Volume 14 Number 13 1986 Nucleic Acids Research

Isolation and characterization of the human catalase gene

F.Quan[1,2], R.G.Korneluk[1,3], M.B.Tropak[1] and R.A.Gravel[1,2]*

[1]Research Institute, Hospital for Sick Children, Toronto, ONT, M5G 1X8 and [2]Department of Medical Genetics, University of Toronto, ONT, M5S 1A8, Canada

Received 5 February 1986; Accepted 25 April 1986

ABSTRACT

Catalase is a tetrameric hemoprotein which degrades H_2O_2. Recombinant phage clones containing the human catalase gene have been isolated and characterized. The gene is 34 kb long and is split into 13 exons. The precise size and location of the exons has been determined. In addition, essentially full length catalase cDNA clones have been isolated and sequenced and used to tentatively identify the 5'-end of the gene. This assignment, if correct, predicts that the region upstream of the gene does not contain a TATA box. This region is GC rich (67%) and contains several CCAAT and GGGCGG sequences which may form part of the promoter. Translation of the catalase mRNA appears to begin immediately upstream of the amino-terminal Ala residue of catalase.

INTRODUCTION

Catalase (E. C. 1. 11. 1. 6; H_2O_2-H_2O_2 oxidoreductase) is an enzyme which catalyzes the decomposition of hydrogen peroxide to oxygen and water. It is found in virtually all aerobic cells and is partly responsible for protecting cells from the toxic effects of hydrogen peroxide (1, 2).

Mammalian catalase occurs as a complex of four identical subunits. Each subunit has a molecular weight of approximately 60 KDa and contains a single hematin (Fe(III)-protoporphyrin IX) group (1, 2). The amino acid sequences of bovine liver and erythrocyte (3) and human erythrocyte catalase (4) have been reported.

In mammalian tissues the highest levels of catalase are found in the liver, kidney, and erythrocytes while the lowest levels are found in connective tissues (1, 2, 5). Shingu et al have reported the absence of catalase activity in human vascular smooth muscle cells and endothelial cells (6). In tissues such as the liver, catalase is found predominantly in peroxisomes. However in mature human erythrocytes catalase is found free in the cytosol (1, 2, 5).

Catalase deficiency was first described by Takahara in 1948 (7).

Acatalasemia is inherited as an autosomal recessive trait and is characterized by an erythrocyte catalase level that varies from 0.2-4% of normal. Catalase activity may or may not be deficient in other tissues. Surprisingly, the manifestation of clinical symptoms is rare and restricted to a progressive oral gangrene (5, 8).

The gene for human catalase has been mapped to chromosome 11, band p13. Wilm's tumor, a common childhood neoplasia can be associated with deletions of chromosome 11 centering around band p13. In some individuals, the deletions are associated with reduced catalase activity (9, 10, 11). The human catalase gene is currently being investigated as a unique marker for a gene or genes located in 11p13 which predispose individuals to this tumor (12).

A knowledge of the structure of the human catalase gene would facilitate studies into the regulation of catalase levels in different tissues and is essential for determining the nature of the mutations which result in catalase deficiency. In addition studies into the etiology of Wilm's tumor would be facilitated by the isolation of probes for 11p13. In this report we describe the isolation and detailed characterization of the human catalase gene. In addition we report the isolation and sequence of essentially full length catalase cDNA clones.

MATERIALS AND METHODS

Genomic Southern Blots

High molecular weight DNA was isolated from human lymphoblasts as described (13). Restriction digests were done using the appropriate core buffers as supplied by International Biotechnologies Incorporated. DNA was run on 0.8% agarose gels and transferred to nitrocellulose filters as described (14). Prehybridization was done at 42°C in 50% deionized formamide, 3X SSC, 0.05M sodium phosphate pH 6.7, 1X Denhardt's and 15 µg/ml denatured salmon sperm DNA for at least 1 hour. Hybridization was done in this buffer containing 10% dextran sulfate and 1-5 $X10^6$ cpm/ml of nick translated probe. Restriction fragments were nick-translated using the Amersham kit and α^{32}P-dCTP (3000 Ci/mmol, Dupont New England Nuclear). Filters were rinsed with 2X SSC, 0.1% SDS at room temperature and washed at 65°C with 0.1X SSC, 0.1% SDS. Autoradiography was done at -70°C with intensifying screens.

Isolation and characterization of phage clones

A library constructed in λCharon4A from a partial HaeIII-AluI

digest of human fetal liver DNA (15) was generously provided by T. Maniatis. The library was plated at a density of 40,000 plaques per 150 mm round petri dish.

A human liver cDNA library in λgt11 (16) was generously provided by S. Woo. This library was plated at a density of 75,000 plaques per 24.3 X 24.3 cm^2 Nunc plate.

Phage DNA was transferred to nitrocellulose filters as described (14) and hybridized to nick-translated probes as described above. Positive clones were picked and rescreened at lower density until single pure plaques could be isolated.

Phage DNA was prepared as described (14). Restriction fragments were subcloned into pSP64 and 65 (17) using the low melting temperature agarose (Bethesda Research Laboratories, BRL) method (18).

DNA Sequencing

Restriction fragments were sequenced using the dideoxy-nucleotide chain termination method (19) with double stranded templates (20) after subcloning into pSP64 and 65. In some cases fragments were sequenced after the generation of a nested set of deletions using exonuclease III (BRL) (21).

Northern Blots

RNA was isolated from cultured cells by guanidinium isothiocyanate extraction followed by CsCl centrifugation (22). RNA was run through 1% agarose, 6% formaldehyde gels and transferred to nitrocellulose as described (14). Prehybridization, hybridization, washing and autoradiography were as described above.

RESULTS AND DISCUSSION

In order to obtain an estimate of the size and complexity of the catalase gene, a Southern blot of human lymphoblast DNA, cut with various restriction enzymes, was probed with a 2.0 kb PstI-SnaBI fragment of pCAT41. pCAT41 is a previously isolated 2.2 kb cDNA clone that extends from amino acid 76 to the carboxy terminal residue of catalase (23). Each digest contains hybridizing bands that total 15-20 kb (Fig. 1).

A human genomic library in λCharon4A was screened with a 1.2 kb PvuII-HindIII fragment from pCAT41. Six overlapping phage (λCAT 4, 13, 14, 21, 23, 78; Fig. 2) spanning a total of 34 kb of DNA were isolated. The orientation of the catalase gene was established using 5′ and 3′ restriction fragments from pCAT41. The region upstream of the gene

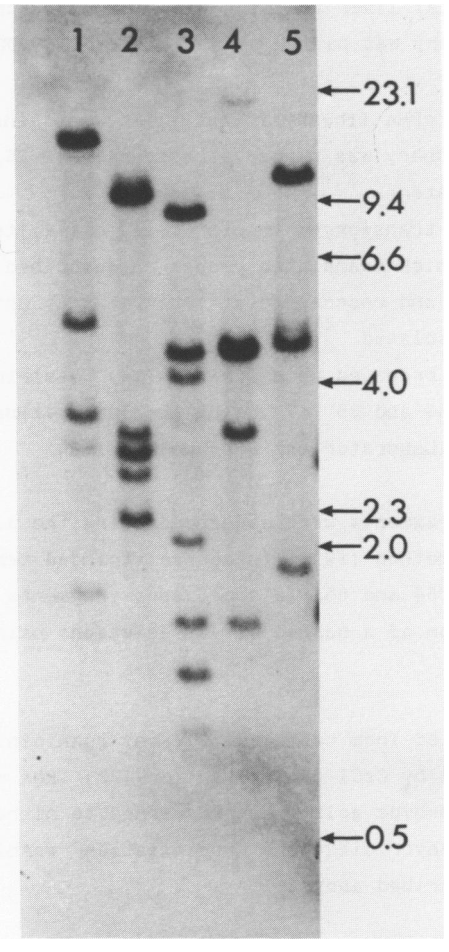

Figure 1. Southern blot of human lymphoblast DNA. DNA was cut with EcoRI (lane 1), HindIII (lane 2), PstI (lane 3), SstI (lane 4), and Xba I (lane 5). The 2.0 kb PstI-SnaBI fragment of pCAT41 was used as probe. The positions of the molecular weight markers are shown on the right.

was isolated by screening the library with genomic fragments. Using an 800 bp ScaI-SnaBI fragment from λCAT13, a phage extending another 4.0 kb, λCAT17, was isolated. Three other phage (λCAT2, 18, 142) were isolated using a 450 bp SmaI fragment from λCAT17. One of these phage, λCAT2, extends an additional 13.5 kb beyond λCAT17 (Fig. 2).

Overlapping restriction fragments were subcloned into pSP64 and 65 for fine structure mapping (Fig. 2). All of the bands in genomic digests hybridizing to pCAT41 were accounted for, indicating that the cloned DNA

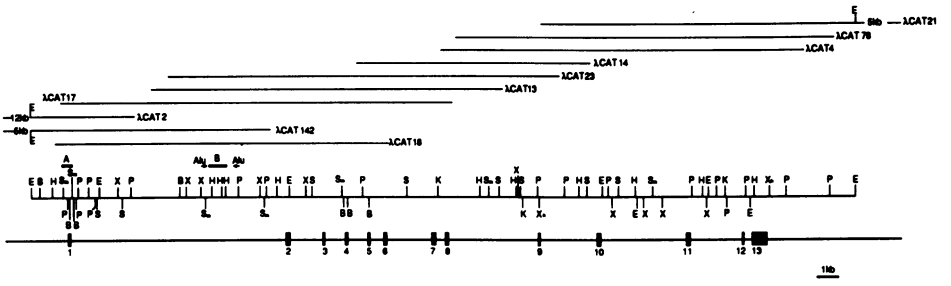

Figure 2. Structure of the human catalase gene. The restriction map of the human catalase gene is shown below the λCharon4A recombinants isolated from the genomic library. Enzymes represented are B, BamHI; E, EcoRI; H, HindIII; K, KpnI; P, PstI; S, SstI; Sm, SmaI; X, XbaI; and Xh, XhoI. No sites were found for SalI. Exons are shown as solid boxes. Fragments A and B are the 450 bp SmaI fragment and the 800 bp ScaI-SnaBI fragment referred to in the text respectively. The Alu sequences flanking fragment B and their orientations are indicated by arrows.

was not rearranged. Exons were located by probing Southern blots of subclones, cut with various combinations of restriction enzymes, with pCAT41. However, because the 5' end of the catalase mRNA is missing from pCAT41, the corresponding exons were found by using cloned genomic fragments to probe Northern blots of HeLa RNA. The fragments hybridizing to pCAT41 or catalase mRNA were sequenced to locate the intron-exon junctions.

As shown in Figure 2, the catalase gene is split into 13 exons by 12 introns and spans a region of approximately 34 kb. The introns range in size from about 400 bp to 10.5 kb. The largest intron separates exons 1 and 2 and contains the 800 bp ScaI-SnaBI fragment that detects a previously described TaqI polymorphism (24). Nucleotide sequencing has shown that this fragment is flanked by two Alu sequences in opposite orientations (data not shown). A previously isolated cDNA clone, pCAT1, contained a 462 bp insertion that had the characteristics of an intron (23). The nucleotide sequence of intron 7 (Fig. 3) corresponds exactly to this 462 bp insertion. This confirms that pCAT1 is an intron containing cDNA clone.

The nucleotide sequence of pCAT41, corresponding to amino acids 76 to the carboxyl terminus of catalase, has been previously reported (23). The complete sequence of the catalase mRNA, together with the intron-exon junctions is now shown in Figure 3. This sequence is in complete agreement with the previously reported sequence with five exceptions. These differences do not change the reported amino acid sequence of catalase. Nucleo-

```
gtcccagggcggcctgaaggatgctgataaccgggagcccccgccctgggttcggctatcccgggcacccc
320                 -300                -280                -260

gggccggcggggcgaggctctccaattgctgggccagagcgggacccttcctttccgcaccctcctgggt
        -240                -220                -200

atctccggtcttcaggcctccttcggagagccctgctccgagcccattgggcttccaatcttggcctgcc
180                 -160                -140                -120

tagcgccgagcagccaatcagaaggcagtcctcccgaggggggcgggacgaggggtggtgctgattggct
        -100                -80                 -60

gagcctgaagtcgccacggactcggggcaacaggcagattTGCCTGCTGAGGGTGGAGACCCACGAGCC
40                  -20                 -1
                                         1
                                       METAlaAspSerArgAspProAlaSerAspG
GAGGCCTCCTGCAGTGTTCTGCACAGCAAACCGCACGCTATGGCTGACAGCCGGGATCCCGCCAGCGACC

10                              20
lnMetGlnHisTrpLysGluGlnArgAlaAlaGln
AGATGCAGCACTGGAAGGAGCAGCGGGCCGCGCAGgtacactctgtgctccccgagcgggcccgaaggtc

cgtttagaaagcgggggcgtcggcaagtaaaggcccggcttctcccggggcggcgcttggagggactgta

ccgcggctcactgggcaggggggatccccttcggtgcagacggacttttacattcgccgaagcaggggag

ggg........ about 10.2kb .......tggcccatcctgtcagattttagtactttggacacagg

aaattaaaaaagagggcagatggtataaacattgcaaagctatgtacccgtgacagtgtaaatgaaaggt

                                                  LysAlaAspValLeuThrThrGlyA
ttgattgtgctaactctcctgcactttctttctgtgttcctgtagAAAGCTGATGTCCTGACCACTGGAG

30                              40                              50
laGlyAsnProValGlyAspLysLeuAsnValIleThrValGlyProArgGlyProLeuLeuValGlnAs
CTGGTAACCCAGTAGGAGACAAACTTAATGTTATTACAGTAGGGCCCCGTGGGCCCCTTCTTGTTCAGGA

                    60                              70
pValValPheThrAspGluMetAlaHisPheAspArgGluArgIleProGluArgValValHisAlaLys
TGTGGTTTTCACTGATGAAATGGCTCATTTTGACCGAGAGAGAATTCCTGAGAGAGTTGTGCATGCTAAA

GlyAlaG
GGAGCAGgtaagtgctgtgt....... about 1.6kb .......aatgtctgagtaatggtctcatg

                                80                              90
                              lyAlaPheGlyTyrPheGluValThrHisAspIleThrLysT
gtaaggatttctgtgtctttctcgttagGGGCCTTTGGCTACTTTGAGGTCACACATGACATTACCAAAT

                        100                             110
yrSerLysAlaLysValPheGluHisIleGlyLysLysThrProIleAlaValArgPheSerThrValA
ACTCCAAGGCAAAGGTATTTGAGCATATTGGAAAGAAGACTCCCATCGCAGTTCGGTTCTCCACTGTTGg

taagttggtttattggcgtgattggtatggcttaactcaacttcacctttgggg...about 1.0kb.

                                    120                             130
                          laGlyGluSerGlySerAlaAspThrValArgAspProArgGlyPhe
......ccatttgaatattgtagCTGGAGAATCGGGTTCAGCTGACACAGTTCGGGACCCTCGTGGGTTT

                            140                             150
AlaValLysPheTyrThrGluAspGlyAsnTrpAspLeuValGlyAsnAsnThrProIlePhePheIleA
GCAGTGAAATTTTACACAGAAGATGGTAACTGGGATCTCGTTGGAAATAACACCCCCATTTTCTTCATCA

rgAspProIleLeu
GGGATCCCATATTGgtaggtaatagagtattttgcactcaacaaatgtttgttgacttaaattgatttca
```

```
..... about 1.0kb.....ttcctgtaaacttagtttttggatttttttctctctttttttcta

     160                              170                              180
     PheProSerPheIleHisSerGlnLysArgAsnProGlnThrHisLeuLysAspProAspMetVa
tttagTTTCCATCTTTTATCCACAGCCAAAAGAGAAATCCTCAGACACATCTGAAGGATCCGGACATGGT

                           190
lTrpAspPheTrpSerLeuArgProGluSerLeuHisGln
CTGGGACTTCTGGAGCCTACGTCCTGAGTCTCTGCATCAGgtatgaaccctttttgccattgtattata

                                                                  ValSerPhe
tcacctgggatgcagt.....about 0.6kb.....tttatatttctgttctttagGTTTCTTTC

     200                              210                              220
LeuPheSerAspArgGlyIleProAspGlyHisArgHisMetAsnGlyTyrGlySerHisThrPheLysL
TTGTTCAGTGATCGGGGGATTCCAGATGGACATCGCCACATGAATGGATATGGATCACATACTTTCAAGC

                           230
euValAsnAlaAsnGlyGluAlaValTyrCysLysPheHisTyrLys
TGGTTAATGCAAATGGGGAGGCAGTTTATTGCAAATTCCATTATAAGgtatgtgtt...about 2.3kb

               240                              250
        ThrAspGlnGlyIleLysAsnLeuSerValGluAspAlaAlaArgLeuSerGlnGluAspPr
....gtagACTGACCAGGGCATCAAAAACCTTTCTGTTGAAGATGCGGCGAGACTTTCCCAGGAAGATCC

      260                              270                              280
oAspTyrGlyIleArgAspLeuPheAsnAlaIleAlaThrGlyLysTyrProSerTrpThrPheTyrIle
TGACTATGGCATCCGGGATCTTTTTAACGCCATTGCCACAGGAAAGTACCCCTCCTGGACTTTTTACATC

                           290                              300
GlnValMetThrPheAsnGlnAlaGluThrProPheAsnProPheAspLeuThrLys
CAGGTCATGACATTTAATCAGGCAGAAACTTTTCCATTTAATCCATTCGATCTCACCAAGgtgagtcagt

aaacaactatattgttttcttttttaagtctcttcttacctaattagaaaaaaaatctagtcaaacaatt

ataataatggggaagtcatatacaaaatacagagggtaccacttcagagtgtcctaagctgtgaatgagt

gcttaccagcatcttacttccacgttcctgtttgtcatttcattgagtatgtgtatgtggcttcatatat

tgttattaacagggaacagattatgaaaagctgatgtacttttttcctggggaaactgtcagtatttacca

cttactattgtgaaagatttaactaaggcactcatcttaaattcttatgttttattggatttaaaaatta

ttttcattggcttgattgtatttgaaatctggtatttttgtgggtagctttgatttccttcagttgattg

                                                                  310
                                    ValTrpProHisLysAspTyrProLeuIleProValGl
cctggtaattgtgaatatgacatcattttcagGTTTGGCCTCACAAGGACTACCCTCTCATCCCAGTTGG

                    320                              330
yLysLeuValLeuAsnArgAsnProValAsnTyrPheAlaGluValGluGlnIleAlaPheAspProSer
TAAACTGGTCTTAAACCGGAATCCAGTTAATTACTTTGCTGAGGTTGAACAGATAGCCTTCGACCCAAGC

          340                              350
AsnMetProProGlyIleGluAlaSerProAspLysMetLeuGln
AACATGCCACCTGGCATTGAGGCCAGTCCTGACAAAATGCTTCAGgtgagcctggtggattgagatgttc

tgagg.....about 4.3kb.....ccattcctatgttatatgttactgcccctagtcagtgtct

                                                  360
                      GlyArgLeuPheAlaTyrProAspThrHisArgHisArgLeuGlyProAsn
attgtatttattactgcagGGCCGCCTTTTTGCCTATCCTGACACTCACCGCCATCGCCTGGGACCCAAT
```

```
   370                                380                                390
TyrLeuHisIleProValAsnCysProTyrArgAlaArgValAlaAsnTyrGlnArgAspGlyProMetC
TATCTTCATATACCTGTGAACTGTCCCTACCGTGCTCGAGTGGCCAACTACCAGCGTGATGGCCCGATGT

ysMetGlnAspAsnGlnG
GCATGCAGGACAATCAGGgtaggcctaaagacgttgggctccccctgcgtgggcagagggcacgtggagc

agatgggcgggaggccagg....about 2.6kb....aaatgcgggaaattaaaaataatatgtgtgcg

ttgtgtttatatctgtgtatgtgtacgtgtgtatttgattaccacttgaatttatttctcatcacagtga

                                                                      40
                                                                   lyGlyAl
ttatttgcagacttacttgacttttcttattcctaagtgcatctgggtggttttgttttgaagGTGGTGC

0                                410                                420
aProAsnTyrTyrProAsnSerPheGlyAlaProGluGlnGlnProSerAlaLeuGluHisSerIleGln
TCCAAATTACTACCCCAACAGCTTTGGTGCTCCGGAACAACAGCCTTCTGCCCTGGAGCACAGCATCCAA

                    430                                440
TyrSerGlyGluValArgArgPheAsnThrAlaAsnAspAspAsnValThrGln
TATTCTGGAGAAGTGCGGAGATTCAACACTGCCAATGATGATAACGTTACTCAGgtaatgacttctcttt

atctgctatggaagtcacctgctaattc.......about 4.0kb.......aattttgttggtgataa

                                                                     450
                                                  ValArgAlaPheTyrValAsnValLeu
actggtgattcaattctctgcacttgctcttttctctgagcagGTGCGGGCATTCTATGTGAACGTGCTG

                                  460                                470
AsnGluGluGlnArgLysArgLeuCysGluAsnIleAlaGlyHisLeuLysAspAlaGlnIlePheIleG
AATGAGGAACAGAGGAAACGTCTGTGTGAGAACATTGCCGGCCACCTGAAGGATGCACAAATTTTCATCC

lnLysLysAla
AGAAGAAAGCGgtgagtctttgtaagctgaagggtgtcctct.......about 2.4kb.......ttg

                               480                                490
                        ValLysAsnPheThrGluValHisProAspTyrGlySerHisIleGl
catttattttcctttggccttagGTCAAGAACTTCACTGAGGTCCACCCTGACTACGGGAGCCACATCCA

                    500
nAlaLeuLeuAspLysTyrAsnAlaGluLysProLys
GGCTCTTCTGGACAAGTACAATGCTGAGAAGCCTAAGgtaagctgggagcagcctggccatgcagaggct

gtgtgtgctggg....... about 0.25kb......gaattctgaattattattttcatttgcataca

tattaaaactgagtaaatatcacgttgctgcccatgaggtgattaacctgctcatcttgttcttttaaaa

              510                                520
   AsnAlaIleHisThrPheValGlnSerGlySerHisLeuAlaAlaArgGluLysAlaAsnLeu***
cagAATGCGATTCACACCTTTGTGCAGTCCGGATCTCACTTGGCGGCAAGGGAGAAGGCAAATCTGTGAG

GCCGGGGCCCTGCACCTGTGCAGCGAAGCTTAGCGTTCATCCGTGTAACCCGCTCATCACTGGATGAAGA

TTCTCCTGTGCTAGATGTGCAAATGCAAGCTAGTGGCTTCAAAATAGAGAATCCCACTTTCTATAGCAGA

TTGTGTAACAATTTTAATGCTATTTCCCCAGGGGAAAATGAAGGTTAGGATTTAACAGTCATTTAAAAAA

AAAATTTGTTTTGACGGATGATTGGATTATTCATTTAAAATGATTAGAAGGCAAGTTTCTAGCTAGAAAT

ATGATTTTATTTGACAAAATTTGTTGAAATTATGTATGTTTACATATCACCTCATGGCCTATTATATTAA

AATATGGCTATAAATATATAAAAAGAAAAGATAAAGATGATCTACTCAGAAATTTTTATTTTTCTAAGGT
```

TCTCATAGGAAAAGTACATTTAATACAGCAGTGTCATCAGAAGATAACTTGAGCACCGTCATGGCTTAAT

GTTTATTCCTGATAATAATTGATCAAATTCATTTTTTTCACTGGAGTTACATTAATGTTAATTCAGCACT

GATTTCACAACAGATCAATTTGTAATTGCTTACATTTTTACAATAAATAATCTGTACGTAAGAACAgaga

tggtattttctttctttcgactccatatgtaactgtaaactgctaccagactcttaatttgaacatcatc

attttcagatgtttacccttaaaaatggaaatgccagtatctcgag

Figure 3. Nucleotide sequence of the 13 exons of the human catalase gene including intron-exon boundaries and flanking sequences. Exon sequences are shown in upper case. The predicted amino acid sequence is shown above the nucleotide sequence. The putative translation initiation codon appears in the upper case. The TGA translation termination codon is indicated by ***. Arrows indicate the positions of the GC boxes. CCAAT sequences, the AATAAA polyadenylation signal, and the oligo-dT tract flanking the gene are underlined.

tides 1403 and 1850 of the previously reported sequence should be G and GC instead of A and T respectively. Due to a typographical error, nucleotide 1823 was reported as a C instead of a G. The other differences are clearly polymorphisms. Nucleotides 1409 and 1497 are C and T respectively in the cDNA clones analyzed, and T and C respectively, in the genomic clones sequenced.

The amino acid sequence predicted by the nucleotide sequence of exons 1 and 2 agrees with the partial amino acid sequence of human erythrocyte catalase reported by Schroeder et al (4). A number of ambiguities in the amino acid sequence can now be resolved. Amino acid 3 is the Ser residue reported to replace one of the Asn or Asp residues found in bovine catalase between residues 2-9. The Gln/Glx ambiguity at residue 12 is a Gln. The sequence of residues 30-31 is found to be Ala-Gly. Other ambiguities and corrections to the amino acid sequence have been previously reported (23).

Immediately upstream of the GCT codon for the amino terminal Ala residue of catalase is an ATG codon. This ATG codon is most likely used for translation initiation. The sequence around it, ACGCTATGG, is a close match to the consensus sequence, CC(A/G)CCATG(G), proposed for translation initiation codons (25), differing from it in only two positions. In addition, the sequence of the region upstream from this ATG codon contains no other ATG codons for 368 bp. The ATG found at position -300 is not in the correct translational reading frame.

Furuta et al have recently reported the isolation and nucleotide sequence of rat liver cDNA clones spanning the length of the rat catalase mRNA (26). These clones also code for a Met residue immediately preceding

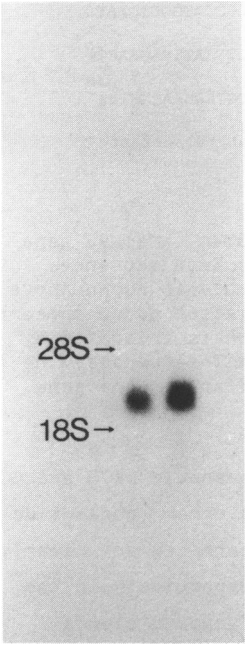

Figure 4. Northern blot of HeLa RNA. 10 and 20 µg of total HeLa RNA was loaded. The 1.2 kb PvuII-HindIII fragment of pCAT41 was used as probe. The positions of the 18S and 28S ribosomal RNA bands are shown.

the Ala residue believed to be the amino terminus of rat liver catalase. These results are consistent with work demonstrating that catalase and all other peroxisomal proteins thus far examined, with the exception of 3-ketoacyl-CoA thiolase, are synthesized at their mature sizes (27-31).

The exact site of transcription initiation has not been determined. S1 nuclease and primer extension analyses have been hampered by the low levels of catalase mRNA in cultured cells and the poor availability of human tissue. Therefore, to define the 5'-end of exon 1, a human liver cDNA library in λgt11 was screened for full length catalase cDNA clones. A fragment carrying a portion of exon 2 was used to isolate pCAT16, a 2.4 kb catalase cDNA that includes 68 bp of 5'-untranslated sequence. The size of the catalase mRNA, as estimated from Northern blots of HeLa RNA (Fig. 4), is approximately 2.4 kb. Therefore pCAT16 is close to the mRNA in size and should contain most of the 5'-untranslated region. The restriction map of pCAT16 is shown in Figure 5.

The rat catalase cDNAs isolated by Furuta et al (26) have a 5'-untranslated region of 83 bp. Residues -83 to -62 of the rat clones (5'-ATTGCCTACCCCGGGTGGAGAC-3') include two blocks of sequence (underlined) identical to sequence found at the 5'-end of pCAT16. Thus the 5'-untrans-

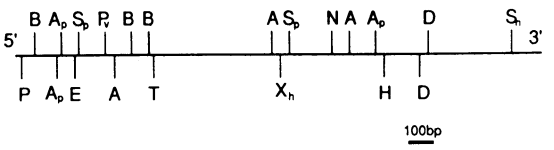

Figure 5. Restriction map of pCAT16. The enzymes represented are A, AvaII; Ap, ApaI; D, DraI; N, NaeI; Pv, PvuII; Sp, SphI; and T, TthIII 1. Other enzymes shown are as in Fig. 2.

lated region of the rat liver catalase cDNA, even though 15 bp longer, actually extends only two nucleotides beyond the end of pCAT16.

The sequence of the 5'-untranslated region of pCAT16 is colinear with the 68 bp of genomic sequence upstream of the translation initiation codon (Fig. 3). While the possibility of another small exon coding for additional 5'-untranslated sequence cannot be excluded the similarity of the 5'-ends of the human and rat catalase cDNA clones makes this less likely. This data allows the tentative assignment of the transcription start site to the region of the 5'-end of pCAT16.

Exon 13 contains the codons for the 21 carboxy terminal amino acid residues, the TGA translation termination signal and the 3'-untranslated region of the catalase mRNA. The position of the translation stop codon is identical to that reported for pCAT41, confirming that human catalase consists of 526 amino acid residues (23). The nucleotide sequence of the rat liver catalase cDNAs determined by Furuta et al also predicts a protein of 526 amino acids (26). The different carboxy terminii for the liver and erythrocyte proteins reported by Schroeder et al (3, 4) are most likely a result of proteolysis during the isolation of the protein. Artifactual proteolytic cleavage of catalase during purification has been reported (32). In addition, Furuta et al have shown that the carboxy terminal amino acid sequence of rat liver catalase, isolated in the presence of a protease inhibitor, matches that predicted by the nucleotide sequence of the cDNA clones (26). Enzyme preparations done in the absence of inhibitors, appear to be shorter and contain different carboxy terminii (26). The 3'-untranslated region of catalase mRNA, as determined from pCAT41, is 628 bp long. A polyadenylation signal, AATAAA, is found 18 bp upstream of the polyadenylation site (33). Polyadenylation occurs at a CA dinucleotide as observed in several other genes (34).

The sequences of the intron-exon junctions are shown in Figure 3. These sequences are a close match to the consensus sequences proposed for

the donor and acceptor sites of introns (35).

The region upstream of most genes transcribed by RNA polymerase II contains a TATA box and a CAAT box found 25-30 bp and about 80 bp respectively, upstream of the transcription initiation site (36, 37). If the tentative assignment of the 5'-end of the gene is correct then the region upstream of the 5'-end of the catalase gene (Fig. 3) does not contain a TATA box. However, the sequence CCAAT is found at positions -97, -126, and -229 relative to the 5'-end of pCAT16. The significance of these sequences, in relation to the catalase gene promoter, given the absence of a downstream TATA box, is unknown.

Another element found in the promoters of some genes is the sequence GGGCGG (GC box). The GC box has been shown to be important in the transcription of the herpes simplex thymidine kinase gene (38) and early and late genes of SV40 (39, 40). The GC box has also been found in the promoters of a number of cellular genes (for review see 41). Sp1, a transcription factor isolated from cultured human cells (42), has been shown to bind at sequences containing GC boxes (41, 43, 44, 45, 46). All Sp1 binding regions contain one or more perfect copies of the sequence GGGCGG (41). A consensus sequence [(G/T)GGGCGG(G/A)G/A)(C/T)] (or its inverse complement) has been proposed for strong Sp1 binding sites (46).

The sequence of the region extending 320 bp upstream of the catalase gene is GC rich (67%) and contains copies of the GC box sequence GGGCGG, or its inverse complement CCGCCC, at positions -71, -281, and -314. Two of the GC boxes found upstream of the catalase gene are found within sequences which closely match the consensus sequence for Sp1 binding sites. The sequence at position -71, GGGGCGGGAC, is a perfect match, while the sequence at position -281, GCCCCGCCCT, differs by only one nucleotide.

Three other genes have been described which have promoters lacking TATA and CAAT boxes and which contain GGGCGG boxes within GC rich upstream sequences. These genes are mouse hypoxanthine phosphoribosyl transferase (47), hamster 3-methylglutaryl coenzyme A reductase (48) and human adenosine deaminase (49). Therefore the GGGCGG motifs found upstream of the human catalase gene may be part of its promoter.

The sequence downstream of the catalase gene is shown in Figure 3. A comparison of the 3'-flanking regions of a number of genes coding for polyadenylated transcripts has revealed a conserved sequence, YGTGTTYY (Y = C or T), found downstream of the polyadenylation site of approximately 67% of the genes examined (50). Many of the genes which lack this octa-

nucleotide sequence contain T rich regions (34, 50). These sequences are thought to be involved in the polyadenylation of mRNA transcripts (50, 51). The catalase gene lacks the YGTGTTYY sequence but has a T rich sequence (10 of 12 residues = T) which begins 10 bp downstream of the polyadenylation site.

ACKNOWLEDGEMENTS

We thank Meri-Jo Anderson for technical assistance. This work was supported by grants from Health and Welfare, Canada and the Medical Research Council of Canada.

[3]Present address: Department of Genetics, Children's Hospital of Eastern Ontario, Ottawa, ONT, K1H 8L1, Canada

*To whom correspondence should be addressed at: Research Institute, Hospital for Sick Children, Toronto, ONT, M5G 1X8, Canada

REFERENCES

1. Deisseroth, A. and Dounce, A. L. (1970) Physiol. Rev., 50, 319-375.
2. Schonbaum, G. R., and Chance, B. (1976) In The Enzymes vol. 13 pp. 363-408 (Boyer, P. D., Ed.), Academic Press, New York.
3. Schroeder, W. A., Shelton, J. R., Shelton, J. B., Robberson, B., Apell G., Fang, R. S. and Bonaventura, J. (1982) Arch. Biochem. Biophys., 214, 397-421.
4. Schroeder, W. A., Shelton, J. R., Shelton, J. B., Apell, G., Evans, L., Bonaventura, J. and Fang, R. S. (1982) Arch. Biochem. Biophys., 214, 422-424.
5. Aebi, H. E. and Wyss, S. R. (1978) In The Metabolic Basis of Inherited Disease, pp. 1792-1807 (Stanbury, J. B., Wyngaarden, J. B. and Fredrickson, D. S., Eds.) McGraw-Hill, New York.
6. Shingu, M., Yoshioka, K., Nobunaga, M. and Yoshida, K. (1985) Inflamm., 9, 309-320.
7. Takahara, S. and Miyamoto, H. (1948) J. Otorhi. Soc. Jpn., 51, 163-164.
8. Ogata, M. and Mizugaki, J. (1979) Hum. Genet., 48, 329-338.
9. Wieacker, P., Mueller, C. R., Mayeroua, A., Grzeschik, K. H. and Ropers, H. H. (1980) Ann. Genet., 23, 73-77.
10. Junien, C., Turleau, C., de Grouchy, J., Said, R., Rethove, M. O., Tenconi, R. and Dufier, J. C. (1980) Ann. Genet. 23, 165-168.
11. Junien, C., Turleau, C., Lenoir, G. M., Phillip, T., Said, R., Despoisse, S., Laurent, C., Rethore, M. O., Kaplan, J. C. and de Grouchy, J. (1983) Cancer Genet. Cytogenet., 10, 51-57.
12. van Heyningen, V., Boyd, P. A., Seawright, A., Fletcher, J. M., Fantes, J. A., Buckton, K. E., Spowart, G., Porteous, D. J., Hill, R. E., Newton, M. S. and Hastie, N. D. (1985) Proc. Natl. Acad. Sci. U. S. A., 82, 8592-8596.
13. Willard, H. F., Smith, K. D., and Sutherland, J. (1983) Nucl. Acids Res., 11, 2017-2038.

14. Maniatis, T., Fritsch, E. F. and Sambrook, J. (1982) Molecular Cloning : A Laboratory Manual. Cold Spring Harbor Laboratory Press, New York.
15. Lawn, R. M., Fritsch, E. F., Parker, R. C., Blake, G. and Maniatis, T. (1978) Cell, 15, 1157-1174.
16. Kwok, S. C. M., Ledley, F. D., DiLella, A. G., Robson, K. J. H. and Woo, S. L. C. (1985) Biochemistry, 24, 556-561.
17. Melton, D. A., Krieg, P. A., Rebagliati, M. R., Maniatis, T., Zinn, K. and Green, M. R. (1984) Nucl. Acids Res., 12, 7035-7056.
18. Frischauf, A. M., Garoff, H. and Lehrach, H. (1980) Nucl. Acids Res., 8, 5541-5549.
19. Sanger, F., Nicklen, S. and Coulson, A. R. (1977) Proc. Natl. Acad. Sci., 74, 5463-5467.
20. Korneluk, R. G., Quan, F. and Gravel, R. A. (1985) Gene, 40, 317-323.
21. Guo, L., Yang, R. C. A. and Wu, R. (1983) Nucl. Acids Res., 11, 5521-5540
22. Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J. and Rutter, W. J. (1979) Biochemistry, 18, 5294-5299.
23. Korneluk, R. G., Quan, F., Lewis, W. H., Guise, K., Willard, H. F., Holmes, M. T. and Gravel, R. A. (1984) J. Biol. Chem., 259, 13819-13823.
24. Quan, F., Korneluk, R. G., MacLeod, H. L., Tsui, L. C. and Gravel, R. A. (1985) Nucl. Acids Res., 13, 8288.
25. Kozak, M. (1984) Nucl. Acids Res., 12, 857-872.
26. Furuta, S., Hayashi, H., Hijikata, M., Miyazawa, S., Osumi, T. and Hashimoto, T. (1986) Proc. Natl. Acad. Sci. U. S. A., 83, 313-317.
27. Furuta, S., Hashimoto, T., Miura, S., Mori, M. and Tatibana, M. (1982) Biochem. Biophys. Res. Commun., 105, 639-646.
28. Robbi, M. and Lazarow, P. B. (1982) J. Biol. Chem., 257, 964-970.
29. Miura, S., Mori, M., Takiguchi, M., Tatibana, M., Furutá, S., Miyazawa, S. and Hashimoto, T. (1984) J. Biol. Chem., 259, 6397-6402.
30. Rachubinski, R. A., Fujiki, Y., Mortensen, R. M. and Lazarow (1984) J. Cell Biol., 99, 2241-2246.
31. Fujiki, Y., Rachubinski, R. A., Mortensen, R. M. and Lazarow, P. B. (1985) Biochem. J., 226, 697-704.
32. Robbi, M. and Lazarow, P. B. (1978) Proc. Natl. Acad. Sci. U. S. A., 75, 4344-4348.
33. Proudfoot, N. J. and Brownlee, G. G. (1976) Nature, 263, 211-214.
34. Birnstiel, M., Busslinger, M. and Strub, K. (1985) Cell, 41, 349-359.
35. Cech, T. R. (1983) Cell, 34, 713-716.
36. Breathnach, R. and Chambon, P. (1981) Ann. Rev. Biochem., 50, 349-383.
37. Benoist, C., O'Hare, K., Breathnach, R. and Chambon, P. (1980) Nucl. Acids Res., 8, 127-142.
38. McKnight, S. L. and Kingsbury, R. (1982) Science, 217, 316-324.
39. Everett, R. D., Baty, D. and Chambon, P. (1983) Nucl. Acids Res., 11, 2447-2464.
40. Hartzell, S. W., Byrne, B. J. and Subramanian, K. N. (1984) Proc. Natl. Acad. Sci. U. S. A., 81, 23-27.
41. Dynan, W. S. and Tjian, R. (1985) Nature, 316, 774-778.
42. Dynan, W. S. and Tjian, R. (1983) Cell, 32, 669-680.
43. Dynan, W. S. and Tjian, R. (1983) Cell, 35, 79-87.
44. Gidoni, D., Dynan, W. S. and Tjian, R. (1984) Nature, 312, 409-413.

45. Jones, K. A. and Tjian, R. (1985) Nature, 317, 179-182.
46. Dynan, W. S., Sazer, S., Tjian, R. and Schimke, R. T. (1986) Nature, 319, 246-248.
47. Melton, D. W., Konecki, D. S., Brennand, J. and Caskey, C. T. (1984) Proc. Natl. Acad. Sci. U. S. A., 81, 2147-2151.
48. Reynolds, G. A., Basu, S. K., Osborne, T. F., Chin, D. J., Gil, G., Brown, M. S., Goldstein, J. L. and Luskey, K. L. (1984) Cell, 38, 275-285.
49. Valerio, D., Duyvesteyn, M. G. C., Dekker, B. M. M., Weeda, G., Berkvens, Th. M., van der Voorn, L., van Ormondt, H. and van der Eb, A. J. (1985) EMBO J., 4, 437-443.
50. McLauchlan, J., Gaffney, D., Whitton, J. L. and Clements, J. B. (1985) Nucl. Acids Res., 13, 1347-1368.