Identification of a conserved sequence in the non-coding regions of many human genes

Lawrence A. Donehower*, Betty L. Slagle, Margaret Wilde¹, Gretchen Darlington¹ and Janet S. Butel

Department of Virology and ¹Department of Pathology, Baylor College of Medicine, Houston, TX 77030, USA

Received September 9, 1988; Revised and Accepted December 19, 1988

Accession no. X13001

ABSTRACT

We have analyzed a sequence of approximately 70 base pairs (bp) that shows a high degree of similarity to sequences present in the non-coding regions of a number of human and other mammalian genes. The sequence was discovered in a fragment of human genomic DNA adjacent to an integrated hepatitis B virus genome in cells derived from human hepatocellular carcinoma tissue. When one of the viral flanking sequences was compared to nucleotide sequences in GenBank, more than thirty human genes were identified that contained a similar sequence in their non-coding regions. The sequence element was usually found once or twice in a gene, either in an intron or in the 5' or 3' flanking regions. It did not share any similarities with known short interspersed nucleotide elements (SINEs) or presently known gene regulatory elements. This element was highly conserved at the same position within the corresponding human and mouse genes for myoglobin and *N-myc*, indicating evolutionary conservation and possible functional importance. Preliminary DNase I footprinting data suggested that the element or its adjacent sequences may bind nuclear factors to generate specific DNase I hypersensitive sites. The size, structure, and evolutionary conservation of this sequence indicates that it is distinct from other types of short interspersed repetitive elements. It is possible that the element may have a cis-acting functional role in the genome.

INTRODUCTION

Virtually all mammalian genes contain various types of repetitive elements in their non-coding regions. The two most abundant classes of interspersed repetitive sequences associated with genes are the long interspersed nucleotide elements (LINEs) and short interspersed nucleotide elements (SINEs) (1-3). Both types of elements have structural properties (flanking direct repeats and 3' A-rich regions) which indicate that they are inserted into different sites in the genome via an RNA intermediate form. Consequently, these repetitive elements have been classified as retrotransposons (2). Typically, these elements show greater interspecies divergence than intraspecies divergence. For example, the dominant SINE families of humans (*Alu*) and mice (B1) are ancestrally related, but show significant differences both in consensus sequence and overall structure (1-3). No clear functional role has been assigned to these types of repetitive elements.

In addition to repetitive elements and non-functional DNA, the non-coding regions of genes possess regulatory elements that allow the appropriate level of expression of the encoded protein. A number of cis-acting gene regulatory elements have been characterized, and some have been shown to bind nuclear factors in a specific manner (4,5). While the majority of these regulatory elements appear to be located 5' of the coding sequences, a number of examples indicate that cis-acting sequences that regulate expression can be present in introns or in the 3' flanking region of a gene (6-10).

In this paper, we report the characterization of a sequence element which appears to be highly conserved among mammals. The sequence is about 70 bp in length and is usually

present once or twice in the non-coding regions of at least thirty human genes and a number of other mammalian genes. It has none of the features typical of SINEs, such as flanking direct repeats or 3' A-rich tracts. The evolutionary conservation of this sequence suggests a cis-acting functional role, and we propose that this element is part of a class of interspersed repetitive elements distinct from other SINEs.

MATERIALS AND METHODS

Molecular Cloning and Sequencing

Standard molecular cloning and nucleotide sequencing methods were used and are fully described in Zhou et al. (11).

Sequence Comparisons

All computer sequence analysis work was performed on the Baylor College of Medicine Molecular Biology Information Resource. The streamlined user interface EuGene, developed by Thomas Shalom, was used to efficiently search GenBank and make sequence comparisons. The GenBank search programs were developed by Charles Thomas and Dan Goldman (12,13). The default parameters (unit cost matrix used, 15 nucleotide minimum length of similarity for acceptance, and only matches with an SD value above 3.5 reported) were used for the search function (12). Programs used for optimal alignments of similar sequences were according to Altschul and Erickson (14) and Lawrence and Goldman (13). The parameters of the Altschul and Erickson (14) alignments were the default parameters (Dayhoff matrix used, a cost penalty of 2.5 for opening a gap, and a cost penalty of 0.5 for each space in a gap).

Preparation of Hepatoma Extracts

Nuclear extracts were prepared from the Hep3B and HepG2 human hepatocellular carcinoma cell lines (15). Cells were trypsinized from eight roller bottles, pelleted by low-speed centrifugation, and suspended in solution A2 (10 mM HEPES, pH 7.2, 0.15 mM spermine, 0.5 mM spermidine, 2 mM EGTA, and 2 mM DTT). The suspended cells were lysed by Dounce homogenization, and the crude nuclei were pelleted twice by low-speed centrifugation in solution A2 and resuspended in 76 ml of solution C1 (10 mM HEPES, pH 7.6, 25 mM KCl, 0.5 mM spermidine, 0.15 mM spermine, 2 mM EDTA, 0.5 mM EGTA, 2 mM DTT, 1.45 M sucrose, and 10% glycerol). The suspension was pelleted at 25,000 rpm for 30 minutes in an SW28 rotor. The pellet was suspended in solution D (10 mM HEPES, pH 7.6, 100 mM KCl, 3 mM MgCl₂, 0.1 mM EDTA, 1 mM DTT, 0.1 mM PMSF, and 10% glycerol) and then diluted to 10 A₂₆₀ units/ml. One-tenth volume of 4 M ammonium sulfate was added, and the mixture was placed on ice with occasional gentle mixing. The lysate was then centrifuged at 45,000 rpm in a Beckman 50Ti rotor for 1 hour, and 0.3 g of ammonium sulfate was added to each ml of supernatant. The precipitated proteins were pelleted in the Sorvall SS34 rotor at 10,000 rpm for 20 minutes and then redissolved in 2.5 ml of solution E (50 mM Tris, pH 8.0, 0.1 mM EDTA, 1 mM DTT, 12.5 mM MgCl₂, and 20% glycerol). The solution was desalted on a Pharmacia PD10 disposable G25 column and frozen as aliquots in liquid nitrogen.

DNase I Footprinting

Footprints were performed as described by Carthew et al. (16), based on the method of Galas and Schmitz (17). Briefly, 10,000 cpm of ³²P-end-labelled DNA (250-bp fragment) were incubated with 10 µg of bovine serum albumin or 10 µg of the Hep3B nuclear extract in solution E for 60 minutes at 30°C. Five ng of DNase I was added to each reaction, and the mixture was incubated on ice for 3 minutes before addition of 1 µl of 0.5 M EDTA, followed by 100 µl of stop buffer (1% SDS, 100 mM Tris-HCl, pH 8.0, 0.5 mg/ml proteinase K, and 50 µg/ml yeast RNA). This mixture was phenol-extracted, chloroform-extracted and ethanol-precipitated prior to denaturation in formamide-dye buffer and loading on a 6% denaturing polyacrylamide gel for electrophoresis followed by autoradiography. Nucleotide sequence determination of the end-labelled fragments for use as markers was performed according to Maxam and Gilbert (18).

RESULTS

Identification of a Sequence Similar to Elements in Other Human Genes

The probe DNA sequence described here (Figure 1C) was isolated as part of a study on the role of hepatitis B virus (HBV) infection in the development of hepatocellular carcinoma (HCC) in humans (11). Genomic DNA derived from a liver tumor tissue of an individual with chronic HBV infection was analyzed by Southern blot hybridization and found to contain two integrated copies of HBV DNA. One of the integrated HBV DNAs and its flanking cellular DNA was molecularly cloned and analyzed extensively by restriction endonuclease mapping and selective DNA sequencing (Figure 1B). A 1.0-kb human flanking DNA fragment (Figure 1B) contained no repetitive DNA sequences after hybridization to human DNA at high stringency and was mapped to the p11.2-p12 region of chromosome 17 (Figure 1A) (11).

The sequenced segments of the human flanking DNA were further analyzed by computer. None of the flanking sequences contained discernible open reading frames. A search of the GenBank data base for possible nucleotide sequence similarities was negative for all of the sequence fragments except one. This one sequence (Figure 1C) showed significant similarity to a number of GenBank entries, and its reverse complement also showed similarity to GenBank sequences, indicating that the element was present in genes in both orientations. Optimal alignment of the HBV flanking sequence (Figure 2A) with 17 GenBank human DNA entries (Figure 2B) revealed a startling degree of similarity extending about 70 nucleotides among most of the entries. No gaps had to be introduced into any of the sequences to provide a better fit. A consensus sequence was derived from these alignments (Figure 2C), in which a given nucleotide position has an identical base for 9 or more of the 18 DNAs listed. Secondary nucleotides present in at least 6 of 18 sequences are indicated below the primary consensus nucleotides. Similarity of greater than 75% (at least 14 of the 18 DNAs contained the consensus nucleotide) was noted at 28 positions (Figure 2C, asterisks). In general, the highest amount of sequence conservation is found near the center of the 70-bp region, with decreasing conservation moving away from this central core sequence. The calculated probability values for obtaining the consensus nucleotides at a given frequency at each position (Figure 2D) are usually very low, indicating a high degree of significance of these similarities throughout most of the aligned sequences.

The most conserved part of the consensus sequence derived in Figure 2C (the central 45-bp portion) was used to further probe the GenBank library for entries with sequence similarities. Additional sequences with significant similarity were obtained, and the sequence alignments for 35 of the most similar entries are presented (Figure 3). The 45-bp consensus (+) probe is displayed at the top of Figure 3, and the region of highest similarity (core similarity) is indicated by double underlining. In the 15-bp core similarity region, 14 of the positions have at least 75% nucleotide identity among the entries. In four positions at least 34 of the 35 entries have the same nucleotide (Figure 3, bottom). When the data from 20 consensus-like sequences in the opposite orientation (Table 2) are included in these comparisons, 55 of 55 sequences have an A at position 33, 54 of 55 sequences have an A at positions 35 and 39, and 53 of 55 sequences have a G at position 38.

Thus far, we have identified over thirty different human genes and eight genes from other mammalian species that contain a consensus-like sequence (Table 1 and Table 2). In addition, three human sequences are represented that are not associated with a particular gene. Two of these sequences are in potential origin of replication regions: the human *ARS1* sequence (19) and the African green monkey SV40 origin-like sequence (20). The standard deviation (SD) value in Table 1 is a measure of the degree of similarity (12). Values of 3.0 or greater are considered to reflect possible similarity, and values above 6.0 are considered to have probable similarity (13). The position of the consensus-like sequence within each gene is indicated (Tables 1 and 2, third column). Many of the genes contain the sequence within an intron, although some genes have the element either in their 5' or 3' flanking regions. In only one case (human interleukin 1) is the consensus-like sequence present within a gene exon. However, in this instance, it is located within the 3' non-coding region of the gene. Interestingly, the human acetylcholine receptor (alpha

Figure 1. Genetic and physical maps showing the location of the original consensus-like sequence in a human hepatocellular carcinoma (HCC). See Zhou et al. (11) for details on molecular cloning, chromosomal localization, and sequencing. (A) Pictorial representation of human chromosome 17. The hepatitis B virus (HBV) flanking sequences shown in B map to the 17p11.2-17p12 region of chromosome 17. (B) Restriction endonuclease map of a cloned 9.0-kb EcoRI fragment derived from HCC genomic DNA containing integrated HBV sequences and human flanking DNA. HBV DNA (middle boxed area) contains two genes (labelled "S" and "C"). Open boxes represent the pre-S gene and hatched boxes represent the gap between pre-S and C. Enzymes used in the map include EcoRI (E), BglII (B), HindIII (H), XhoI (X), and XbaI (Xb). The 1.0-kb BglII-EcoRI fragment, used for chromosome mapping, is represented by a closed box at the right of the map. Regions that have been sequenced are indicated by lines below the map. (C) The sequence containing the consensus-like element is shown.

subunit) gene contains the consensus-like sequence as part of a 49-bp tandem direct repeat. Finally, the number of nucleotides in the central 15-bp core similarity region (double underlined nucleotides in Figure 3 consensus sequence) that are identical to the consensus are indicated for each entry (Tables 1 and 2, fourth column). Fourteen entries have identity with the consensus sequence in at least 14 of 15 nucleotide positions in Table 1. Similar results are shown in Table 2 (5 of 20 entries have at least 14 of 15 nucleotides which match the core consensus), which lists consensus-like sequences found in the reverse orientation (to those in Table 1) in a number of human and mammalian genes.

Two of the human genes listed in Table 1, human myoglobin and human *N-myc*, have mouse counterpart genes that are also reported. To test for relative evolutionary conservation of the consensus-like sequence within the human and mouse myoglobin genes, we performed a homology matrix comparison between the second intron (and first and second exons) of each gene (Figure 4). Only matches over 10 nucleotides with a standard deviation value above 3.0 are shown. As expected, the mouse and human exons 2 and 3 are highly homologous, while intron 2 contains only a few scattered regions of homology. Two of these homologous stretches correspond exactly to the match ($SD = 5.4, 3.3$) between the consensus-like sequences in the human and mouse intron 2. The homology between the consensus-like elements does not extend on either side. There are only two other regions

Figure 3. Sequences from GenBank with significant similarity to the consensus sequence. The central 45 nucleotides (nucleotides 14-58) of the consensus sequence derived in Figure 2C are shown at the top. This sequence was used to search GenBank for other similar sequences. Thirty-five DNAs that revealed a significant degree of similarity are aligned with the consensus sequence. (Each DNA is 60 bp in length with the 45-bp region of similarity in the middle). The gene or DNA element from which each listed sequence was obtained is identified at the left. The number of entries that shared identity with the consensus nucleotide at each position is indicated at the bottom. Positions that show a frequent secondary nucleotide are indicated below the consensus nucleotide. The 15 central nucleotides that exhibit the highest degree of similarity (core similarity) are emphasized in the consensus sequence by double underlining.

of homology in intron 2 near the 3' portion. Human *N-myc* and mouse *N-myc* also displayed homology between their consensus-like sequences (data not shown).

Binding of Nuclear Factors to the Consensus-Like Sequence

If the consensus-like sequence plays a functional role in the cell, then it might be expected to specifically bind trans-acting nuclear factors as do a number of other cis-acting regulatory elements. To explore this possibility, we incubated a small DNA fragment (a 250-bp HindIII-EcoRI fragment from the 1.0-kb probe DNA) containing the consensus-like sequence with a nuclear extract prepared from Hep3B hepatoma cells and subjected the reaction to a DNase I footprinting assay. Evidence of specific protein binding may be more

Table 1. Characteristics of 35 (+) consensus-like sequences

Gene or DNA Sequence	SD ^a	Location	Core Similarity ^b
1) human adenosine deaminase	9.7	intron 1	14/15
2) human protein C	9.6	5' flanking	15/15
3) human beta-tubulin	9.3	intron 3	12/15
4) human alpha-1-antitrypsin	9.1	5' flanking	12/15
5) human haptoglobin (alpha-2)	8.8	intron 5	13/15
6) human myoglobin	8.7	intron 2	14/15
7) X chromosome sequences (near DMD)	8.4	unknown	13/15
8) human carbonic anhydrase II	8.3	intron 1	14/15
9) African green monkey SV40 <i>ori</i> like	8.1	unknown	13/15
10) human <i>c-sis</i>	7.5	5' flanking	13/15
11) human haptoglobin-related	7.1	intron 5	13/15
12) human <i>N-myc</i>	7.0	intron 2	15/15
13) human factor IX	6.7	intron 3	11/15
14) human prothrombin	6.7	intron 1	12/15
15) human acetylcholine receptor (alpha)	6.6	3' flanking (rpt) ^c	11/15
16) human apolipoprotein A-1	6.6	3' flanking	14/15
17) human alpha-fetoprotein	6.5	intron 3	13/15
18) human serum prealbumin	6.5	intron 3	13/15
19) mouse myoglobin	6.5	intron 2	14/15
20) human apolipoprotein C-III	6.3	intron 2	13/15
21) rabbit poly Ig receptor	6.3	3' flanking	13/15
22) human fibrinogen	6.3	intron 6	13/15
23) human interferon beta-3	6.0	unknown	12/15
24) human immune interferon gamma	5.9	intron 3	11/15
25) human opsin	5.8	intron 1	11/15
26) bovine acetylcholine receptor (alpha)	5.8	3' flanking	11/15
27) human <i>ARSI</i> sequence	5.6	rep enhancer	13/15
28) human enkephalin B	5.4	intron 3	11/15
29) rat thyrotropin beta subunit	5.0	intron 1	12/15
30) human protein C gene	4.8	intron 3	14/15
31) human interleukin 1	4.6	exon 7 (3' UT) ^d	11/15
32) bovine pancreatic trypsin inhibitor	4.2	3' flanking	11/15
33) mouse <i>N-myc</i>	4.2	intron 2	12/15
34) human T-cell receptor beta chain	4.2	D-J region	14/15
35) mouse glial fibrillary acid protein	3.6	intron 7	14/15

^aSD measures the degree of similarity to the consensus sequence. See text for additional details.

^bCore similarity represents the number of nucleotides in each consensus-like sequence (first number) which are identical to the most highly conserved central 15 nucleotides of the consensus sequence (Figure 3, double underlined nucleotides).

^cThe consensus-like sequences of the alpha subunit gene are within a 49-bp tandem direct repeat.

^dThe consensus-like sequence is within the 3' untranslated region.

readily observable at the higher resolution afforded by this particular procedure. The DNA-binding reactions were performed in an excess of the simple alternating copolymer duplex poly(dI-dC) with a constant amount of ³²P-labelled 250-bp fragment and 10 ug of Hep3B extract.

Table 2. Characteristics of 20 (-) consensus-like sequences^a

	Gene	SD	Location	Similarity	
				Core	Total
1)	human interleukin 2	10.4	5' flanking	13/15	39/45
2)	human <i>c-sis</i>	9.9	5' flanking	14/15	37/45
3)	bovine acetylcholine receptor ^b	9.1	3' flanking	14/15	37/45
4)	human beta crystallin	8.8	intron 4	13/15	37/45
5)	human myoglobin	8.7	intron 1	13/15	36/45
6)	human adenosine deaminase	8.4	5' flanking	14/15	34/45
7)	human protein C	8.2	5' flanking	12/15	36/45
8)	human prothrombin	8.0	intron 12	12/15	36/45
9)	human alpha-fetoprotein	7.6	intron 3	12/15	34/45
10)	human dystrophin	7.4	intron 7	13/15	34/45
11)	human alkaline phosphatase	7.3	intron 9	13/15	29/45
12)	human apolipoprotein C-III	7.2	intron 3	13/15	35/45
13)	human epsilon globin	6.9	5' flanking	13/15	32/45
14)	human aldolase B gene	6.7	5' flanking	14/15	35/45
15)	human prolactin gene	6.7	5' flanking	12/15	33/45
16)	human tissue plasminogen act.	6.3	intron 4	15/15	29/45
17)	human interleukin 1	6.2	intron 4	11/15	34/45
18)	human myoglobin	5.9	intron 2	13/15	33/45
19)	human opsin	5.9	3' flanking	11/15	31/45
20)	mouse myoglobin	4.2	intron 2	12/15	30/45

^aThese consensus-like elements are in the reverse orientation of those shown in Table 1 and are homologous to the opposite strand of the 45 nucleotide consensus sequence at the top of Figure 3. The (-) strand consensus probe is:

5'-TACTTAACCTCTCTGAGCCTCAGTTTCCTCATCTGTAAAATGGGG-3'.

^bBeta subunit.

Both strands of the 250-bp fragment exhibited DNase I hypersensitive cleavage sites in the presence of the Hep3B extract (Figure 5A). These sites were not apparent when the fragment was incubated with DNase I in the presence of bovine serum albumin. Interestingly, the strongest hypersensitive sites on each strand appear to be at the same position and were within the consensus-like sequence. These strong hypersensitive bands were consistently observed in three separate DNase I footprinting experiments as well as with an extract derived from the HepG2 hepatoma cell line (data not shown). Other less hypersensitive sites were present on one strand at the other end of the consensus-like sequence. No obvious protected regions were evident (Figure 5A), although three possible regions of decreased DNase I cleavage appeared, all outside of the consensus-like sequence. The relationship of the DNase I hypersensitive sites to the DNA sequence is illustrated (Figure 5B).

DISCUSSION

As a result of sequence analysis of human DNA associated with an integrated HBV genome in a hepatoma, we have discovered a sequence element that appears to be highly conserved in a number of human and other mammalian genes. The sequence itself displays some interesting features, including purine-rich tracts alternating with pyrimidine-rich regions. The 5' region of some of the consensus-like sequences has the potential to form stem-loop structures, but the significance of these putative secondary structures is unclear.

The consensus-like sequences do not appear to have similarity to any previously described human repetitive DNA element and do not share the properties of a short interspersed nucleotide element (SINE), typified by the human *Alu* repeats. The consensus-like sequence observed in human genes does not contain an A-rich tract at its 3'

Figure 4. Homology matrix comparison between human and mouse myoglobin in the intron 2 (and adjacent exons) sequences. Mouse myoglobin sequences (top) are matched with human myoglobin sequences (left). Sequences greater than 10 nucleotides in length which show homology at an SD value of 3.0 will generate a diagonal line in this matrix display. The distance between each vertical and horizontal line in the matrix is equal to 100 nucleotides. The consensus-like sequences in each intron are represented by a closed box with an arrow. Note that the consensus-like sequences are conserved between the human and mouse genes.

end nor is it bordered by direct repeats, features characteristic of SINEs (1-3). In most cases, the consensus-like sequence is present once or twice per gene, whereas dozens of SINEs can be located within a single gene (1-3), indicating significantly lower copy numbers of the consensus-like sequence. However, it is possible that our screen underestimated the number of consensus-like sequences, since the more divergent members of this family might not be picked up by our computer searches.

The consensus-like sequences are conserved among a variety of mammalian species, whereas rodent and human SINEs, although ancestrally related, have significant differences in overall structure and primary sequence (3). Whereas most of the genes containing the consensus-like sequences are human (Table 1 and 2), a significant number of rodent and other mammalian species contain this repeat. The predominance of human genes in Tables 1 and 2 may be due to the greater number of human genes sequenced (including intron and flanking sequences). Alternatively, the non-human consensus-like sequences may be somewhat diverged from their human counterparts. To test this possibility, we derived a consensus sequence from six consensus-like elements in rat and mouse genes. This rodent consensus sequence was identical to the 15-bp core region and nearly identical to the entire 45 nucleotides of the consensus sequence in Figure 3. This test is consistent with the results of the homology matrix comparison of Figure 4, which shows that the consensus-like sequences in intron 2 of the mouse and human myoglobin gene are conserved, whereas the adjacent non-coding sequences share no homology. These data suggest that the human and mouse consensus-like elements are conserved over some 60-85 million years of evolution. Non-mammalian vertebrate genes in GenBank contained no significant sequence similarities to the consensus sequence.

The differences in structure and evolutionary conservation between the consensus-like

Figure 5. DNase I footprint analysis of a 250-bp HBV flanking DNA fragment containing the consensus-like sequence. The fragment tested was from the 3' end of the 1.0-kb fragment indicated in Figure 1B. DNase I footprinting was performed as described in the Materials and Methods section. (A) Lanes 1-5 contain the negative strand end-labelled at the EcoRI site. Lanes 6-11 contain the positive strand end-labelled at the HindIII site. Maxam and Gilbert sequencing reactions for C+T (lanes 1 and 6), G+A (lanes 2 and 7), and A+C (lane 8) are used as markers. Untreated control DNA fragments are in lanes 3 and 9. Control reactions in which the DNA fragments were incubated with bovine serum albumin before DNase I treatment are shown in lanes 4 and 10. Lanes 5 and 11 contain fragments incubated with Hep3B nuclear extract prior to DNase I treatment. Hypersensitive cleavage sites by DNase I in the presence of Hep3B extracts are indicated by arrowheads. The dominant hypersensitive site on each strand is shown by a large arrowhead. The regions of the fragments containing the consensus-like sequence are designated by vertical lines next to the marker lanes. (B) Summary of footprinting assay in the vicinity of the consensus-like sequence on both strands. The same DNase I hypersensitive cleavage sites are indicated by arrowheads, as in panel A. Underlined nucleotides represent the 45-bp consensus region. Double underlined nucleotides represent the 15 bp of core similarity.

repeats and the typical SINEs indicate that the former should be placed in a separate class of repetitive elements. Other investigators have reported non-*Alu* types of short interspersed repeats (21-23). Some of these are apparently retrotransposons and are inserted (22). Another group of interspersed repeats is composed of tandem repetitions of short oligonucleotides, which are highly conserved (3,24,25). The consensus-like sequences reported here show no evidence of being inserted elements and are not repeated oligonucleotides but are highly conserved. We therefore propose that the consensus-like sequences form a different class of interspersed repetitive elements.

One possible explanation for the evolutionary conservation of the consensus-like repeat is that these sequences play a functional role in the cell. The presence of a highly conserved core sequence is consistent with the idea that the sequence binds a nuclear factor or factors. Comparison of the consensus sequence to a list of sequences associated with nuclear factor binding (5) did not uncover any similarities except possibly in one case. A portion of the HBV genome to which liver-specific factors bind (26) showed identity at 11 of 13 positions with nucleotides 15-27 (Figure 2C) of the consensus sequence. This particular sequence is directly adjacent to the dominant DNase I hypersensitive sites in the consensus-like sequence of the HBV flanking DNA (Figure 5B).

In two instances, the consensus-like element is associated with sequences that display potential origin of replication function. In one example, it is adjacent to a sequence in the African green monkey genome which is similar to the SV40 origin of replication (20). In the other case, the consensus-like sequence is part of the human *ARS1* DNA (19). *ARS1* is a sequence of human DNA that allows replication of yeast integrative plasmids as autonomously replicating elements in yeast cells (19). Within *ARS1* is a 325-bp segment that is necessary for maximal expression of the autonomous replication phenotype, analogous to an enhancer. Interestingly, the consensus-like sequence is positioned in the center of this 325-bp segment having replication enhancer activity. It is not known, however, whether the consensus-like sequence is part of the active portion of the replicative enhancer. It should also be noted that another cloned human sequence, *ARS2*, which also allows autonomous replication in yeast, does not appear to contain the consensus-like sequence.

The DNase I footprinting experiments suggested that nuclear factors in extracts derived from two different liver cell lines bound the DNA fragment containing the consensus-like element and altered its susceptibility to DNase I. Liver cell nuclear extracts were employed because the original HBV flanking sequence was isolated from hepatoma tissue and a number of the genes containing the consensus-like sequence are liver-specific. The footprinting results suggest at least a specific DNA conformational change, since DNase I hypersensitive sites were consistently noted in the same position of the 250-bp fragment. Whether specific binding of discrete factors to the DNA occurred is somewhat unclear, although preliminary gel retardation binding assays suggested specific factor binding (data not shown). Interestingly, the observed DNase I hypersensitive sites were localized to both ends of the consensus-like sequence. This phenomenon of DNase I hypersensitivity has frequently been observed at nucleotide positions adjacent to sites of binding by transcriptional factors (27,28). There was some evidence of DNase I protection of specific domains of the fragment, but the putative protected regions were outside of the consensus-like sequences. Further experiments must be performed to clarify whether there is a specific interaction of nuclear factors with the consensus-like sequence. The primary question is whether this well conserved sequence has a function in gene expression, DNA replication, or some other cellular event.

ACKNOWLEDGEMENTS

We thank Charlie Lawrence, Sandy Honda, Debbie Wilson and Jim Kelly for helpful discussions and Joyce Evans for help with the manuscript.

This work was supported in part by Public Health Service grant CA37257 from the National Cancer Institute and by National Research Service Award CA09197.

*To whom correspondence should be sent

REFERENCES

1. Singer, M.F. (1982). *Cell* **28**, 433-434.
2. Weiner, A.M., Deininger, P.L., and Efstratiadis, A. (1986) *Ann. Rev. Biochem.* **55**, 631-661.
3. Schmid, C.W. and Shen, C.-K.J. (1986) In R.J. MacIntyre (ed.), *Molecular Evolutionary Genetics*, Plenum, New York, pp. 323-358.
4. Gluzman, Y. (ed.) (1985) *Eucaryotic Transcription: The Role of cis- and trans-acting Elements in Initiation*, Cold Spring Harbor Laboratory, New York.
5. Wingender, E. (1988) *Nucleic Acids Res.* **16**, 1879-1902.
6. Coulombe, B., Ponton, A., Daigneault, L., Williams, B.R.G. and Skup, D. (1988) *Mol. Cell. Biol.* **8**, 3227-3234.
7. Reisman, D., Greenberg, M., and Rotter, V. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 5146-5150.
8. Behringer, R.R., Hammer, R.E., Brinster, R.L., Palmiter, R.D., and Townes, T.M. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 7056-7060.
9. Brinster, R.L., Allen, J.M., Behringer, R.R., Gelinas, R.E., and Palmiter, R.D. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 836-840.
10. Bornstein, P., McKay, J., Morishima, J.K., Devarayalu, S., and Gelinas, R.E. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 8869-8873.
11. Zhou, Y.-Z., Slagle, B.L., Donehower, L.A., vanTuinen, P., Ledbetter, D.H., and Butel, J.S. (1988) *J. Virol.* **62**, 4224-4231.
12. Lawrence, C.B., Goldman, D.A., and Hood, R.H. (1986) *Bull. Math. Biol.* **48**, 569-583.
13. Lawrence, C.B. and Goldman, D.A. (1988) *CABIOS* **4**, 25-33.
14. Altschul, S.F. and Erickson, B.W. (1986) *Bull. Math. Biol.* **48**, 603-616.
15. Knowles, B., Howe, C.C., and Aden, D.P. (1980) *Science* **209**, 497-501.
16. Carthew, R.W., Chodosh, L.A., and Sharp, P.A. (1985) *Cell* **43**, 439-448.
17. Galas, D. and Schmitz, A. (1978) *Nucleic Acids Res.* **5**, 3157.
18. Maxam, A. and Gilbert, W. (1980) *Meth. Enzymol.* **65**, 499-560.
19. Montiel, J.F., Norbury, C.J., Tuite, M.F., Dobson, M.J., Mills, J.S., Kingsman, A.J., and Kingsman, S.M. (1984) *Nucleic Acids Res.* **12**, 1049-1067.
20. Saffer, J.D. and Lerman, M.I. (1983) *Mol. Cell. Biol.* **3**, 960-964.
21. Watanabe, Y., Tsukaga, T., Notake, M., Nakanishi, S., and Numa, S. (1982). *Nucleic Acids Res.* **10**, 1459-1469.
22. Spence, S.E., Young, R.M., Garner, K.J., and Lingrel, J.B. (1985). *Nucleic Acids Res.* **13**, 2171-2186.
23. Sun, L., Paylson, K.E., Schmid, C.W., Kady, K.L., and Leinwand, L. (1984). *Nucleic Acids Res.* **12**, 2669-2690.
24. Miesfeld, R., Krystal, M., and Arnheim, N. (1981). *Nucleic Acids Res.* **9**, 5931-5447.
25. Proudfoot, N.J., Gill, A., and Maniatis, T. (1982). *Cell* **31**, 553-563.
26. Shaul, Y. and Ben-Levy, R. (1987) *EMBO J.* **6**, 1913-1920.
27. Wildeman, A.G., Zenke, M., Schatz, C., Takahashi, K., Barrera-Saldana, H., Grundstrom, T., Wintzerith, M., Matthes, H., Vigneron, M., and Chambon, P. (1985) In Gluzman, Y. (ed), *Eukaryotic Transcription: The Role of cis- and trans-acting Elements in Initiation*, Cold Spring Harbor Laboratory, New York, pp. 19-26.
28. Cereghini, S., Piette, J., Kryszke, M.-H., and Yaniv, M. (1985) In Gluzman, Y. (ed), *Eukaryotic Transcription: The Role of cis- and trans-acting Elements in Initiation*, Cold Spring Harbor Laboratory, New York, pp. 78-85.