Identification and sequence analysis of the 5' end of the major chicken vitellogenin gene

John B.E.Burch

Department of Genetics, Fred Hutchinson Cancer Research Center, 1124 Columbia Street, Seattle, WA 98104, USA

Received 27 September 1983; Revised and Accepted 6 December 1983

ABSTRACT

We have precisely determined the positions of the first three exons for the major chicken vitellogenin gene (VTG II) by a combination of S1 nuclease protection, primer extension and DNA sequencing experiments. In addition, we have determined the nucleotide sequences of the 5' flanking nuclease hypersensitive sites that we have previously shown are induced during the estrogen mediated activation of the VTG II gene in liver (1). One of these sites is found to be nearly identical to the enhancer core sequence of SV40. A computer assisted analysis of the DNA sequences upstream from the VTG II gene has revealed four short (7 to 9 base pair) sequence elements that are present in similar positions flanking the other major estrogen inducible gene for liver, very low density apolipoprotein II (apoVLDL II). For VTG II, these sequences are located between two of the induced nuclease hypersensitive sites that are liver specific. Sequences homologous to one element, located approximately 100 base pairs upstream from the mRNA cap sites of the VTG II and apoVLDL II genes, are also observed for three estrogen inducible genes that are expressed in the oviduct, although for each of these genes the sequence falls further upstream, between -220 and -200. We suggest that these conserved sequences may be important in mediating the tissue specific responses of these genes to estrogen.

INTRODUCTION

A fundamental problem of biology is the coordinate regulation of sets of genes during development. Of particular interest in this context are the sets of hormonally regulated genes which, in addition to being capable of responding to hormone, must be programmed to do so in a tissue specific way. A well documented case in point is the set of genes which code for the vitellogenin proteins in oviparous vertebrates (for review, see 2). In chickens, these genes are transcriptionally regulated by estrogen in liver cells and yet are totally unresponsive to the same primary stimulus in the oviduct, despite the presence of functional estrogen receptors which mediate the expression of the egg white genes in these cells. Conversely the egg white genes are not expressed in liver, with the exception of conalbumin which is expressed in both tissues (3).

In principal it is not unreasonable to suppose that the tissue specific transcriptional potential of such genes could be determined as a consequence of the genes being either in relaxed domains or buried within highly condensed regions of chromatin in a particular tissue. Based on an analysis of the chromatin structure, this would appear to account for how the major chicken vitellogenin gene (VTG II) is held silent in certain cell types such as fibroblasts and erythrocytes. In oviduct, however, the situation is more complex because in this tissue the chromatin structure of the VTG II domain has many features that are observed in the hormone naive liver and are more often associated with potentially active genes (1). This suggests that an additional level of control is required to account for the tissue specific regulation of this gene by hormone.

When the VTG II gene is transcriptionally activated in hormone naive liver cells by estrogen, a series of associated chromatin structural changes are observed in the 5' flanking region. The earliest event (which is roughly coincident with the appearance of detectable mRNA) is the generation of nuclease hypersensitive sites that are located approximately 0.00, 0.32, 0.76 and 0.94 kb upstream from the 5' end of the gene. Subsequent to this event, a single Msp I site within the VTG II gene domain (at -0.61 kb) becomes detectably undermethylated. Interestingly, in oviduct cells (but not other tissues) this Msp I site is also found undermethylated and is proximal to the 5' hypersensitive site which in liver is only observed during periods of hormone stimulation, suggesting that the VTG II gene domain may not only be "marked" in oviduct cells but also may actually "sense" the presence of hormone. However, in oviduct the two gene-proximal hypersensitive sites are not induced and the gene remains transcriptionally silent.

We would like to understand how these observations are functionally related to the tissue specific expression of this gene. In order to approach this question, we have carried out a sequence analysis, in conjunction with S1 nuclease mapping and primer extension experiments, to determine the organization of the 5' region of the VTG II gene. In addition, a comparison of this sequence with the 5' flanking sequence of the other major estrogen responsive liver gene, very low density apolipoprotein II (apoVLDL II), has revealed a set of homologies which we suggest may be relevant to the coordinate regulation of these two genes. For VTG II, this set of homologies are bounded by the pair of liver specific induced hypersensitive sites that are apparently necessary but not sufficient for transcription. The nucleotide sequences of these sites, as well as the other estrogen responsive

hypersensitive sites, are also presented.

MATERIALS AND METHODS

Construction and sequencing of M13 clones:

As a convenient source of DNA for subcloning into M13 vectors we initially constructed pBR322 derivatives which contained either the 1.1 kb or 0.5 kb EcoRI fragments from λVTG10 or the 3.6 kb Bam HI fragment of λVTG40 (1). These fragments were digested with a variety of restriction enzymes (see Fig. 1) and ligated into similarly digested M13mp 8/9 (4) or M13mp 10/11 vectors (J. Messing, University of Minnesota, unpublished; a gift from Dr. Richard Gelinas). After transforming 71/18 or JM101 cells, recombinants were picked from indicator plates (5) and single stranded recombinant phage was prepared (4) and sequenced using the dideoxy method of Sanger (6). As a primer for these sequencing reactions, we used a 17mer synthesized by Dr. Peter Seeburg and kindly made available by Dr. Richard Gelinas.

Preparation of single stranded probes:

Single stranded probes for nuclease S1 and primer extension experiments to map the RNA were obtained in basically two different ways. To prepare uniformly labelled probes we annealed a synthetic primer to M13 recombinant constructions which contained asymetric inserts (i.e., with different restriction sites at the two ends) and extended them in the presence of ^{32}P-dCTP and ^{32}P-TTP using Klenow fragment essentially as described above for sequencing, except that only deoxynucleotides were used. After 30 min. the unincorporated nucleotides were removed by passage through a 1cc syringe of Sephadex G50 fine (7). For each, the eluate was then digested with an appropriate restriction enzyme in order to introduce a cut at the downstream insert/vector junction. The resultant single stranded labelled fragment was isolated on a denaturing polyacrylamide gel (7).

Alternatively, double stranded fragments were either 5' end labelled by sequential treatment with phosphatase and kinase, or were 3' end labelled by filling in recessed 3' ends with Klenow fragment. In each case, the single stranded probes were obtained by electrophoresis in strand separation gels (7).

Mapping the RNA using S1 nuclease:

Liver polysomal RNA was prepared using the method of Palmiter (8) and 50 μg aliquots were hybridized to single stranded probes (1-5 X 10^4 CPM per reaction) in 30 μl hybridization buffer (10 mM Tris, 1 mM EDTA, 0.3M NaCl, pH 7.5) for 2 hrs at 65°C. After digesting the residual single stranded regions

for 15 min. at room temperature with 10.0 U/ml nuclease S1, the reaction was quenched and the protected fragments were precipitated with ethanol, resuspended in 99% formamide dye mix and electrophoresed on denaturing acrylamide gels (7).

Mapping the 5' ends of RNA by primer extension:

After coprecipitating 50 µg of liver polysomal RNA and 1-5 X 10^4 CPM single stranded primer with ethanol, the samples were resuspended in 20µl hybridization buffer (see above) and incubated at 65°C for 1 hr.. At this point 50 µl aliquots of reverse transcriptase cocktail were added, the reaction was allowed to proceed for 1 hr. at 37°C and then the samples were ethanol precipitated and resuspended in 99% formamide dye mix for analysis on a sequencing gel.

Sequence homology search:

In order to search for sequence homologies between the 5' flanking sequences of VTG II (this manuscript) and either apoVLDL II (9, 10), conalbumin (11), ovalbumin (12), lysozyme (13) or albumin (10), we took advantage of the two dimensional matrix analysis program of Pustell and Kafatos (14) kindly made available to us by Jim Wallace and Dr. Richard Gelinas. Typically we searched for 9 base sequences and had the computer print out all matches which received a minimum value of 70 (see 14). For the VTG II x apoVLDL II analysis we examined 750 and 475 base pairs, respectively, of 5' flanking sequences whereas only approximately 270 base pairs were examined for albumin and the egg white protein genes.

Precise mapping of 5' flanking VTG II nuclease hypersensitive sites:

The preparation of liver nuclei (from either egg laying hens or estradiol stimulated roosters), digestion *in situ* by endogenous nucleases and isolation of DNA were all as described previously (1). The same is true of the subsequent analysis by Southern blotting with two exceptions. First, agarose gels were electrophoresed in duplicate in order to yield two nitrocellulose filters; this allows one to examine the fragments looking from each end of the parent fragment by indirect end labelling. Secondly, instead of using λ x Hind III and ϕX x Hae III standards we prepared a mixture of partial (7) or complete restriction enzyme digests of the parent 3.6 kb Bam fragment (in cloned form from pVTG412) covering the 5' end of the VTG II gene. The particular restriction enzymes were chosen based on their proximity to the nuclease hypersensitive sites. Thus, instead of estimating the positions of the hypersensitive sites relative to the ends of the parental restriction fragments, we mapped these sites by interpolation

between closely spaced pairs of restriction sites.

In addition, we found it essential to include 10 μg of carrier DNA (Bam HI digested erythrocyte genomic DNA) along with the trace amounts of cloned DNA in order to avoid electrophoresis artifacts. Based on a comparison of λ x Hind III markers in the presence and absence of this amount of genomic DNA (which is a typical sample for Southern blotting genomic chicken DNA) it is clear that these artifacts can be rather severe (for example, as much as 0.2 kb in the case of the 4.45 kb marker). We estimate that our precision using this method is probably at least as good as +/- 15 base pairs (see Results).

RESULTS

I. Sequencing the 5' portion of the chicken VTG II gene.

The approximate position of the 5' end of the VTG II gene relative to restriction sites has independently been determined in two laboratories using the method of EM heteroduplex analysis (15, 16, 17). This position roughly coincides with the assignments that we made for the locations of the set of hypersensitive sites which are induced in response to hormone (1). Since both the putative 5' end of the gene and the induction specific hypersensitive sites map within two contiguous Eco RI fragments of 1.1 kb and 0.5 kb, we sequenced this entire region using M13 recombinants (4) in conjunction with the dideoxy sequencing procedure (6). After a preliminary restriction site analysis of this region by partial restriciton digestion of end labeled fragments (data not shown), we chose to use the restriction enzymes Eco RI, Pst I, Msp I, Sau3A and/or Hind III in order to generate sets of overlapping fragments which were cloned directly into M13 vectors and sequenced (Material and Methods). We were able to routinely resolve 200-300 nucleotides from each sequencing run and we obtained sequence data for both DNA strands. These sequence data are presented in Figures 1 and 3.

II. Mapping the start site for transcription.

We began our analysis by carrying out S1 nuclease protection experiments using subfragments from the 0.55 kb Eco RI fragment that we had sequenced (Fig. 1) since we anticipated that this fragment would encode the 5' end of the gene. Since the two fragments generated by digesting this Eco RI fragment at the single Hind III site were each cloned into M13 vectors for sequencing, we used these templates to obtain uniformly labeled single stranded probes which would be complementary to any mRNA transcribed from this region. These two probes (see cartoon for Fig. 2A) were independently hybridized to total liver polysomal RNA isolated from either laying hens

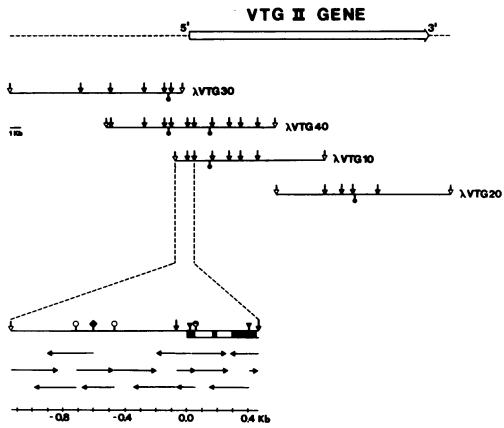

Figure 1: Cloning and sequencing strategy. The lambda clones which span the chicken VTG II gene were isolated from a genomic library as described previously (1). The approximate position of the 3' end of gene relative to these clones (which is indicated at the top of the figure) is based on EM heteroduplex analysis carried out in other laboratories (15,16) whereas the precise placements of the 5' end and the first three exons (depicted as solid boxes in the lower figure) are derived from the present analysis. The 1.6 kb of DNA corresponding to the two Eco RI fragments from the left end of the λVTG 10 insert was digested with various restriction enzymes and subcloned into M13 vectors (4). Fifteen such constructions were sequenced using the dideoxy method of Sanger (6) as indicated in the lower portion of the figure where the arrows indicate the direction and extend of sequence derived from each clone. The complete sequence obtained from this analysis is presented in Fig. 3. The relevant restriction sites have been marked using the symbols: , Eco RI; , artificial Eco RI linker; , Pst I; , Bam HI; , Msp I; , Sau 3A and , Hind III.

(lanes 3-4, 9-10), a 19 day embryo (lanes 5, 11) or a rooster (lanes 6, 12) and then digested with S1 nuclease. The protected fragments were electrophoresed on a denaturing polyacrylamide gel and exposed to x-ray film. This procedure revealed three protected DNA fragments when the experiment was carried out using mRNA from the livers of laying hens, where VTG II is expressed, but not when the mRNA derived from other sources (i.e., embryonic or rooster livers), where the gene is not expressed. The sizes of these fragments were determined to be approximately 58, 26, and 158 nucleotides, respectively. Since mild conditions were employed for the S1 nuclease digestion, we anticipated that these sizes might represent slight overestimates for the sizes of the first three exons of the gene (see below). As can be seen from Fig. 2A, two additional larger bands are also present but since they are observed in the absence of exogeneous mRNA (lanes 2, 8), we presume that they are artifacts and hence we shall not discuss them further.

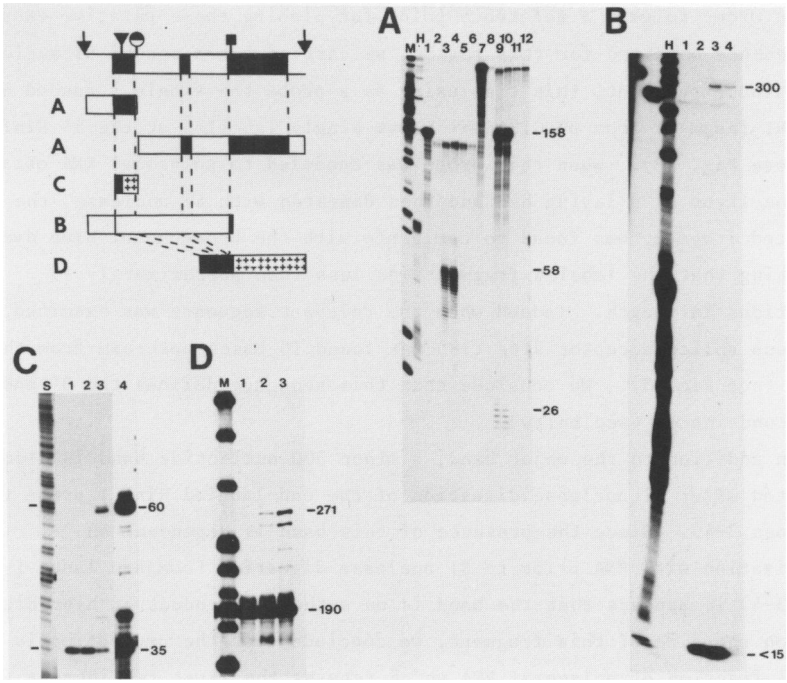

Figure 2: Mapping the start site of transcription and the first three exons for the chicken VTG II gene. The cartoon in the upper left hand corner summarizes the positions of the first three exons (solid boxes) and corresponding introns (open boxes) relative to restriction sites (▼ , Eco RI; ▼ , Sau 3A; ▼ , Hind III; and ▼ , Hinf I). The underlying figures are schematic representations of the data presented in panels A-D. Panel A: The uniformly labelled single stranded Hind III x Eco RI probes shown in lanes 1 and 7 were hybridized to 50 µg each of liver polysomal RNA obtained from laying hens (lanes 3-4, 9-10), 19 day embryos (lanes 5, 11) or adult roosters (lanes 6, 12) prior to S1 nuclease digestion and analysis on a sequencing gel. Controls in which no RNA was added are shown in lanes 2 and 8. The markers for this, and the other panels are: M, Msp I x pBR322 and H, Hinf I x pBR322. Panel B: The single stranded Hinf I x Eco RI probe was labelled at its 5' Hinf I site, hybridized to 50 µg polysomal RNA from an egg laying hen (lanes 3-4) or no RNA (lanes 1-2) at either 55° (lanes 1, 3) or 65° (lanes 2, 4), digested with S1 nuclease and analyzed on an 8% sequencing gel. Panel C: The 35 nucleotide Hind III x Sau 3A primer was hybridized to either no RNA (lane 1) or 50 µg liver polysomal RNA from either a rooster (lane 2) or a laying hen (lane 3) and extended using cold nucleotides and reverse transcriptase. A non-overexposed version of lane 3 is presented in lane 4. Lane "s" contains a dideoxy G sequencing tract of a totally unrelated template which is shown to provide a reference molecular weight ladder. Panel D: Primer extension analysis using the uniformly labelled 190 base Eco RI x Sau 3A single stranded fragment to prime 1, 3 and 10 µg (lanes 1-3 respectively) of hen liver polysomal RNA.

In order to gain a reference point for placing these putative exons on the sequence obtained for this region, we carried out a second S1 nuclease protection experiment, this time using as a probe the single stranded Hinf I x Eco RI fragment from pVTG126 which was singly labelled at the 5' Hinf I site (see Fig. 2B). When this probe was annealed to polysomal RNA obtained from the liver of a laying hen and then digested with S1 nuclease, the major protected fragment was found to comigrate with the bromophenol blue dye, indicating that the labeled fragment was less than approximately 15 nucleotides in length. Indeed when the relevant sequence was examined, a consensus splice acceptor site (18) was found 10 bases upstream from the Hinf I site (see Fig. 3). We conclude that this sequence defines the 3' end of the second intron (see below).

In addition to the major band, a minor 300 nucleotide band is also protected after S1 nuclease digestion of the end-labeled Hinf I probe (Fig. 2B, lanes 3-4). Since the presence of this band is dependent on hybridization with RNA prior to S1 nuclease digestion (compare lanes 1-2 with lanes 3-4) it appears that the band is an authentic product of hybridization. Based on the size of this fragment, we conclude that the protection is due to a small fraction of polysomal RNA which retains the first two introns (see below). Interestingly, there is no evidence of any message which retains only the second (and not the first) intron, perhaps suggesting an ordered processing for these two introns.

In order to convincingly argue that the exons we mapped by S1 nuclease protection are indeed at the 5' end of the message, we carried out primer extension analysis. Since we knew that the Hinf I site lay within an exon (see above) we began by preparing a uniformly labeled single stranded probe which terminated at its 5' end with an Eco R1 site and at its 3' end with the Hinf I site (see Fig. 2D). This primer was annealed to laying hen liver polysomal RNA and then extended to the terminus of the mRNA using AMV reverse transcriptase in the presence of cold nucleotides. As can be seen in Fig. 2D, the major band at 271 nucleotides is due to the primer having been extended by 81 nucleotides. Whereas no products are observed that are larger than this, several intermediate size bands are present which we presume are attributable to pause sites for the enzyme. This interpretation is supported by a second primer extension experiment which will be presented below (see Fig. 2C).

Given that the 3' end of the primer was defined by the Hinf I site, 7 nucleotides of the extended product could be accounted for by the exon within

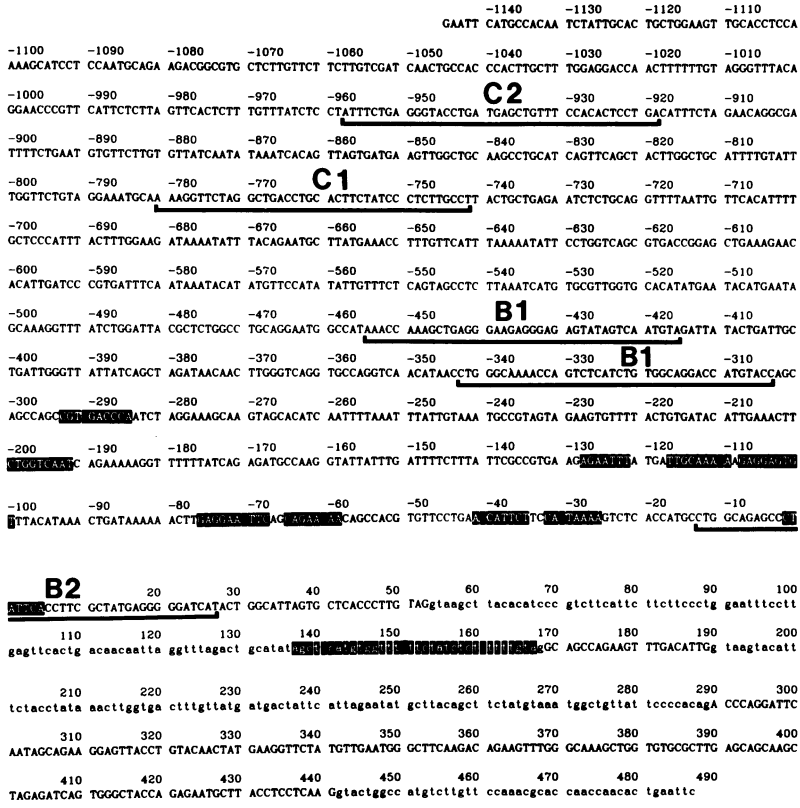

Figure 3: The nucleotide sequence for the 5' region of the chicken VTG II gene. The dideoxy method of Sanger (6) was used to sequence the M13 constructions (4) presented in Fig. 1. The sequence that was obtained is numbered here relative to the start site of transcription which is defined as position 1. The introns are denoted by lower case letters. Several sequences of interest have been highlighted for emphasis: TATA box homologies at positions -26 and -59; weak homologies to CAAT boxes at position -68 and -100; cap sequences at positions -35 (silent) and centered about position 1; four sequence elements that are shared between the 5' flanking regions of VTG II and apoVLDL II (CGTGACCCA, CTGGTCAAT, AGATTT and TTGCAAAA, respectively); and, within the first intron, two overlapping sequences that are each homologous to the consensus sequence of Mulvihill et al. (44). The nuclease hypersensitive regions that flank the 5' end of the gene are underlined (see Figure 4).

which the Hinf I site lies. Thus, the balance is 74 nucleotides of which we know at most 58 are accounted for by the S1 protected fragment derived from the smaller Eco RI x Hind III fragment shown in Fig. 2A since this fragment lies upstream. This suggested that the small exon observed in Fig. 2A (lanes

9-10) lies between the Hinf I and Hind III sites, and furthermore that the Hinf I site would thus lie within the large (approximately 158 base) exon.

Working within these guidelines we examined the sequence and found consensus splice junctions at each of the expected positions (see Fig. 3). In particular we found a donor splice site downstream from the Hinf I site which defines a 152 base pair (third) exon. Secondly, we found a 21 base pair (second) exon located between the Hind III and Hinf I sites which is flanked at both ends by consensus splice sites. Finally, just upstream of the Hind III site there is a splice donor site which in turn is preceeded by a consensus transcriptional start site 53 base pairs upstream (19), again in agreement with the S1 nuclease protection data. More importantly, the sum of the first two putative exons is exactly the same (74 nucleotides) as predicted from the primer extension analysis presented in Fig. 2D and yields a continuous open reading frame. Moreover, the reading frame is consistent with the fact that the protein is secreted (20) in that the initiation methionine residue is followed by an arginine and then a string of 14 consecutive non-polar amino acids.

As a definitive test of our assignment for the transcriptional start site, we made a 5' end labeled single stranded primer that we predicted would hybridize to mRNA within the first exon sequence and would be extended 25 bases to the 5' end of the message by reverse transcriptase. As shown in Fig. 2C, this is exactly what we observed. In particular, the 35 base primer labeled at the 5' Hind III end was extended exactly 25 bases beyond the 3' Sau3A end by reverse transcriptase in the presence of cold nucleotides. This start site maps exactly to the adenine residue within the consensus cap sequence (see Fig. 3). As can be seen from the overexposed autoradiogram in lane 3 of Fig. 2C, aside from weak bands which extend by 1, 2 and 3 bases beyond what we believe to be the true start site, there are no start sites which would inidicate initiation at positions further upstream from this cap site. As expected, the primer was not extended after annealing to rooster liver polysomal RNA (lane 2 of Fig. 3).

Finally, we note that our exon/intron assignments are in excellent agreement with previous EM heteroduplex analysis (15).

III. Mapping the positions for the 5' flanking nuclease hypersensitive sites.

In a previous publication (1) we showed that a set of nuclease hypersensitive sites is established in the region upstream from the VTG II gene when transcription is induced in the liver by estradiol treatment of either embryos or roosters. These same sites are also observed in the livers

of laying hens. Interestingly, once induced, two of the sites (denoted B1 and B2) are stable in that they persist after hormone withdrawal whereas the third site, which we are now able to resolve as a pair of sites (denoted C1 and C2; see below), is only present during periods of hormone mediated expression. As an initial step in trying to better understand the significance of these sites we undertook to map the sites relative to the DNA sequence that we had obtained for these regions.

Despite the inherent limitations of sizing DNA fragments by Southern blot analysis, we have been able to obtain quite precise data, as evidenced by the coincidence (to within 20 base pairs) for the map locations of the 5' flanking hypersensitive sites deduced using upstream and downstream probes. The analysis makes use of our knowledge of the exact locations (based on the DNA sequence) of restriction sites within the region of interest which provide a set of closely spaced standards. The strategy then was merely to locate the hypersensitive sites by determining the sizes of subbands based on interpolations between adjacent restriction sites.

The data that we obtained using this analysis are presented in Fig. 4. As indicated in the bottom portion of the figure, each of the sites are characterized by rather precise double stranded cuts. Several features of the sequences in the immediate vicinities of the hypersensitive sites deserve comment. First, it is clear that there is no correlation between the nuclease hypersensitive sites and AT-rich regions which might have been expected simply on the basis that such regions are thermodynamically less stable than GC-rich regions. Considering 40 base pairs of sequence centered about each of the sites defined in Fig. 4, we find that these regions are only 45-53% AT as opposed to even larger regions elsewhere which are greater than 80% AT rich but are not nuclease sensitive. A similar conclusion has been drawn based on fine mapping the nuclease hypersensitive region flanking the chicken adult β-globin gene (21).

Secondly, it has recently been demonstrated that many regions which are hypersensitive to nucleases <u>in situ</u> are also reactive to both nucleases (22) and single strand specific chemical probes (23) when present in the form of naked supercoiled DNA. Since inverted repeats (24,25) direct repeats (26,27) and Z DNA (28) sequences have all been shown to be capable of forming S1 nuclease sensitive structures in supercoiled DNA, we examined the regions defining the 5' flanking VTG II hypersensitive sites for such sequences. Although several of these kinds of sequence elements are present (for example multiple short direct repeats as well as a 7 base pair tract of alternating

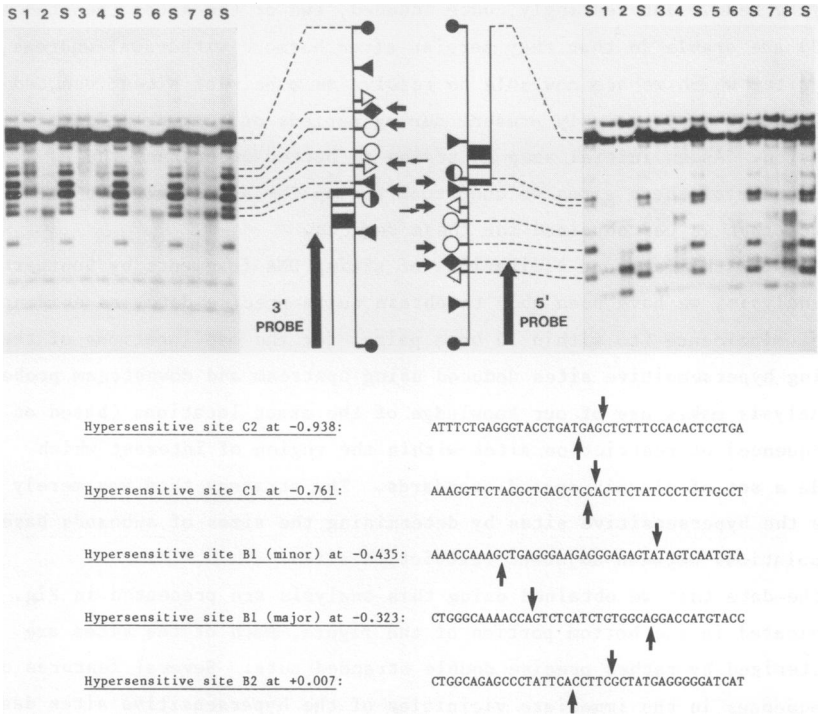

Figure 4: Precise determination of the estradiol induced VTG II 5' flanking nuclease hypersensitive sites. Hypersensitive sites within the 3.6 kb Bam HI fragment were mapped using the upstream and downstream probes as depicted in the top figure. In addition to the prominant subbands generated by in situ digestion at the hypersensitive sites (indicated by the horizontal arrows in the top figure), a set of weak bands are also observed in each lane which we have subsequently identified as being due to the use of a contaminated aliquot of Bam HI. Since these bands fall outside the bracketed region of interest, however, they represent only a cosmetic problem and in no way interfere with the analysis. For each filter, the pairs of in situ digested genomic samples (lanes 1-2, laying hen #1; lanes 3-4, laying hen #2; lanes 5-6, laying hen #3; lanes 7-8, estradiol treated rooster) are flanked by restriction site standards (see text) which are indicated by the following symbols: , AvaII; , Bam HI; , Eco RI; , Hind III; , Pst I and , XbaI. The precise positions of the 5' flanking nuclease hypersensitive sites are shown with respect to the DNA sequence in the lower portion of the figure. The horizontal arrows above and below each sequence indicate the positions determined from the 3' and 5' probes, respectively.

purines and pyrimidines), their occurence is not particularly striking when compared to the rest of the flanking DNA. Thus, whereas it may be that these features contribute to the structural discontinuities at these positions, they clearly do not account for why these particular sites are cut whereas

other flanking sequences are relatively nuclease resistant.

Thirdly, given that enhancers (reviewed in 29) in SV40, proviral LTRs and the immunoglobulin heavy chain gene have been found to contain nuclease hypersensitive sites (30, 31, and 32, respectively), it is perhaps worth noting that one of the induced hypersensitive sites (site C2) contains an 11 base pair region which is identical in the central 9 positions with the proposed enhancer core sequence of SV40 (29). When the search was broadened to include reasonable (i.e. 6 out of 8 base pair matches) homologies with the enhancer core consensus (which has been proposed based on a comparative analysis of many viral enhancer elements (33)), many additional elements were found flanking the VTG II gene. Whereas most of these are solitary, in the regions that correspond to the pair of B1 hypersensitive sites the homologies are clustered next to one another, as in the region defined as the immunoglobulin heavy chain gene enhancer (34).

IV. Comparison of the 5' flanking region of VTG II with other estrogen regulated genes.

As mentioned in the introduction, estradiol stimulation of embryos or roosters elicits the expression in liver cells of a set of genes that are normally induced in the livers of laying hens in response to the same stimulus. Since there is no evidence that the genes which comprise the liver specific set are evolutionally related to one another we decided to compare the flanking sequences of these genes as a means of identifying sequence components (other than the familiar CAAT and TATA homologies) that might be involved in mediating their coordinate expression.

VTG II is now the second member, along with apoVLDL II (9, 10), of the estrogen responsive liver specific genes for which the 5' end has been assigned and sequenced. Since approximately 500 base pairs of 5' proximal flanking sequence is available for apoVLDL II, we compared this with a similar region for VTG II using the two dimensional matrix program of Pustell and Kafatos (14).

This analysis revealed four similarities between the upstream flanking regions of these two genes which are summarized in Fig. 5. Immediately 5' to the apoVLDL II CAAT box the sequences AGAATTT and TTGCAAAA are found separated from another by two nucleotides. These same two sequences (separated in this case by four nucleotides) are also present in a similar position for VTG II. The other two elements identified by this analysis were found to be homologous to the CAAT consensus sequence (12), although the elements are located approximately 200 and 290 base pairs upstream from the

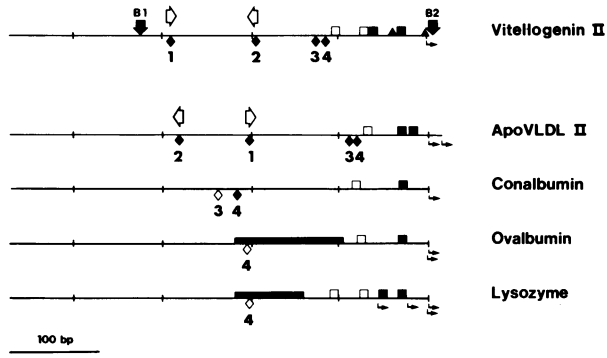

Figure 5: Regions of sequence homology are present upstream from the mRNA cap sites for VTG II and other estrogen responsive chicken genes. The 5' flanking sequences of VTG II (Fig. 3) were compared with similar regions for apoVLDL II (9,10) as well as conalbumin (11), ovalbumin (12), lysozyme (13) and albumin (10) using a computer matrix analysis program (14). The results of this analysis are represented schematically in this figure along with the positions of the consensus promoter elements. The genes are aligned at the right with respect to their mRNA cap sites. As noted in the text, VTG II contains two sets of consensus promoter elements (i.e., cap and TATA sequences, with weak homologies to the CAAT consensus, represented by the symbols ▲, ■ and □, respectively) although the upstream cap sequence is transcriptionally silent. In addition to these familiar sequences, VTG II and apoVLDL II share additional regions of homology. The solid diamonds indicate the subset of four such sequences (elements 1-4 are CGTGACCCA, CTGGTCAAT, AGAATTT and TTGCAAAA, respectively) that are located in similar positions for the two genes. The open horizontal arrows indicate that a core of element 2 (TGGTC) is the inverted repeat of a sequence present in element 1. Note also that element 2 is highly homologous with the CAAT consensus (12). As depicted in the lower portion of the figure, homologies to element 4 are also present in the region between 220 and 200 base pairs upstream from the mRNA cap sites for each of the three egg white genes examined, although in some cases (open diamonds) the homologies are not exact. Conalbumin also contains a sequence (GGAATTT) nearly identical to element 3 which is located 14 base pairs upstream from its copy of element 4. None of these sequences are found flanking the chicken albumin gene. The solid vertical arrows represent the two estrogen induced liver specific hypersensitive sites (sites B1 and B2) that flank this set of conserved sequence elements for VTG II. The solid horizontal bars for ovalbumin and lysozyme represent the hormone responsive regions which have recently been identified for these genes (quoted in 43).

mRNA cap site of each of the genes instead of the usual 70-80 base pairs. Element 2 is, in fact, a better CAAT homology than either of the two "CAAT sequences" found in the expected positions closer to the VTG II gene. In addition, although the positions of the elements are very similar for the two genes, their order is reversed; i.e. element #1 (CGTGACCCA) is the extreme 5' element for VTG II whereas element #2 (CTGGTCAAT) is the extreme 5' element

for apoVLDL II. Note that the CAAT homology for element 1 reads away from the gene (i.e. the two elements constitute an imperfect indirect repeat).

When the analysis was extended to include comparisons between approximately 300 base pairs of VTG II flanking DNA and similar regions from each of three different chicken egg white protein genes (i.e. ovalbumin, lysozyme and conalbumin), a perfect copy of element 4 (TTGCAAAA) was identified at position -216 of conalbumin. Upon closer examination, related sequences were also found at similar positions for both lysozyme and ovalbumin; TTGCAACA at -202 and TTAGCAGAA at -205, respectively. Curiously, whereas element 3 (AGAATTT) is located close to element 4 in both VTG II and apoVLDL II, the only egg white gene which possesses a similar upstream sequence is conalbumin, albeit the element (GGAATTT) in this case is not perfect and is separated by 14 base pairs instead of only 2 or 4. Although other different regions of homology were found between VTG II and other genes examined, including the serum albumin gene (10), they were neither in similar positions nor were shared between sets of coordinately regulated genes.

DISCUSSION

The sequence organization of the 5' portion of the major vitellogenin gene (VTG II) has several interesting aspects that are worth noting. First, in common with several other estrogen regulated genes such as lysozyme (13), ovomucoid (35,36) and apoVLDL II (9, 10), VTG II is found to contain two overlapping sets of sequences which are each homologous to consensus promoter elements (see Fig. 3). Whereas in the other cases examined both promoters are functional (albeit to varying extents), as evidenced by the presence of RNA species with the appropriate 5' ends, this is not true of VTG II despite the fact that the redundant promoter elements are highly homologous to one another. Our failure to detect any evidence of transcriptional initiation from the upstream promoter would appear to be very neatly accounted for, however, by the guanine residue (instead of a thymine) in position 3 of the respective TATA box as this exact transversion was shown to have a drastic effect on the transcriptional efficiency of the chicken conalbumin gene (37, 38).

The conservation of these tandom homologous promoter elements suggests that they may be advantageous despite the fact that for VTG II they are not in themselves competent with respect to transcription initiation. Considering the proximity of these sequences to the functional set it seems reasonable to suggest, for example, that they may serve instead to localize

putative transcriptional factors in the vicinity of the true promotor and in this way contribute favorably to more efficient transcription in vivo.

In analyzing the DNA sequence downstream from the mRNA cap site, it seems reasonable to presume that the ATG sequence at position 14 serves as the translation initiation codon since this is the first such triplet encountered (39). This supposition is in all likelihood correct as evidenced by the fact that this assignment yields a continuous open reading frame for the region we have sequenced thus far (the first three exons, which code for approximately 71 amino acids). Moreover, the predicted N-terminus for this protein is compatible with the fact that VTG II is a secreted protein (20). In particular, the initiator methionine is followed by an arginine and then a block of 14 contiguous non-polar amino acids.

Assuming that translation does begin at this position, it is evident that the untranslated leader sequence for this message is quite short. Whereas shorter leaders have been documented (40), the size in this case is particularly interesting in light of the fact that the half life for the VTG II message (but not liver messages in general) is known to vary greatly in the presence and absence of hormone (41, 42). The size of the leader sequence would suggest that this property is probably encoded elsewhere. Clearly these predictions can be tested experimentally.

As mentioned in the Introduction, VTG II is one member of a small set of genes which are synthesized exclusively in liver in response to estrogen stimulation. Since the gene which codes for the other major estrogen inducible species, apoVLDL II, has recently been sequenced (9, 10), we made a comparison between the 5' flanking sequences of VTG II and apoVLDL II to identify common sequences which might be regulatory elements involved in the coordinate expression of these two genes.

A computer assisted search (14) revealed a number of sequences that are found within 500 base pairs upstream from the mRNA cap sites for both VTG II and apoVLDL II. Particularly intriguing are the subset of four such sequences that are present in similar positions relative to the mRNA cap sites for the two genes (Fig. 5). Two of these elements are located approximately 200 and 290 base pairs upstream from the mRNA cap sites and by nucleotide sequence are clearly related to the CAAT consensus (12), although in the case of element 1 the sequence reads away from the gene. The other pair of elements (separated by only 2-4 base pairs) is located just upstream from the set of consensus promoter elements for each of the two genes (Fig. 5). This pair of sequences is also observed flanking the conalbumin gene,

whose transcription is regulated by estrogen in both liver and oviduct (3). Weaker homologies for the element closest to the gene (but not the adjacent element) were also found for the two other egg white genes examined (ovalbumin and lysozyme). For each of these three genes which are expressed in oviduct, however, the sequence is present further upstream than for VTG II and apoVLDL II, i.e. approximately 220 to 200 base pairs in front of the respective mRNA cap sites. Interestingly, these positions coincide with the 5' boundaries for the hormone responsive regions recently determined for both lysozyme and ovalbumin (quoted in 43).

We also searched the VTG II gene region for the 19 base pair consensus originally presented by Mulvihill et al (44) as a potential progesterone receptor DNA binding site. Although more recent experiments have led to a refinement of this hypothesis (45), the large number of homologous copies around the various egg white genes (44) as well as copies near the gene for apoVLDL II (10) remains an intriguing observation. Whereas the function of this sequence remains obscure, it is worth noting that VTG II (which is not responsive to progesterone) contains two overlapping copies of this sequence at the 3' end of the first intron (see Fig. 3).

We are currently in the process of determining whether the conserved sequence elements that we have identified for VTG II and apoVLDL II are of functional importance in vivo. In addition, we would like to better understand how the tissue specific induction of the 5' flanking nuclease hypersensitive sites which bracket these homologies is related to the differential expression of VTG II. The mapping and sequence determinations that we have presented in this manuscript have allowed us to focus more closely on these sites although we still are unable to rationalize the precise digestion pattern that we observe. However, the fact that the nuclease hypersensitive site located 938 base pairs upstream from the mRNA cap site is identical in 9 out of 11 positions with the core SV40 enhancer sequence (29) suggests a function for at least one of the previously identified estrogen induced hypersensitive sites which we are currently testing experimentally.

ACKNOWLEDGEMENTS

I would like to thank Pei Feng Cheng and Brian Keegan for their expert technical assistance, Helen Devitt for preparing the manuscript, Jim Wallace for making the computer analysis possible and Drs. Steve McKnight, Mark Groudine, Colin Casimir and Harold Weintraub for critically reading the

manuscript. The generous gift of synthetic M13 primer from Drs. Richard Gelinas and Peter Seeburg is also gratefully acknowledged. This work was supported by a National Institutes of Health postdoctoral fellowship (GM08337). Additional funds were provided by National Institutes of Health grants awarded to Harold Weintraub, in whose lab this work was carried out.

REFERENCES

1. Burch, J.B.E. and Weintraub, H. (1983) Cell 33, 65-76.
2. Wahli, W., Dawid, I.B., Ryffel, G.U. and Weber, R. (1981) Science 212, 298-304.
3. McKnight, G.S., Lee, D.C. and Palmiter, R.D. (1980) J. Biol. Chem. 255, 148-153.
4. Messing, J. and Vieira, J. (1982) Gene 19, 269-276.
5. Gronenborn, B. and Messing, J. (1978) Nature (London) 272, 375-377.
6. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463-5467.
7. Maniatis, T., Fritsch, E.F., and Sambrook, J. (1982) Molecular cloning: a laboratory manual (Cold Spring Harbor, New York: Cold Spring Harbor Laboratory).
8. Palmiter, R.D. (1974) Biochemistry 13, 3606-3615.
9. van het Schip, A.D., Meijlink, F.C.P.W., Strijker, R., Gruber, M., van Vliet, A.J., van de Klundert, J.A.M. and AB, G. (1983) Nucleic Acids Res. 11, 2529-2540.
10. Hache, R.J.G., Wiskocil, R., Vasa, M., Roy, R.N., Lau, P.C.K. and Deeley, R.G. (1983) J. Biol. Chem. 258, 4556-4564.
11. Cochet, M., Gannon, F., Hen, R., Maroteaux, L., Perrin, F., and Chambon,P. (1979) Nature (London) 282, 567-579.
12. Benoist, C., O'Hare, K., Breathnach, R., and Chambon, P. (1980) Nucleic Acids Res. 8, 127-142.
13. Grez, M., Land, H., Giesecke, K., Schutz, G., Jung, A., and Sippel, A.E. (1981) Cell 25, 743-752.
14. Pustell, J. and Kafatos, F.C. (1982) Nucleic Acids Res. 10, 4765-4782.
15. Arnberg, A.C., Meijlink, F.C.P.W., Mulder, J., vanBruggen, E.F.J., Gruber, M., and AB, G. (1981) Nucleic Acids Res. 9, 3271-3286.
16. Wilks, A.J., Cozens, P.J., Mattaj, I.W. and Jost, J.-P. (1981) Gene 16, 249-259.
17. Wilks, A.J., Cozens, P.J., Mattaj, I.W., and Jost, J.-P. (1982) Proc. Nat. Acad. Sci. USA 79, 4252-4255.
18. Breathnach, R. and Chambon, P. (1981) Ann. Rev. Biochem. 50, 349-383.
19. Busslinger, M., Portman, R., Irminger, J.C., and Birnstiel, M.L. (1980) Nucleic Acids Res. 8, 957-977.
20. Kreil, G. (1981) Ann. Rev. Biochem. 50, 317-348.
21. McGhee, J.D., Wood, W.I., Dolan, M., Engel, J.D., and Felsenfeld, G. (1981) Cell 27, 45-56.
22. Larsen, A., and Weintraub, H. (1982) Cell 29, 609-622.
23. Kowhi-Shigematsu, T., Gelinas, R., and Weintraub, H. (1983) Proc. Natl. Acad. Sci. USA 80, 4389-4393.
24. Lilley, D.M.J. (1980) Proc. Natl. Acad. Sci. USA 77, 6468-6472.
25. Panayotatos, N., and Wells, R.D. (1981) Nature (London) 289, 466-470.
26. Hentschel, C.C. (1982) Nature (London) 295, 714-716.
27. Glikin, G.C., Gargiulo, G., Rena-Descalzi, L., and Worcel, A. (1983) Nature (London) 303, 770-774.
28. Singleton, C.K., Klysik, J., Stirdivant, S.M., and Wells, R.D. (1982)

Nature (London) 299, 312-316.
29. Khoury, G., and Gruss, P. (1983) Cell 33, 313-314.
30. Cremisi,C. (1981) Nucleic Acids Res. 9, 5949-5964.
31. Groudine, M., Eisenman, R. and Weintraub, H. (1981) Nature (London) 292, 311-317.
32. Storb, U., Arp, B. and Wilson, R. (1981) Nature (London) 294, 90-92.
33. Weiher, H., Konig, M., and Gruss, P. (1983) Science 219, 626-631.
34. Gillies, S.D., Morrison, S.L., Oi, V.T., and Tonegawa, S. (1983) Cell 33, 717-728.
35. Lai, E.C., Roop, D.R., Tsai, M.J., Woo, S.L.C., and O'Malley, B.W. (1982) Nucleic Acids Res. 10, 5553-5567.
36. Gerlinger, P., Krust, A., LeMeur, M., Perrin, F., Cochet, M., Gannon, F., Dupret, D., and Chambon, P. (1982) J. Mol. Biol. 162, 345-364.
37. Wasylyk, B., Derbyshire, R., Guy, A., Molko, D., Roget, A., Teoule, R., and Chambon, P. (1980) Proc. Natl. Acad. Sci. USA 77, 70245-7028.
38. Grosschedl, R., Wasylyk, B., Chambon, P., and Birnstiel, M.L. (1981) Nature (London) 294, 178-180.
39. Kozak, M. (1981) Curr. Top. Microbiol. Immun. 93, 643-646.
40. Kelly, D.W., Coleclough, C., and Perry, R.P. (1982) Cell 29, 681-689.
41. Wiskocil, R., Bensky, P., Dower, W., Goldberger, R.F., Gordon, J.I., and Deeley, R.G. (1980) Proc. Natl. Acad. Sci. USA 77, 4474-4478.
42. Brock, M.L., and Shapiro, D.J. (1983) Cell 34, 207-214.
43. Parker, M. (1983) Nature (London) 304, 687-688.
44. Mulvihill, E.R., LePennec, J.-P., and Chambon, P. (1982) Cell 28, 621-632.
45. Davison, B.L., Mulvihill, E.R., Egly, J.M., and Chambon, P. (1983) CSHSQB 47, 965-976.

Note added in proof: Two other laboratories (Geiser, et al, J. Biol. Chem. 258, 9024-9030; Walker, et al, EMBO J. 2, 2271-2279) recently described the sequence organization of the 5' end of the chicken vitellogenin gene in agreement with the data presented here.