A cDNA-encoded GAGA protein to activate transcription, GAGA binding site-dependent transcription was quantitated in vitro and in vivo. We constructed a series of reporter plasmids in which synthetic promoters containing single or multiple GAGA sites were placed upstream of an Adh basal (−33) promoter and a CAT reporter gene (18). These plasmids were transfected into S2 tissue culture cells, in which endogenously synthesized GAGA protein activated CAT expression (Fig. 6A). Similar relative levels of activation were observed with in vitro transcription assays (Fig. 6A). Despite the presence of GAGA protein in S2 cells, significant stimulation (19- to 38-fold) of target promoters was also observed when the level of GAGA protein was increased by cotransfection (Fig. 6B). The degree of stimulation was dependent on both the level of GAGA protein and the number of GAGA binding sites. We conclude that the 5A cDNA encodes the GAGA transcription factor.

DISCUSSION

In this report, we describe a biochemical approach to investigating the mechanism of transcriptional control of the *Drosophila engrailed* gene. Our previous work had identified eight sequence elements in the proximal promoter region of the *engrailed* gene which are required to promote optimal levels of transcription from the *engrailed* promoter (48). In this report, we describe the purification of the protein GAGA, which binds to seven of these sites. We also describe the cloning of the GAGA gene, the developmental profile of GAGA gene expression, and the immunohistochemical localization of GAGA protein in embryos.

Isolation of the GAGA protein by DNA affinity chromatography yielded a set of proteins whose electrophoretic migration in SDS ranged in apparent molecular size between

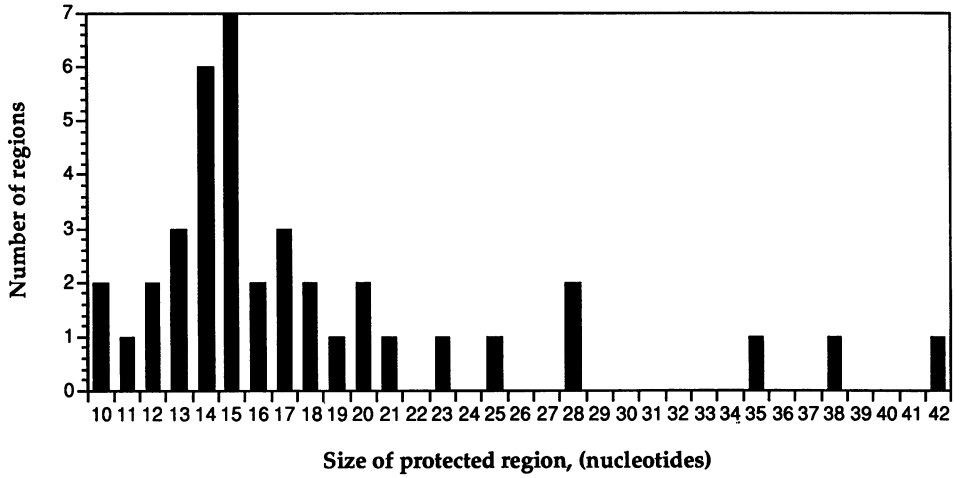

FIG. 8. Sizes of GAGA protein binding sites. Shown is a plot of the sizes of regions that were protected from DNase I digestion by GAGA protein in the proximal promoter regions of *engrailed* (48), *eve* (45), *E74* (51), *ftz* (52), *Krüppel* (28), *Ubx* (6), *hsp26* (35), *hsp70* (33), and the histone H3 and H4 genes (20). These size estimates are approximate, since the numbers of protected nucleotides in the complementary strands at each site are not necessarily equivalent, and most of the data are presented for one strand only.

67 and 90 kDa. The origin of this diversity is unclear. A similar profile of protein species was observed by Biggin and Tjian from an affinity purification protocol that employed a GAGA site from the proximal promoter region of the *Drosophila Ubx* gene (6). Extraction of individual species from an SDS-gel indicated GAGA site binding to be associated with several, but not all, of the species in their purified preparation. Two explanations for the multiple GAGA species might be considered. Alternative states of glycosylation could variably alter electrophoretic mobility, and evidence for glycosylation of GAGA has been reported. Modification of three of the species in the purified preparation of Biggin and Tjian (6) by galactosyltransferase indicates that these proteins contain a terminal O-linked *N*-acetylglucosamine residue (24). Variable glycosylation could therefore account for all or part of the diversity of GAGA. Alternatively, it is possible that glycosylation of GAGA is uniform and that the multiplicity of GAGA species is a consequence of variations in primary sequence. It might be relevant that two other *Drosophila* genes, *BR-C* and *tramtrack*, encode zinc finger domain-containing proteins that have an N-terminal domain which is highly related to GAGA (Fig. 3). Both the *BR-C* and *tramtrack* genes produce multiple protein products by alternative RNA splicing, generating a set of proteins that have a common N-terminal region and different C-terminal zinc finger-containing domains (14, 44). If this mechanism for producing multiple transcription factors from a single locus is conserved among other members of this family, at least part of the diversity of GAGA proteins might be similarly explained. Indeed, recent studies of Adf-2, a protein that binds to the Adh distal promoter and is thought to repress Adh transcription (5), indicate that Adf-2 is GAGA and that several different forms of Adf-2/GAGA may be encoded by a heterogeneous population of cDNAs. The identified differences are in the putative C-terminal domains (4).

The possibility that GAGA proteins with different primary sequences exist suggests an intriguing explanation for our experiments analyzing GAGA protein-dependent activation of transcription in S2 cells. These cells express GAGA endogenously, yet appropriate reporter constructs can nevertheless be stimulated by transfection of the GAGA cDNA clone 5A (Fig. 6). Although GAGA protein may simply be limiting in S2 cells, a more interesting possibility is that the form of GAGA present in S2 cells differs from cDNA 5A and that the 5A form has greater affinity for the GAGA elements in the *engrailed* promoter region.

The relatedness of the structures of the GAGA, tramtrack, and BR-C proteins is intriguing. GAGA, tramtrack, and BR-C are thought to be transcription factors which bind DNA through a zinc finger-containing C-terminal domain. All are evolutionarily related by 120-residue N-terminal domains whose structure-function motifs have not yet been determined. The members of this *Drosophila* gene family number between 15 and 25 (42). We do not yet understand the function of the conserved N-terminal domain but are investigating whether the related N-terminal domains of this protein family mediate homo- or heteromultimerization.

GAGA protein is a transcription activator. Its known targets include genes such as *Krüppel*, *eve*, *ftz*, *Ubx*, and *engrailed* that are expressed during early embryogenesis. Other targets include *hsp26* and *hsp70*, the *actin 5C*, laminin B2, Adh, and histone H3 and H4 gene, and *E74*, genes that are expressed abundantly at various developmental stages. We do not know whether different genes are activated by different forms of GAGA, whether different putative protein partners target GAGA to different genes, or whether the varied levels of GAGA protein alter its effective specificity. As discussed above, GAGA protein in S2 tissue culture cells is present at less than saturating levels for activation of *engrailed* gene promoter sites (Fig. 6B). Two testable possibilities are (i) that the level of GAGA protein is regulated in a manner that influences target selection and (ii) that the form of endogenous GAGA protein in S2 cells is not the most efficient form of GAGA activator for *engrailed*.

GAGA binding sites have been identified by DNase I footprinting in the proximal promoter regions of 10 *Drosophila* genes (6, 20, 28, 45, 48, 51, 52). Inspection of these promoter regions reveals that (i) GAGA binding sites cluster in proximal promoter regions of these genes, and the number of distinct binding sites in each ranges between two and seven (Fig. 7); (ii) 9 of the 10 genes have a GAGA binding site placed 35 to 50 bases upstream of the transcription start

(Fig. 6); and (iii) the sizes of the protected sites range between approximately 10 and 42 bases (Fig. 8). The protected regions are 14 to 15 bases at most of the sites, and although we do not yet know what the stoichiometry of GAGA protein is at any of these sites, at least two to three molecules probably bind to the larger sites. It is attractive to speculate that GAGA protein binds to clustered sites in a cooperative fashion, especially given the two charge clusters in its primary sequence. We observed that adding multiple GAGA binding sites to synthetic promoters augmented GAGA-dependent stimulation of transcription in transfected cells (Fig. 6). Thus, the multiplicity of binding sites may be functionally significant and may relate to the proposed role of GAGA protein in allowing transcription in the face of potential inhibitory effects of histones, HMG proteins, etc. (13). The benefit of clearing the proximal promoter regions of proteins with nonspecific repressive activities could be proportional to the size of the cleared area.

ACKNOWLEDGMENTS

We thank Doug Brutlag for the λgt11 expression cDNA library, Ulrike Heberlein for plasmid pD-33CAT, F. William Studier for pAR3038 and *E. coli* BL21(DE3), and Mark Krasnow for the expression vector pPac. We thank Tim Karr for help with the immunostaining, Mary Rykowski for in situ mapping of the GAGA gene, and Jim Kealy for HPLC purification of GAGA CNBr fragments and peptide sequencing. We thank S. Elgin, C. Benyajati, and F. Karch for communicating results prior to publication.

This work was supported by grants from the National Institutes of Health.

REFERENCES

1. **Abarzua, P., W. Soeller, and K. Marians.** 1984. Mutational analysis of primosome assembly sites. J. Biol. Chem. **259:** 14286–14292.
2. **Ausubel, F. M., R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidman, J. A. Smith, and K. Struhl (ed.).** 1992. Current protocols in molecular biology, vol. 2. John Wiley & Sons, New York.
3. **Bejsovec, A., and A. A. Martinez.** 1991. Roles of wingless in patterning the larval epidermis of Drosophila. Development **113:**471–485.
4. **Benyajati, C.** Personal communication.
5. **Benyajati, C., A. Ewel, J. McKeon, M. Chovav, and E. Juan.** 1992. Characterization and purification of Adh distal promoter factor 2, Adf-2, a cell-specific and promoter-specific repressor in Drosophila. Nucleic Acids Res. **20:**4481–4489.
6. **Biggin, M. D., and R. Tjian.** 1988. Transcription factors that activate the *Ultrabithorax* promoter in developmentally staged extracts. Cell **53:**699–711.
7. **Brown, J. L., S. Sonoda, H. Ueda, M. P. Scott, and C. Wu.** 1991. Repression of the Drosophila fushi tarazu (ftz) segmentation gene. EMBO J. **10:**665–674.
8. **Brown, J. L., and C. Wu.** 1993. Repression of *Drosophila* pair-rule segmentation genes by ectopic expression of *tramtrack*. Development **117:**45–58.
9. **Busturia, A., and G. Morata.** 1988. Ectopic expression of homeotic genes caused by the elimination of the *Polycomb* gene in Drosophila imaginal epidermis. Development **104:**713–720.
10. **Chen, C.-N., and T. B. Kornberg.** Unpublished data.
11. **Chi, H. C., D. Juminaga, S. Y. Wang, and C. F. Hui.** 1991. Structure of the Drosophila gene for the laminin B2 chain. DNA Cell Biol. **10:**451–466.
12. **Chomczynski, P., and N. Sacchi.** 1987. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Anal. Biochem. **162:**156–159.
13. **Croston, G. E., L. A. Kerrigan, L. M. Lira, D. R. Marshak, and J. T. Kadonaga.** 1991. Sequence-specific antirepression of histone H1-mediated inhibition of basal RNA polymerase II transcription. Science **251:**643–649.
14. **DiBello, P. R., D. A. Withers, C. A. Bayer, J. W. Fristrom, and G. M. Guild.** 1991. The Drosophila broad-complex encodes a family of related proteins containing zinc fingers. Genetics **129:**385–397.
15. **DiNardo, S., J. Kuner, J. Theis, and P. O'Farrell.** 1985. Development of embryonic pattern in *D. melanogaster* as revealed by accumulation of the nuclear *engrailed* protein. Cell **43:**59–69.
16. **DiNardo, S., E. Sher, J. Heemskerk-Jongens, J. Kassis, and P. O'Farrell.** 1988. Two-tiered regulation of spatially patterned *engrailed* gene expression during Drosophila embryogenesis. Nature (London) **332:**604–609.
17. **Dorn, R., J. Szidonya, K. Günter, M. Sehnert, H. Taubert, E. Archoukieh, B. Tschiersch, H. Morawietz, G. Wustmann, G. Hoffmann, and G. Reuter.** 1993. P transposon-induced dominant enhancer mutations of position-effect variegation in *Drosophila melanogaster*. Genetics **133:**279–290.
18. **England, B. P., U. Heberlein, and R. Tjian.** 1990. Purified *Drosophila* transcription factor, Adh factor-1 (Adf-1), binds to sites in several *Drosophila* promoters and activates transcription. J. Biol. Chem. **265:**5086–5094.
19. **Fjose, A., W. McGinnis, and W. Gehring.** 1985. Isolation of a homeobox-containing gene from the *engrailed* region of Drosophila and the spatial distribution of its transcript. Nature (London) **313:**284–289.
20. **Gilmour, D. S., G. H. Thomas, and S. C. Elgin.** 1989. Drosophila nuclear proteins bind to regions of alternating C and T residues in gene promoters. Science **245:**1487–1490.
21. **Hafen, E.** 1986. Rubin lab methods book.
22. **Harrison, S. D., and A. A. Travers.** 1990. The tramtrack gene encodes a Drosophila finger protein that interacts with the ftz transcriptional regulatory region and shows a novel embryonic expression pattern. EMBO J. **9:**207–216.
23. **Hoey, T., and M. Levine.** 1988. Divergent homeo box proteins recognize similar DNA sequences in Drosophila. Nature (London) **332:**858–861.
24. **Jackson, S. P., and R. Tjian.** 1988. O-glycosylation of eukaryotic transcription factors: implications for mechanisms of transcriptional regulation. Cell **55:**125–133.
25. **Kadonaga, J. T., and R. Tjian.** 1986. Affinity purification of sequence specific binding proteins. Proc. Natl. Acad. Sci. USA **83:**5889–5893.
26. **Karlin, S.** Personal communication.
27. **Karr, T. L., M. J. Weir, Z. Ali, and T. Kornberg.** 1989. Patterns of *engrailed* protein in early Drosophila embryos. Development **105:**605–612.
28. **Kerrigan, L. A., G. E. Croston, L. M. Lira, and J. T. Kadonaga.** 1991. Sequence-specific transcriptional antirepression of the Drosophila *Krüppel* gene by the GAGA factor. J. Biol. Chem. **266:**574–582.
29. **Kornberg, T., I. Siden, P. O'Farrell, and M. Simon.** 1985. The *engrailed* locus of Drosophila: in situ localization of transcripts reveals compartment-specific expression. Cell **40:**45–53.
30. **Krasnow, M. A., E. E. Saffman, K. Kornfeld, and D. S. Hogness.** 1989. Transcriptional activation and repression by Ultrabithorax proteins in cultured Drosophila cells. Cell **57:**1031–1043.
31. **Lawrence, P. A.** 1992. The making of a fly. Blackwell Scientific Publications, Oxford.
32. **Lawrence, P. A., and P. Johnston.** 1989. Pattern formation in the Drosophila embryo: allocation of cells to parasegments by even-skipped and fushi tarazu. Development **105:**739–746.
33. **Lee, H., K. W. Kraus, M. F. Wolfner, and J. T. Lis.** 1992. DNA sequence requirements for generating paused polymerase at the start of hsp70. Genes Dev. **6:**284–295.
34. **Levis, R., P. Bingham, and G. M. Rubin.** 1982. Physical map of the *white* locus of *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **79:**564–568.
35. **Lu, Q., L. L. Wallrath, H. Granok, and S. C. Elgin.** 1993. $(CT)_n$ $(GA)_n$ repeats and heat shock elements have distinct roles in chromatin structure and transcriptional activation of the Drosophila *hsp26* gene. Mol. Cell. Biol. **13:**2802–2814.
36. **Maniatis, T., E. F. Fritsch, and J. Sambrook.** 1982. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

37. **Manoukian, A. S., and H. M. Krause.** 1992. Concentration-dependent activities of the even-skipped protein in Drosophila embryos. Genes Dev. **6**:1740–1751.
38. **Martinez Arias, A., N. Baker, and P. Ingham.** 1988. Role of segment polarity genes in the definition and maintenance of cell states in the Drosophila embryo. Development **103**:157–170.
39. **Moazed, D., and P. H. O'Farrell.** 1992. Maintenance of the engrailed expression pattern by Polycomb group genes in Drosophila. Development **116**:805–810.
40. **Mortin, M. A., R. Zuerner, S. Berger, and B. J. Hamilton.** 1992. Mutations in the second-largest subunit of Drosophila RNA polymerase II interact with Ubx. Genetics **131**:895–903.
41. **Poole, S. J., L. M. Kauvar, B. Drees, and T. Kornberg.** 1985. The engrailed locus of Drosophila: structural analysis of an embryonic transcript. Cell **40**:37–43.
42. **Read, D., and T. B. Kornberg.** Unpublished data.
43. **Read, D., M. Levine, and J. L. Manley.** 1992. Ectopic expression of the Drosophila tramtrack gene results in multiple embryonic defects, including repression of *even-skipped* and *fushi tarazu*. Mech. Dev. **38**:183–195.
44. **Read, D., and J. L. Manley.** 1992. Alternatively spliced transcripts of the *Drosophila tramtrack* gene encode zinc finger proteins with distinct DNA binding specificities. EMBO J. **11**:1035–1044.
45. **Read, D., T. Nishigaki, and J. L. Manley.** 1990. The *Drosophila even-skipped* promoter is transcribed in a stage-specific manner in vitro and contains multiple, overlapping factor-binding sites. Mol. Cell. Biol. **10**:4334–4344.
46. **Rosenberg, A. H., B. N. Lade, D. S. Chui, S. W. Lin, J. J. Dunn, and F. W. Studier.** 1987. Vectors for selective expression of cloned DNAs by T7 RNA polymerase. Gene **56**:125–135.
47. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74**:5463–5467.
48. **Soeller, W. C., S. J. Poole, and T. B. Kornberg.** 1988. In vitro transcription of the Drosophila *engrailed* gene. Genes Dev. **2**:68–81.
49. **Studier, F. W., and B. A. Moffat.** 1986. Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. J. Mol. Biol. **189**:113–130.
50. **Tabata, T., S. E. Eaton, and T. B. Kornberg.** 1992. The Drosophila *hedgehog* gene is expressed specifically in posterior compartment cells and is a target of *engrailed* regulation. Genes Dev. **6**:2635–2645.
51. **Thummel, C.** 1989. The Drosophila *E74* promoter contains essential sequences downstream from the start site of transcription. Genes Dev. **3**:782–792.
52. **Topol, J., C. R. Dearolf, K. Prakash, and C. S. Parker.** 1991. Synthetic oligonucleotides recreate Drosophila *fushi tarazu* zebra-stripe expression. Genes Dev. **5**:855–867.
53. **Wood, W. I., J. Gitschier, L. A. Lasky, and R. M. Lawn.** 1985. Base composition-independent hybridization in tetramethylammonium chloride: a method for oligonucleotide screening of highly complex gene libraries. Proc. Natl. Acad. Sci. USA **82**:1585–1588.