Volume 17 Number 16 1989 Nucleic Acids Research

Rat gene 33: analysis of its structure, messenger RNA and basal promoter activity

Nancy B.Chrapkiewicz, Charles M.Davis, David T.-W.Chu, Catherine M.Caldwell and Daryl K.Granner*

Department of Molecular Physiology and Biophysics, Vanderbilt University Medical School, Nashville, TN 37232, USA

Received May 8, 1989; Revised and Accepted July 21, 1989 EMBL accession nos X07266, X07267

ABSTRACT

Several overlapping cDNA and genomic DNA clones corresponding to mRNA^{g33} and gene 33, respectively, were isolated and characterized. The mRNA^{g33} molecule is 2971 nt in length, exclusive of a poly(A+) tail, and encodes a putative 459 amino acid protein (49,946 daltons). The 13.2 kbp gene consists of four exons, three introns, and has two initiation sites located 27 and 30 bp downstream from a TATA box.

Transfection of H4IIE cells with a fusion gene 33 (-1900 to +32)/luciferase construct (pSL330A) gave rise to readily detectable luciferase activity. In addition, primer extension analysis of the gene 33/luciferase mRNA transcribed in these experiments showed that transcription initiates in the gene 33 DNA at two sites consistent with those found *in vivo*. Analysis of the effect of 5′ deletions on basal promoter activity showed that, in relation to the promoter activity of pSL330A, ~40% of the activity is lost between -1643 and -1050, another ~40% between -550 and -475, and deletion to -55 causes a total loss of detectable luciferase activity.

INTRODUCTION

In 1985 Lee et al. reported the isolation of a cDNA, designated p33, from a hydrocortisone-induced rat liver cDNA library (1). This cDNA corresponds to the mRNA transcribed from gene 33 (mRNA^{g33}), which encodes a protein of unknown function. Studies of the hormonal regulation of mRNA^{g33} in rat liver indicated that both the amount of mRNA^{g33} and its rate of synthesis are increased by insulin, hydrocortisone and cAMP (1). Moreover, detailed studies in hepatoma cells both in our laboratory and others have shown that insulin, dexamethasone, and phorbol esters increase the amount of mRNA^{g33} approximately 10-fold, 20-fold, and 20-fold, respectively (2-6), and that the rate of transcription of gene 33 is also increased by these three effectors (4-6).

Our particular interest in gene 33 is as a complement to previous studies in which insulin has been shown to inhibit the transcriptional activity of the phosphoenolpyruvate carboxykinase (PEPCK) gene in hepatoma cells (7,8). We recently examined the concurrent effect of insulin on gene 33 and PEPCK gene transcriptional activity in H4IIE cells. The results showed that the two genes were reciprocally regulated simultaneously at physiologic concentrations

of insulin (4). This system therefore provides a very powerful tool for the study of insulin action as well as examination of how a hormone can simultaneously alter transcription of two genes in opposite directions in the same cell.

Further study of the mechanism of insulin action requires a detailed analysis of the gene 33 and the PEPCK gene promoters. Examination of the PEPCK promoter for an Insulin Regulatory Element(s) (IRE) has already begun and, although the IRE is not well defined as yet, a fusion gene containing 600 bp of 5′ flanking sequence from the PEPCK gene has been shown to respond to insulin in the same manner as the endogenous gene (9). In order to identify the DNA sequences that mediate the transcriptional control of gene 33, and to learn how they interact with various factors to regulate transcription, the structure of the gene and its promoter must first be defined. In this paper we present data on the primary structures of gene 33, its mRNA and protein, the sequence of 1383 bp of 5′ flanking DNA, and analysis of the basal promoter activity of 1900 bp of 5′ flanking gene 33 DNA in fusion gene constructions.

MATERIALS AND METHODS

Materials

Restriction enzymes were obtained from International Biotechnologies, Inc., Promega Corp., New England Biolabs, and Bethesda Research Labs. AMV reverse transcriptase was purchased from Life Sciences, Inc., Klenow fragment came from Promega Corp. and deoxynucleotides and dideoxynucleotides from P-L Biochemicals. S1-nuclease and Luciferin were purchased from Boehringer-Mannheim. Radioisotopes were obtained from Amersham Corporation or New England Nuclear.

General Methods

DNA fragments used as probes were radioactively labelled by random oligonucleotide priming (10), using an Amersham Multiprime DNA labelling system. Bacteriophage DNA preparations, alkaline lysis plasmid preparations, and 5′ end-labelling of oligonucleotides were all performed by standard procedures (11). Total cell RNA was obtained from H4IIE cells by the method of Chirgwin et al. (12) and the poly(A^+) fraction was isolated as previously described (13).

Libraries/Screening/Subcloning

The partial cDNA p33 used to screen libraries was generously provided by Dr. F. Kenney, University of Tennessee at Oak Ridge (1). Two λgt11 rat liver cDNA libraries were screened for clones corresponding to $mRNA^{g33}$ (11). One of these libraries was obtained from Dr. M. Mueckler, Washington University Medical Center (14), and the other was purchased from Clontech Laboratories,

Inc. The designations p2M, p5M, p7M (isolated from the Mueckler library) and p14C (isolated from the Clontech library) refer to the cDNA inserts of plasmids p2M.9E1, p5M.9E17, p7M.9E18 and p14C.9E2, respectively (see Figure 1).

Gene 33 clones were obtained from three different rat genomic DNA libraries; these included an EMBL3 library (15) obtained from Dr. G. Scherer (German Cancer Research Center, Heidelberg, Germany), a Charon 4A library (16), and an unamplified rat genomic DNA library constructed in our laboratory following the methods described by Frischauf et al. (17). The designations λS7, λB1, and λD1 refer to the bacteriophage clones containing gene 33 inserts. The bacterial plasmids pλB1.8E15, pλB1.9E14, pλB1.9E13, and pλS7.BS11 (shown in Figure 3A) are the subclones of λB1 and λS7 which were used to characterize gene 33 by S1 nuclease mapping, primer-extension analysis, and sequencing.

DNA Sequence Analysis

Single stranded DNA obtained from the pEMBL or Bluescript subclones by superinfection with helper phage M13K07 (18) was sequenced by the Sanger method (19) using custom oligonucleotide primers and by sequencing from restriction enzyme sites using universal primers. All primers were synthesized on a model 380A Applied Biosystems DNA synthesizer.

S1-Nuclease Mapping

Definition of exon boundaries by S1-nuclease protection was performed as previously described by Beale et al. (20). The S1-mapping of the initiation site of the gene was done essentially as described by Green et al. (21). The 5′ end-labelled oligonucleotide P33.P16 was annealed to single-stranded DNA of plasmid pλB1.9E14 and second strand synthesis was carried out using Klenow fragment. After a 15 min synthesis at 37° C the sample was extracted with phenol/chloroform and the aqueous fraction was divided between three tubes containing 10 µg of tRNA, 10 µg poly(A^+) RNA from untreated cells, or 10 µg poly(A^+) RNA from insulin-treated cells. The rest of the procedure employed was as described by Berk and Sharp (22) except that the S1-nuclease concentration was 750 units/ml and the products were electrophoresed on 8% polyacrylamide gels containing 7 M urea.

Primer-Extension Analysis

To determine the initiation site of gene 33, the primer P33.P16 (see Figure 2) was 5′ end-labelled and hybridized to 10 µg of mRNA by incubation at 95° C for 5 min then 65° C for 3 h in a buffer containing 50 mM Tris at pH 8.0, 250 mM NaCl, and 1 mM EDTA. The buffer was then adjusted to 35 mM Tris at pH 8.3, 150 mM NaCl, 10 mM DTT, 10 mM $MgCl_2$, and 0.5 mM dNTPs, 10 units of AMV reverse transcriptase was added, and synthesis was carried out for 60 min

at 42° C. To determine the site of mRNA initiation in the gene 33/luciferase fusion gene pSL330A (see Figure 6), a 30 base primer called PL1 (5′-GCCTTTCTTTATGTTTTTGGCGTCTTCCAT-3′) was synthesized. PL1 is complementary to a luciferase sequence in the fusion mRNA such that its 5′ end anneals 92 bases from the expected 5′ end of the mRNA (based upon the determination of the predominant initiation site for gene 33 in Figure 4). The 5′ end-labelled PL1 was hybridized to 50 µg of total RNA isolated from H4IIE cells transfected with pSL330A; prior to hybridization the RNA was treated twice with DNase I to remove any residual plasmid DNA. Hybridization was carried out by incubation at 65° C for 1 h in 20 mM Tris at pH 7.5, 250 mM NaCl, and 1 mM EDTA. The buffer was then adjusted to 50 mM Tris at pH 7.5, 33 mM NaCl, 10 mM DTT, 3 mM $MgCl_2$, 0.5 mM dNTPs, 2 units of AMV reverse transcriptase was added, and synthesis was carried out at 37° C for 60 min. After synthesis the reactions were ethanol precipitated and electrophoresed on 6-8% polyacrylamide gels containing 7 M urea.

Plasmid Constructions

The plasmids pSV0AL (containing the luciferase insert L) and pSV232AL-AΔ5′ (containing the luciferase insert L-AΔ5′) were obtained from Jeffrey DeWet and colleagues at the University of California at San Diego (23). Since the luciferase insert L-AΔ5′ is the preferred reporter gene, L-AΔ5′ was removed from pSV232AL-AΔ5′ along with the SV40 intron and polyadenylation sequence by digestion with *Hind*III and *Bam*HI and inserted in place of the corresponding sequences in pSV0AL. Also, the original *Bam*HI and *Hind*III sites 5′ of luciferase were destroyed and new restriction enzyme sites were added such that the resulting plasmid, pSV232PL2L contains the polylinker *Kpn*I/*Sac*I/*Sma*I/*Hind*III/*Bgl*II 5′ of L-AΔ5′.

Plasmid pSL330A was constructed by insertion of the -1900 to +32 *Bgl*II/*Bgl*II fragment of gene 33 into the *Bgl*II site in the polylinker of pSV232PL2L. The 5′ deletion mutants pSL330A-1a, pSL330A-3a, pSL330A-4a, pSL330A-5b, and pSL330A-5c, were generated by digestion of pSL330A with *Sac*I and *Hind*III, then digestion with Exonuclease III followed by religation using a Promega Corp. Erase-a-Base system. pSL331A was generated by digestion of pSL330A with *Sma*I then religation, and pSL332A by digestion of pSL330A with *Sma*I and *Sal*I followed by fill-in of the *Sal*I site with Klenow and then religation. All these plasmids are diagrammed in Figure 7.

Transfection

H4IIE cells were co-transfected with 25 µg of a gene 33/luciferase plasmid and 25 µg of RSVCAT as described by Quinn et al. (24). Cells were harvested 18 h after DMSO shock, resuspended in 250 mM Tris at pH 7.8 containing 1 mM PMSF, sonicated, and cellular debris pelleted by centrifugation at 12,000 rpm

Figure 1. Restriction enzyme map and sequencing strategy of cDNA^{g33} clones. The cDNA designated p33, given to us by Dr. F. Kenney, is shown along with the cDNA inserts of clones p14C.9E2, p2M.9E1, p5M.9E17, and p7M.9E18 (see Materials and Methods). The sequencing strategy employed is depicted by the arrows beneath the clones; the length of the arrow represents the number of bases sequenced. Arrows pointing to the right indicate sense strand sequence, arrows pointing to the left indicate anti-sense strand sequence. The triangles mark the locations of the splice sites. The restriction enzyme map was confirmed by the completed cDNA^{g33} sequence. Abbreviations used are: A, *Ava*I; AII, *Ava*II; B, *Bgl*II; C, *Cla*I; H, *Hind*III; P, *Pst*I; PII, *Pvu*II; X, *Xba*I.

for 15 min. The supernatant was then used to assay for CAT activity by the method of Nordeen et al. (25) and luciferase activity by the method of DeWet et al. (23) using an AMINCO photometer coupled to a chart recorder.

RESULTS

Isolation of cDNA clones corresponding to mRNA^{g33}

Using the partial cDNA designated p33 (Figure 1) as the probe, two cDNA libraries were screened for clones containing mRNA^{g33} sequences. Two clones containing inserts longer than p33 were initially obtained and subcloned into pEMBL19. The plasmid p14C.9E2 contains a 2150 bp cDNA (p14C) whereas plasmid p2M.9E1 contains a 2800 bp cDNA (p2M). A comparison of the DNA fragments generated by digesting p2M, p14C and p33 with various combinations of the restriction endonucleases *Pst*I, *Xba*I, *Ava*II, and *Hind*III, showed that the 3′ 1000 bp of all three had identical restriction maps, as did the adjacent (5′ direction) 1000 bp of p14C and p2M (Figure 1).

To confirm that p14C and p2M contain sequences complementary to mRNA^{g33}, RNA blot transfers of poly(A)$^+$ RNA isolated from untreated cells, or insulin-treated cells were probed with p33 or subfragments of p14C and p2M. Probes from all three cDNAs hybridized to an mRNA ~3000 nt long, and in all cases mRNA^{g33} was several times more abundant in mRNA isolated from the insulin-

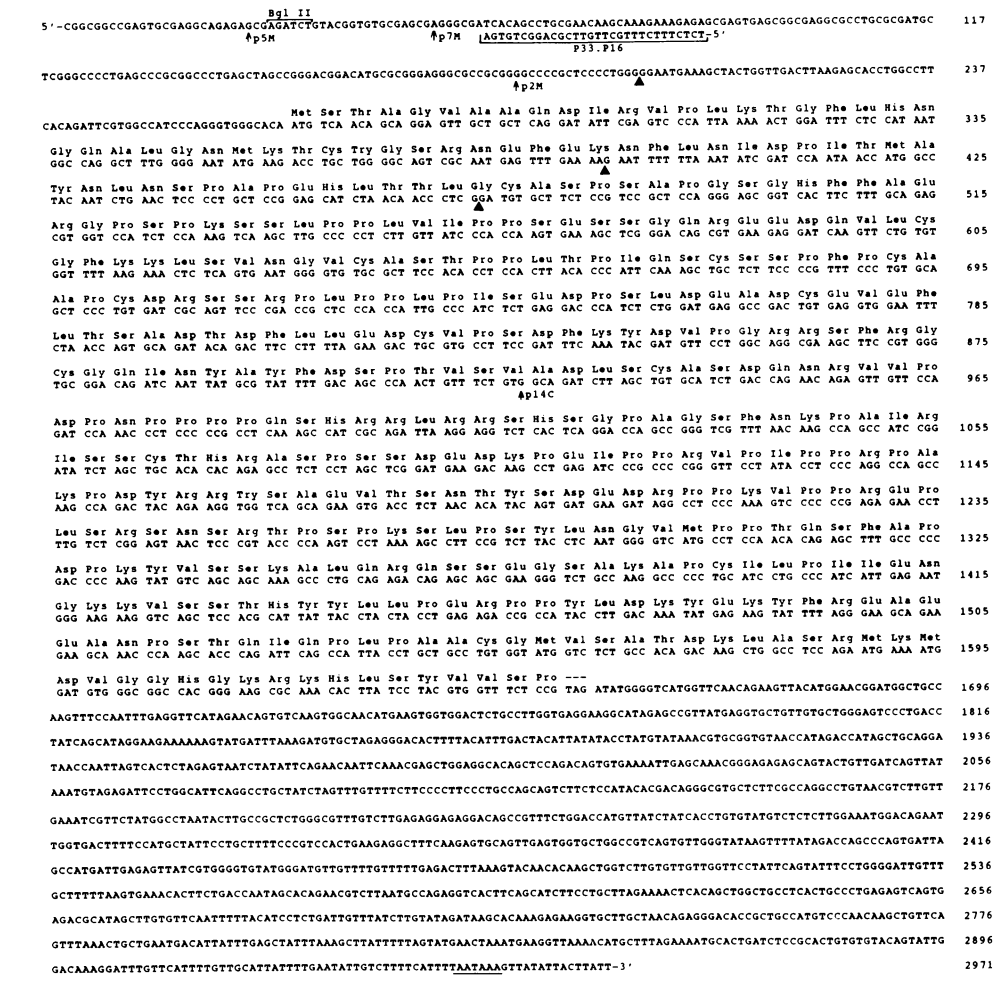

Figure 2. The cDNA sequence homologous to mRNA^{g33} and the deduced amino acid sequence of protein 33. All but 24 nt of the sequence homologous to mRNA^{g33} was determined by sequencing p2M, p5M, p7M, and p14C, as shown in Figure 1. The 5′ ends of these four cDNA clones are marked by the arrows. The last 24 bases at the 5′ end were obtained from genomic DNA sequence, based on the demonstrated transcription initiation site (see Figure 4). The splice sites are marked with triangles. The location and sequence of the synthetic primer P33.P16 (used to determine the transcription initiation site) and the 5′-most *Bgl*II site are bracketed. The polyadenylation site, which starts at position 2951, is underlined, and the deduced 459 amino acid sequence, which starts at nucleotide 270, is shown above the codons.

treated cells (data not shown). These results showing the regulation of the mRNA and its size are in agreement with previously published reports on mRNA^{g33} (1,2).

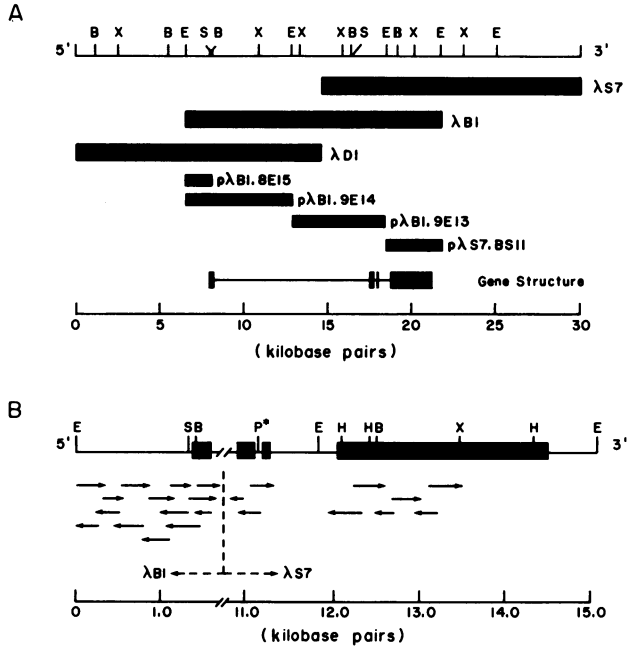

Figure 3. Characterization of gene 33. The inserts of three genomic clones λS7, λB1, and λD1 are shown relative to one another below a partial restriction enzyme map of this portion of genomic DNA in Panel A. The inserts of plasmid subclones pλB1.8E15, pλB1.9E14, pλB1.9E13, and pλS7.BS11 are also depicted. A scale representation of the exon/intron structure of gene 33 is illustrated below the subclones. The four exons are represented by the solid bars, the introns by the connecting lines. In Panel B a 15 kbp segment of the genomic DNA is expanded to show the exon/intron structure in more detail. This stretch of DNA contains 1.4 kbp of 5′ flanking DNA, the 13.2 kbp gene, and 0.4 kbp of 3′ flanking DNA. The four exons are again represented by the solid bars, and the introns by the connecting lines between the bars. The break in the scale (in intron 1) represents about 9 kbp of DNA. The sequencing strategy is shown by the arrows; the length of the arrow represents the number of bases sequenced. Arrows pointing to the right indicate sense strand sequence, arrows pointing to the left indicate anti-sense strand sequence. The sequence indicated to the left of the dashed line was obtained from λB1 subclones while that to the right of the dashed line was obtained from λS7 subclones. Abbreviations used are: B, *Bgl*II; E, *Eco*RI; P, *Pst*I (* not all sites shown); S, *Sal*I; X, *Xba*I.

A primer-extension analysis, using poly(A$^+$)RNA isolated from insulin-treated cells indicated that p2M lacked ~200 bp of mRNA^{g33} sequence (data not shown). Consequently, the cDNA libraries were rescreened using the 5′-most 370 bp of p2M as the probe. Six clones were obtained and their cDNA inserts were subcloned for characterization. One, designated p7M, contains an additional 140 bp of mRNA^{g33} sequence. Another of the clones, p5M, contains an additional 27 bp beyond the 5′ end of p7M (see Figure 1).

cDNA^{g33} Sequence

The sequence of all but the 5′-most 24 bases of mRNA^{g33} was determined by sequencing p14C, p2M, p5M, and p7M as shown in Figure 1. Over 96% of the cDNA^{g33} sequence (including 100% of the sequence in the presumptive coding region) was determined on both strands of cDNA. The sequence of the 5′-most 24 nt of mRNA^{g33}, not contained in any of our cDNAs, was deduced from genomic clones of gene 33. The addition of this sequence relied upon the determination of the start site of transcription (see below). The entire cDNA^{g33} sequence representative of mRNA^{g33} is depicted in Figure 2. The location of the 5′ ends of the cDNA molecules are indicated by arrows.

Deduced Amino Acid Sequence of Protein 33

Analysis of the cDNA^{g33} sequence showed that there was only one large open reading frame which would code for a protein of approximately 53 kDa, the size expected for protein 33 based on the hybrid-selected *in vitro* translation of mRNA^{g33} reported by Lee et al. (1). This open reading frame is 1377 nt long, and the decoded 459 amino acid sequence, shown in Figure 2, would encode a protein of 49,946 daltons.

Cloning of Gene 33 DNA

Having obtained a set of cDNAs representing almost the entire mRNA^{g33} molecule, we screened three genomic DNA libraries for clones containing gene 33 DNA. Three clones were obtained as shown in Figure 3A. Clone λS7 was isolated from an EMBL3 genomic library (15), using both p33 and the 5′-most 800 bp of p2M as hybridization probes. Analysis of blot transfers of restriction enzyme-digested λS7 DNA, probed with several subfragments of p2M, provided the orientation of the genomic clone as well as the information that there were approximately 2 kbp of DNA in which the restriction enzyme map of λS7 exactly matched the 3′ end of p2M and p14C [compare the *Hind*III, *Bgl*II, and *Xba*I restriction enzyme sites between 12 and 14.5 kbp, illustrated in Figure 3B, with these same restriction enzyme sites as shown in Figure 1]. Thus λS7 contains the 3′ region of gene 33. Clone λB1 was obtained from a Charon 4A library (16) using p7M as the hybridization probe. This clone contains an additional 8 kb of sequence extending in the 5′ direction, including a *Bgl*II site not present in λS7. This *Bgl*II site, located 1411 bp from the 5′ end of the clone, corresponds to the *Bgl*II site at the 5′ end of the cDNA p5M. Lastly, clone λD1 was obtained from the rat genomic cDNA library constructed in our laboratory. It contains an additional 6 kbp of 5′ flanking sequence.

Identification of the Transcription Initiation Site of Gene 33

In order to determine whether λB1 contained the transcription initiation site, S1-nuclease mapping, primer-extension analysis, and DNA sequencing were

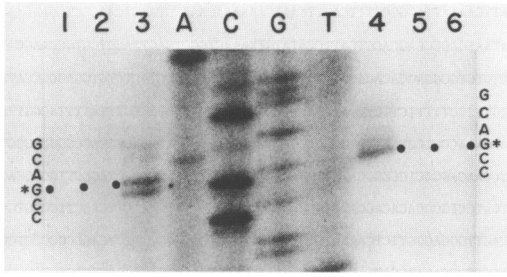

Figure 4. Determination of the transcription initiation site of gene 33. Sequencing reactions, primer-extension analysis, and S1-nuclease protection analysis were all performed using oligonucleotide primer P33.P16 so that the results could be directly compared. The primer-extension analysis, performed as described in Materials and Methods, is shown in lanes 1-3. The templates for reverse transcriptase synthesis were: lane 1, 10 µg tRNA; lane 2, 10 µg poly(A^+) RNA from untreated H4IIE cells; lane 3, 10 µg poly(A^+) RNA from H4IIE cells treated for 90 min with insulin. The next four lanes show the sequence ladder generated using P33.P16 to prime dideoxynucleotide sequencing of single stranded DNA of plasmid pλB1.9E14. The S1-nuclease protection analysis is shown in lanes 4-6. Using 5′ end-labelled P33.P16 as a primer, a highly radioactive single stranded DNA was synthesized, annealled to RNA, and digested with S1-nuclease as described in Materials and Methods. The RNAs used were: lane 4, 5 µg poly(A^+) RNA from H4IIE cells treated for 1 hour with insulin; lane 5, 5 µg poly(A^+) RNA from untreated H4IIE cells; lane 6, 5 µg tRNA. The major site of initiation is indicated by the asterisk.

all performed from a single primer, P33.P16, which is complementary to a sequence at the 5′ end of p7M (see Figure 2). The results are shown in Figure 4. The primer-extension reaction yielded four products in the form of two doublets, a prominent one of 85/86 nt and another, less intense, of 88/89 nt (lane 3). All of these products were more abundant when mRNA isolated from insulin-treated cells was used (compare lanes 3 and 2), and were completely absent in the tRNA control (lane 1). The doublet products produced by this analysis indicate that gene 33 has two transcription start sites located 86 and 89 nt from the 5′ end of P33.P16 (26) and it appears that the former is used preferentially. The S1-nuclease protection assay supports this conclusion. The most prominent S1-nuclease-protected band is the same size as the 86 nt primer-extended product (see lanes 4-6). A comparison with the DNA sequence showed that the 86 nt primer-extended product corresponds to a guanine residue on the anti-sense strand. We have assigned this base as position +1 (see Figure 5).

Characterization of Gene 33

Having located the transcription initiation site of gene 33, we determined the sequence of both strands of the 5′-most 1582 bp of λB1 DNA (Figures 3B and 5). This includes 1383 bases of 5′ flanking DNA, the first exon, and 6 bases

Figure 5. Partial sequence of gene 33 and 1383 bases of 5′ flanking DNA.
The sequence of 1383 bp of 5′ flanking DNA is shown along with the first 205 bp of gene 33 sequence. The 5′ flanking DNA and exon sequences are designated in uppercase letters and the intron sequences are designated in lower case letters. The transcriptional promoter elements found between positions -100 and -1 are boxed. These include a TATA box, two potential SP1 binding sites, and a reverse CAAT/NF1 binding site. Sequences with similarity to described regulatory elements are underlined and numbered (1, CRE; 2, GRE; 3, AP1 binding site).

of the first intron. Additional genomic DNA sequence was obtained from λS7 DNA (see Fig. 3B) using primers complementary to the cDNA sequence. Also, the 3′ end of gene 33 was analyzed by the S1-nuclease protection method (data not shown). This data enabled us to verify the cDNA sequence in the protein coding region and to determine the exon/intron structure of gene 33. The gene is composed of four exons of 197, 194, 77, and 2503 bp separated by three introns of ~9200 bp, 101 bp, and ~850 bp in the 5′ to 3′ orientation (Figure 3B). The exon/intron boundaries were located by comparing the genomic and cDNA sequences at points of DNA sequence divergence, and by determining the best fit with splice site consensus sequences (27).

Comparison of the genomic DNA sequence with the first 2168 bp of the cDNA sequence showed one discrepancy, which was located in the 5′ non-coding mRNA sequence. The dinucleotide GA, found at position +87/88 in the genomic sequence, is not present in either p5M, p7M, or p64M (another cDNA). These two bases are marked with asterisks in Figure 5. Since all three cDNAs lack these two bases, we presume that the difference between the two sequences is

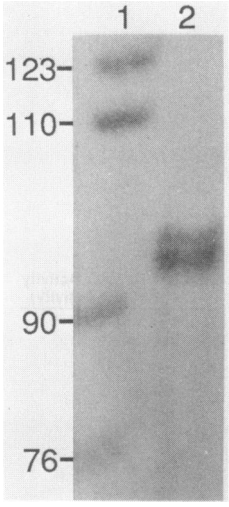

Figure 6. Determination of the Initiation Site of the Gene 33/Luciferase Fusion mRNA. Two T-150 flasks of H4IIE cells were transfected with 50 µg of pSL330A and then replated. Four hours later total cellular RNA was isolated, treated twice with DNase I, and primer-extended using oligonucleotide PL1 as described in Materials and Methods. Lane 1 contains end-labelled *Msp*I digested pBR322 as size markers. The primer-extended products are in lane 2.

due to microheterogeneity in gene 33 rather than to copying errors by reverse transcriptase, therefore the two bases were not added to the cDNA representation of the mRNA^{g33} sequence in Figure 2. Consequently the length of the first exon is 197 bp based on the cDNA sequence but is 2 bases longer in the genomic sequence (Figure 5).

Analysis of the Gene 33 Promoter

Analysis of the promoter activity of gene 33 was initiated by construction of the gene 33/luciferase plasmid pSL330A (see Materials and Methods and Figure 7). Transient transfection of H4IIE cells with pSL330A resulted in readily detectable, DNA concentration-dependent, expression of functional luciferase (data not shown). To determine whether, at the mRNA level, the luciferase activity was the result of transcription from the gene 33 initiation site designated +1 in Figure 5, or from some other site or sites, primer-extension analysis was performed using the oligonucleotide PL1. This oligonucleotide would be expected to give primer-extended products 92 and 95 bp in size if the mRNA synthesis begins in the gene 33 sequences of pSL330A at the two sites of gene 33 initiation *in vivo*. As can be seen in Figure 6, the fusion mRNA clearly initiates at two sites consistent in location (size) with the *in vivo* initiation sites of gene 33. Thus in plasmid pSL330A, luciferase is under control of the gene 33 promoter.

To further delineate the role of the 1900 bp of 5′ flanking DNA in pSL330A on the strength of the basal promoter activity, successive 5′ deletion of the gene 33 DNA was performed. The resulting plasmids, shown in Figure 7, were then assayed for their relative promoter activity by co-transfection with

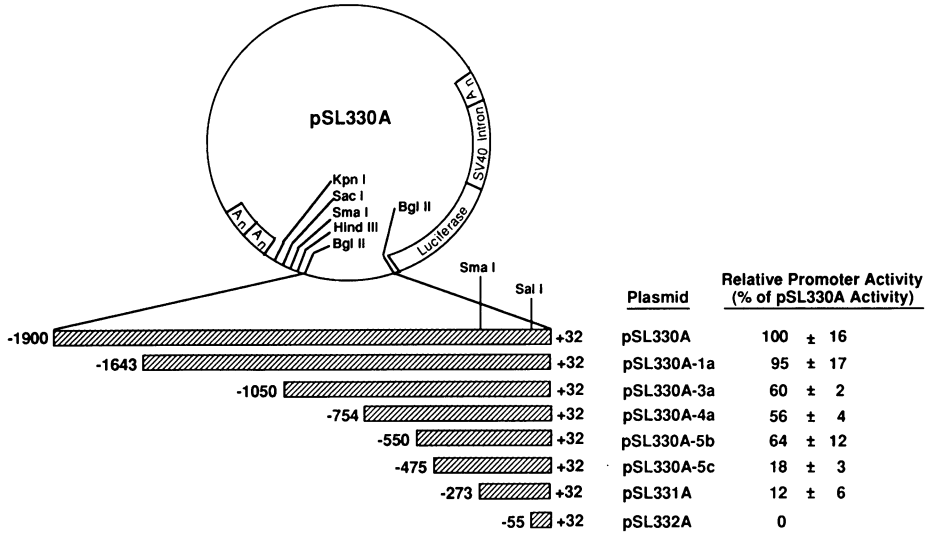

Figure 7. 5′ Deletion Analysis of the Gene 33 Basal Promoter Activity. A diagram of pSL330A, which contains gene 33 DNA from -1900 to +32 (see Materials and Methods), is shown along with a representation of the 5′ flanking DNA of the 7 deletion mutants pSL330A-1a, pSL330A-3a, pSL330A-4a, pSL330A-5b, pSL330A-5c, pSL331A, and pSL332A. The relative promoter activities were determined by co-transfecting H4IIE cells with each of the plasmids plus RSVCAT and then normalizing the resulting luciferase activities with the CAT activities. The data shown is the combined result of two experiments in which each plasmid was assayed in duplicate. The promoter activities are expressed as the percentage of the activity of pSL330A ± half of the range.

RSVCAT to normalize the luciferase activities, and then comparison of the luciferase activities relative to the pSL330A value. The results in Figure 7, which represent the combined data of two experiments, show that the promoter activity is reduced approximately 40% by deletion of the sequences between -1643 and -1050, reduced another 40% by deletion of the sequences between -550 and -475, and is undetectable if the sequences are deleted to -55.

DISCUSSION

Since the report by Lee et al. in 1985 in which p33 was first described (1), a substantial number of studies on the regulation of gene 33 have been published. It is now well documented that the transcriptional activity of gene 33 is increased by glucocorticoids, phorbol esters, and insulin in hepatoma cells (4-6) and that gene 33 is developmentally regulated in rat liver (28,29). Despite the interest in this gene, however, no progress has been made toward identifying the function of gene 33 and its protein product. As

an initial step in this direction, the putative 459 amino acid sequence decoded from the mRNA^{g33} sequence (shown in Figure 2) was examined. Apart from the observation that the 14% proline content is three times that found in the average protein (64 residues total, and four in a row at one point) and the tryptophan content is low (0.4% vs. the average of 1.4%) there is nothing particularly notable about the composition of this protein (30). Several computer analyses showed that the protein does not appear to contain a large hydrophobic region or signal peptide which might suggest a secreted or membrane bound protein, and there does not appear to be any significant clustering of positively charged amino acids which might indicate a function such as DNA binding. There is, however, one region of negative charge spanning 31 amino acids of unknown significance. Lastly, comparison of the amino acid sequence with protein sequences in the National Biomedical Research Foundation protein data base and NEW Database bank revealed no similarity to any known protein. Thus the identity of the gene 33 protein remains unknown.

Based upon the cDNA^{g33} sequence, and the placement of the initiation site of transcription, the mRNA^{g33} molecule was determined to be 2971 nt in length exclusive of the poly(A+) tail. The molecule can be divided into three segments, a 269 nt 5′ non-coding region which is quite large in comparison to most eukaryotic mRNAs (31), the 1377 nt protein coding region, and a 1325 nt 3′ non-coding region (see Figure 2). This cDNA^{g33} sequence was compared with data in the GenBank but no similarities to known molecules were found. Gene 33 was determined to be ~13.2 kbp in size with a structure of four exons (197, 194, 77, and 2503 bp) separated by three introns (~9200, 101, and ~850 bp). While the exon/ intron structures of class II genes are quite diverse, ranging from those without introns to those with dozens, this gene is rather unusual in that the ~9200 bp first intron contains 70% of the total DNA in the gene, and the 2503 bp fourth exon contains 85% of the total mRNA sequence. In contrast, there is also a small intron (101 bp) and a small exon (77 bp). These structures of gene 33 and its mRNA are in substantial agreement with those described previously by Tindall et al. (32). Although there are slight differences in the size of the first and fourth exons (and therefore mRNA^{g33}), and in two base designations in the 5′ flanking sequence, it is obvious that we have both identified the same gene.

The promoters of genes transcribed by RNA polymerase II contain one or more of several basic transcriptional elements within the first 150 bp of 5′ flanking sequence. Notable are the TATA box, which precisely positions the start site of transcription (27), the CAAT/NF1 binding site, which interacts with CAAT transcription factor (CTF) or CAAT binding protein (CBP) to affect the efficiency of transcription (33-35), and transcription factor SP1 binding

sites, which can influence the efficiency of transcription but can also function in the absence of a TATA box to initiate transcription, usually at multiple sites (36). Different combinations of these elements form a functional transcription unit. The sequence TATAAAT, which matches the consensus TATA box both in sequence and position (27), occurs at position -30 to -24 in gene 33. Slightly further 5′ there are two potential SP1 binding sites which overlap one another. The consensus recognition sequence for SP1 binding is (G/T)GGGCGG(G/A)(G/A)(C/T) with the six underlined bases forming a more conserved core (36). The sequence GGGGCGGCGG at position -68 to -59 has an 80% similarity to this total sequence and a 6/6 base match with the core, while the second site at position -73 to -64 (GAGGCGGGGC) has a 90% overall similarity and 5/6 base match with the core. Finally, a reverse complement CAAT box sequence, ATTGG, occurs at position -88 to -84. The locations of these four elements are marked in Figure 5.

Transfection of H4IIE cells with pSL330A showed that 1900 bp of 5′ flanking gene 33 DNA is functional as a promoter and confirmed that the initiation site we determined for gene 33 *in vivo* was correct. Further analysis of the basal promoter activity by successive 5′ deletion of the pSL330A gene 33 sequences provided two interesting observations. First, there are two regions of sequence which, even though they do not contain any of the three common promoter elements described above, enhance the promoter activity. These are located between -1643 and -1050, and between -550 and -475. Deletion of each of these regions resulted in a 40% loss in basal promoter activity relative to the activity expressed by pSL330A. Second, deletion of the sequences from -273 to -55, which removes both the CAAT box and the two putative SP1 binding sites, reduces the promoter activity to below detectable levels. Therefore, the DNA sequence between -273 and -55 is critical to the efficient promoter activity of gene 33 (presumably due to one or more of the CAAT/SP1 elements), and there are at least two regions of sequence upstream of the CAAT box which appear to enhance the basal promoter activity of gene 33.

As already noted, the transcription of gene 33 is regulated by cAMP (1), glucocorticoids (1,5), phorbol esters (6) and insulin (1,4-6). In most cases these molecules mediate their effects through specific DNA elements. Cyclic AMP regulatory elements, which have the consensus sequence TGACGTCA, are usually found within 200 bp of the initiation site of responsive genes (37). The possible CRE sequence CGACGTCA (a 7/8 nt match to the consensus) is located at position -54 to -47 in gene 33. Glucocorticoid regulatory elements have the consensus sequence GGTACANNNTGTTCT (38). In genes regulated by glucocorticoids in liver these elements generally are located further from the

transcription initiation site than CREs, have more sequence variability, and are usually duplicated (38-40). Two potential glucocorticoid elements, located at positions -749 to -735 (CGTCCCCTTTGTCCT; 8/12) and -714 to -700 (TCACCTTGCTGTTCT; 7/12), are present in the gene 33 sequence. These sequence elements are not high percentage matches, however, there are examples of GREs with even lower similarity to the consensus sequence which are still functional, the PEPCK gene being a case in point (40). Recent studies of the 5′ flanking DNA of genes regulated by phorbol esters have shown that some of them, including the human metallothionein II_A and collagenase genes, contain an enhancer element that binds the transcription factor AP-1 (41,42). The binding sites for AP-1 have been shown to center around the sequence TGACTCA (42), and there is growing evidence to suggest that AP-1 is involved in mediating the inducibility of these genes in response to phorbol esters (41,42). The best similarity to a potential AP-1 binding site in gene 33 is the sequence TGGCTCA located at position -1246 to -1240. Although this is located much further 5′ than seen in collagenase, stromelysin, and metallothionein II_A (41-43), there is evidence that this element can function at distances of over 2 kbp in fusion gene constructions (41). Experiments are currently in progress to determine whether any of these putative regulatory elements (underlined in Figure 5) are functional in controlling expression of the gene 33 transcription unit.

To date there are few well-defined systems available in which to study the mechanism by which insulin regulates the rate of gene transcription. Attempts to identify insulin response elements have only been reported for PEPCK (9), c-fos (44), amylase (45,46), growth hormone (47), and glyceraldehyde-3-phosphate dehydrogenase (48). The analysis of gene 33 and its basal promoter activity reported in this paper now enables us to pursue the definition of the insulin regulatory element of gene 33. It will allow us to analyze a positive-acting insulin response element, and given the observation that transcription of gene 33 and phosphoenolpyruvate carboxykinase are regulated in a reciprocal manner by insulin (4), will allow us to analyze differential regulation of gene expression by this hormone in a single cell type.

ACKNOWLEDGEMENTS: We thank K. Lee and F. Kenney for generously supplying the cDNA p33, Deborah Caplenor for secretarial help, and Tony Weil, Patrick Quinn, Richard O'Brien, and Mark Magnuson for critically revieving the manuscript. This research was supported by NIH grant DK 35107 and by the Vanderbilt Diabetes Research and Training Center (DK 07061). CMD is a trainee in the Vanderbilt Medical Scientist Training Program (T32-GM 07347).

ABBREVIATIONS: PEPCK, cytosolic phosphoenolpyruvate carboxykinase (GTP) (EC 4.1.1.32); $mRNA^{g33}$, messenger RNA coding for protein 33; $cDNA^{g33}$, cloned DNA corresponding to $mRNA^{g33}$; kbp, kilobase pairs; bp, base pair; nt, nucleotides;

p2M.9E1, p14C.9E2, etc., bacterial plasmids that contain cDNA^{g33}; p2M, p14C, etc., cDNA^{g33} inserts; λS7, λB1, etc., bacteriophage containing DNA of gene 33; pλS7.8E15, etc., bacterial plasmids containing gene 33 DNA.

*To whom correspondence should be addressed at: 607 Light Hall, Vanderbilt University Medical School, Nashville, TN 37232, USA

REFERENCES

1. Lee, K., Isham, K.R., Stringfellow, L., Rothrock, R., and Kenney, F.T. (1985) J. Biol. Chem. **260**, 16433-16438.
2. Messina, J.L., Hamlin, J., and Larner, J. (1985) J. Biol. Chem. **260**, 16418-16423.
3. Messina, J.L., Hamlin, J., and Larner, J. (1987) Endocrinology **121**, 1227-1232.
4. Chu, D.T., Davis, C.M., Chrapkiewicz, N.B., and Granner, D.K. (1988) J. Biol. Chem. **263**, 13007-13011.
5. Messina, J.L. (1989) Endocrinology **124**, 754-761.
6. Weinstock, R.S., and Messina, J.L. (1988) Endocrinology **123**, 366-372.
7. Granner, D., Andreone, T., Sasaki, K., and Beale, E. (1983) Nature **305**, 549-551.
8. Sasaki, K., Cripe, T.P., Koch, S.R., Andreone, T.L., Petersen, D.D., Beale, E.G., and Granner, D.K. (1984) J. Biol. Chem. **259**, 15242-15251.
9. Magnuson, M.A., Quinn, P.G., and Granner, D.K. (1987) J. Biol. Chem. **262**, 14917-14920.
10. Feinberg, A.P., and Vogelstein, B. (1983) Analytical Biochem. **132**, 6-13.
11. Maniatis, T., Fritsch, E.F., and Sambrook, J. (1982) Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor.
12. Chirgwin, J., Przybyla, A., MacDonald, R., and Rutter, W. (1979) Biochemistry **18**, 5294-5299.
13. Krystosek, A., Cawthon, M., and Kabat, D. (1975) J. Biol. Chem. **250**, 6077-6084.
14. Mueckler, M.M., and Pitot, H.C. (1985) J. Biol. Chem. **260**, 12993-12997.
15. Shinomiya, T., Scherer, G., Schmid, W., Zentgraf, H., and Schutz, G. (1984) Proc. Natl. Acad. Sci. USA **81**, 1346-1350.
16. Sargent, T.D., Wu, J., Sala-Trepat, J.M., Wallace, R.B., Reyes, A.A., and Bonner, J. (1979) Proc. Natl. Acad. Sci. USA **76**, 3256-3260.
17. Frischauf, A., Lehrach, H., Poustka, A., and Murray, N. (1983) J. Mol. Biol. **170**, 827-842.
18. Dente, L., Cesareni, G., and Cortese, R. (1983) Nucleic Acids Res. **11**, 1645-1655.
19. Sanger, F., Nicklen, S., and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA **74**, 5463-5467.
20. Beale, E.G., Chrapkiewicz, N.B., Scoble, H.A., Metz, R.J., Quick, D.P., Noble, R.L., Donelson, J.E., Biemann, K., and Granner, D.K. (1985) J. Biol. Chem. **260**, 10748-10760.
21. Green, S., Walter, P., Kumar, V., Krust, A., Bornert, J., Argos, P., and Chambon, P. (1986) Nature **320**, 134-139.
22. Berk, A., and Sharp, P. (1977) Cell **12**, 721-732.
23. DeWet, J.R., Wood, K.V., DeLuca, M., Helinski, D.R., and Subramani, S. (1987) Mol. Cell. Biol. **7**, 725-737.
24. Quinn, P.G., Wong, T., Magnuson, M.A., Shabb, J.B., and Granner, D.K. (1988) Mol. Cell. Biol. **8**, 3467-3475.
25. Nordeen, S.K., Green, P.P., and Fowlkes, D.M. (1987) DNA **6**, 173-178.
26. Shelness, G.S., and Williams, D.L. (1984) J. Biol. Chem. **259**, 9929-9935.

27. Breathnach, R., and Chambon, P. (1981) Ann. Rev. Biochem. 50, 349-383.
28. Rothrock, R., Lee, K., Isham, K.R., Johnson, A.C., and Kenney, F.T. (1987) Biochem. Biophys. Res. Comm. 144, 1182-1187.
29. Johnson, A.C., Lee, K., Isham, K.R., and Kenney, F.T. (1988) J. Cell. Biochem. 37, 243-253.
30. Doolittle, R.F. (1986) Of URFS and ORFS: A Primer on How to Analyze Derived Amino Acid Sequences. University Science Books, Mill Valley, California.
31. Kozak, M. (1984) Nucleic Acids Res. 12, 857-872.
32. Tindal, M.H., Lee, K., Isham, K.R., Cadilla, C., and Kenney, F.T. (1988) Gene 71, 413-420.
33. Jones, K.A., Kadonaga, J.T., Rosenfeld, P.J., Kelly, T.J., and Tjian, R. (1987) Cell 48, 79-89.
34. Graves, B.J., Johnson, P.F., and McKnight, S.L. (1986) Cell 44, 565-576.
35. McKnight, S., and Tjian, R. (1986) Cell 46, 795-805.
36. Kadonaga, J.T., Jones, K.A., and Tjian, R. (1986) Trends in Biochem. 11, 20-23.
37. Montminy, M.R., Sevarino, K.A., Wagner, J.A., Mandel, G., and Goodman, R.H. (1986) Proc. Natl. Acad. Sci. USA 83, 6682-6686.
38. Jantzen, H., Strahle, U., Gloss, B., Stewart, F., Schmid, W., Boshart, M., Miksicek, R., and Schutz, G (1987) Cell 49, 29-38.
39. Danesch, U., Gloss, B., Schmid, W., Schutz, G., Schule, R., and Renkawitz, R. (1987) EMBO J. 6, 625-630.
40. Petersen, D.D., Magnuson, M.A., and Granner, D.K. (1988) Mol. Cell. Biol. 8, 96-104.
41. Angel, P., Imagawa, M., Chiu, R., Stein, B., Imbra, R.J., Rahmsdorf, H.J., Jonat, C., Herrlich, P., and Karin, M. (1987) Cell 49, 729-739.
42. Lee, W., Mitchell, P., and Tjian, R. (1987) Cell 49, 741-752.
43. Matrisian, L.M., Leroy, P., Ruhlmann, C., Gesnel, M.C., and Breathnach, R. (1986) Mol. Cell. Biol. 6, 1679-1686.
44. Stumpo, D.J., Stewart, T.N., Gilman, M.Z., and Blackshear, P.J. (1988) J. Biol. Chem. 263, 1611-1614.
45. Osborn, L., Rosenberg, M.P., Keller, S.A., and Meisler, M.H. (1987) Mol. Cell Biol. 7, 326-334.
46. Osborn, L., Rosenberg, M.P., Keller, S.A., Ting, C., and Meisler, M.H. (1988) J. Biol. Chem. 263, 16519-16522.
47. Prager, D., and Melmed, S. (1988) J. Biol. Chem. 263, 16580-16585.
48. Alexander, M.C., Lomanto, M., Nasrin, N., and Ramaika, C. (1988) Proc. Natl. Acad. Sci. USA 85, 5092-5096.