Molecular cloning, mapping, and DNA sequencing techniques have been used to characterize the functional human dihydrofolate reductase (DHFR) gene. The gene is about 30 kilobases in length. Its coding portions are separated into 6 exons, the intron-exon boundaries of which are identical to those of the previously characterized mouse DHFR gene. The 5 introns vary in length from 362 to 12,000 base pairs. The position of the DHFR gene promoter was identified as being shortly upstream from the initiation codon in an in vitro transcription reaction by polymerase II. A DHFR minigene was constructed in a plasmid expression vector by combining a DNA fragment containing exon 1, intron 1, and a small part of exon 2 from the functional gene, with a second DNA fragment containing exons 2-6 from a processed intronless gene, the coding sequences of which are identical to those of the normal locus. Transcription initiation from the DHFR promoter was localized to a position 71 ± 2 base pairs upstream from the initiation codon, both in monkey kidney cells transfected with vectors containing the DHFR minigene, and in human HeLa cells. This single upstream from the initiation codon, both in monkey kidney cells transfected with vectors containing the DHFR minigene, and in human HeLa cells. This single transcription start and the three previously identified polyadenylation sites account for the transcription initiation and the three previously identified polyadenylation sites account for the transcription initiation and the three previously identified polyadenylation sites account for the transcription initiation and the three previously identified polyadenylation sites account for the transcription initiation and the three previously identified polyadenylation sites account for the transcription initiation and the three previously identified polyadenylation sites account for the transcription initiation and the three previously identified polyadenylation sites account for the transcription initiation and the three previously identified polyadenylation sites account for the transcription initiation and the three previously identified polyadenylation sites account for the transcription initiation and the three previously identified polyadenylation sites account for the transcription initiation and the three previously identified polyadenylation sites account for the transcription initiation and the three previously identified polyadenylation sites account for the transcription initiation and the three previously identified polyadenylation sites account for the transcription initiation and the three previously identified polyadenylation sites account for the transcription initiation and the three previously identified polyadenylation sites account for the transcription initiation and the three previously identified polyadenylation sites. The mouse DHFR gene has been extensively characterized by Schimke and his colleagues (9, 10). The coding sequence for DHFR mRNA in the mouse is 561 nucleotides in length and the longest prevalent mRNA species is about 1500 nucleotides; there are six exons in the mouse gene that are distributed over 31 kb of chromosomal DNA. Multiple mouse DHFR mRNA species have been identified (11). Several of these reflect the use of different polyadenylation sites yielding mRNA molecules that differ in the length of their 3′ untranslated region (12). In addition, there is apparent heterogeneity in the initiation site for transcription (13). One start site may be several hundred bp upstream from the initiation codon. The functional mRNA derived from this start site apparently has a splice junction in the 5′ untranslated region. Genomic mapping studies of the Chinese hamster DHFR gene indicate that its general organization is similar to that of the mouse (14) although the hamster DHFR transcriptional unit has not yet been defined.

We have previously cloned and characterized the 3′ half of the normal human DHFR gene (15). The splice junctions that define exons 4, 5, and 6 were shown to be identical in position to those of the corresponding exons of the mouse gene. The 3′ untranslated region of some human DHFR mRNA molecules was found to be 2900 nucleotides in length. Other studies have demonstrated human DHFR mRNA species of lengths of 800, 1000, and 3800 nucleotides (16, 17), and human DHFR cDNA clones derived from mRNAs with the long 3′ untranslated region have been obtained (16, 18).

During our study of the genomic human DHFR gene, we obtained several clones that contained intronless DHFR genes derived from processed mRNA molecules (15). Study of these genes has provided insight into the origin of pseudogenes and, in addition, has provided insights into the structure and function of the normal gene. The coding portion of one of these pseudogenes, hDHFR-ψ₁, was shown to be identical in its sequence to exons 4, 5, and 6 of the normal human DHFR gene and also to the sequence of a cDNA clone (19). The single nucleotide difference originally reported has been found to be due to an error in reading the sequencing gel. The
homology between this processed pseudogene and the functional DHFR gene extends 2.9 kb from the termination codon. Thus, the mRNA from which this pseudogene was derived had the long 3′ untranslated region and presumably was a 3800-nucleotide mRNA molecule. Despite its sequence identity to the normal gene, hDHFR-ϕ, is unlikely to be functional since most processed intronless genes do not include the promoter, and hDHFR-ϕ is not amplified in cells in which methotrexate resistance reflects amplification of the functional gene. We have also characterized two additional processed intronless DHFR pseudogenes (15), one of which is interrupted by a member of the moderately repetitive Alu DNA sequence family. A fourth processed intronless gene has recently been described (18). The other three DHFR pseudogenes are only 85–90% homologous to the coding sequence of the normal gene and contain several inphase termination codons. Study of these genes and cDNA clones (16, 18) has led to identification of two additional polyadenylation sites, one of which is just downstream from the end of the coding sequence.

The purpose of these studies is to characterize the functional DHFR gene and to define the transcriptional unit of this gene. By cloning and Southern blotting, we have determined the length of the locus and determined the relative position of the six exons. A single 5′ start site for DHFR mRNA has been defined and the promoter region of the human gene has been characterized by DNA sequence analysis and functional studies. The sequence of the 5′ flanking, coding, and 3′ untranslated regions of the human DHFR gene has been compared to the corresponding portions of the mouse gene.

MATERIALS AND METHODS

Reagents—The several restriction endonucleases used for these studies were obtained from New England Biolabs (Beverly, MA) or Bethesda Research Laboratories (Rockville, MD) and used under conditions specified by the suppliers. Polynucleotide kinase was obtained from P-L Biochemicals (Milwaukee, WI), bacterial alkaline phosphatase from Worthington Biochemicals (Freehold, NJ), DNA ligase from Bethesda Research Laboratories. Isolations according to the method of Maxam and Gilbert (32, 33). The DNA sequence of both strands has been compared to the corresponding portions of the mouse gene.

Human DHFR Gene

Probes are identified under "Results."

pH79 (Boehringer Mannheim), and human DNA from a methotrexate-resistant HeLa cell line kindly given to us by R. E. Kelly, Baylor College of Medicine. This cell line contains approximately 80 copies of the functional DHFR gene/cell. The genomic DNA was partially restricted with MboI and size-selected for 35–50 kb fragments by sucrose gradient centrifugation. Five μg of human DNA was ligated with 50 μg of vector DNA. The BamHI restricted vector DNA had been treated with bacterial alkaline phosphatase (20) to reduce self ligation. Packaging extracts were prepared as described (29) and the ligated DNA was packaged, transfected into the Escherichia coli strain HB 101, amplified approximately 50-fold, and plated on ampicillin plates as described by Grosved et al. (27, 28). One × 106 colonies were obtained per μg of insert DNA. The library was screened as described with the 342-bp probe. Among the several clones that gave a positive signal was one, cos DHFR-1, that was subsequently shown to contain the 5′ end of the DHFR gene.

DNA Sequencing—DNA fragments for sequencing were isolated by standard methods (20). 5′ or 3′ labeling was performed with polynucleotide kinase (30) or DNA polymerase I (Klenow fragment) (31), respectively. Sequencing was performed according to the method of Maxam and Gilbert (32, 33). The DNA sequence of both strands was determined over the entire region sequenced.

In Vitro Transcription—A HeLa cell extract, prepared according to the method of Manley et al. (34), was obtained from Bethesda Research Laboratories. The in vitro transcription reactions were performed under standard conditions in the presence of [α-32P]UTP. Electrophoresis of glyoxalated RNA in agarose gels followed by autoradiography was as described (34).

Expression Vectors Containing the DHFR Minigene—The 1.8-kb EcoRI fragment that contains the 5′ end of the gene was subcloned into the EcoRI site of ϕR322, yielding the plasmid ϕDHFR-1.8. All subsequent constructions involved EcoRI-SalI fragments obtained by digesting appropriate plasmid DNA either partially or completely with EcoRI and completely with SalI. A 5500-bp EcoRI-SalI fragment from pDHFR-1.8 was ligated to a 4650-bp EcoRI-SalI fragment from pUC21 (15) and cloned to give a plasmid in which the 5′ end of the normal DHFR gene was linked, in proper orientation, to the remainder of the coding sequences from hDHFR-ϕ. A 6450-bp EcoRI-SalI fragment from this plasmid containing the intact minigene was ligated to a 3100-bp EcoRI-SalI fragment from pLTN1 (35) or a 6100-bp EcoRI-SalI fragment from pLTN3 (35). The products of these ligations were cloned and the plasmids pCN8 and pCN9 identified by restriction mapping of isolated plasmid DNA. Plasmid DNA was introduced into the monkey kidney Cos cell line by the calcium phosphate precipitation technique and RNA was harvested 48 h later by described methods (35).

RNA Analysis—Cos cells, transfected with one of the expression vectors, or HeLa cells in which the DHFR gene was amplified, were lysed in guanidinium hydrochloride and the RNA and DNA was purified by CsCl buoyant density centrifugation (35). Primer extension analysis of total cellular RNA was performed by standard methods (36). Total cellular RNA was also annealed to a 5′ end-labeled probe and the probe fragments protected from S1 nuclease were characterized by gel electrophoresis (35, 39).

RESULTS

Map of the Functional DHFR Locus—A clone, cos hDHFR-1, was identified in the cosmid library by the colony hybridization screening procedure of Grosved et al. (27, 28) and subsequently shown to contain the 5′ end of the DHFR gene. Briefly, this was accomplished as follows. Restriction of purified DNA from this clone with EcoRI released fragments of 17.8, 8.7, 8.5, 4.7, 4.3, 2.1, 1.8, 1.7, and 0.77 kb in length. Standard Southern blotting techniques were performed using the probes, human Ex2–6, vector, and mouse DHFR coding sequence. The 1.8-kb EcoRI fragment was shown to contain exon 1 and the 8.5-kb fragment, both exon 2 and vector DNA (Fig. 1A).

The order of the EcoRI fragments and the relative position of the HindIII and BamHI sites upstream from the DHFR gene were originally obtained by study of a recombinant λ phage clone from a methotrexate-resistant cell by Cowan and

3 N. Anagonou, T. Shimada, and A. W. Nienhuis, manuscript in preparation.

T. Shimada and M.-J. Chen, manuscript in preparation.
Goldsmith and verified in our cosmid clone by single and double digestions of cosmid DNA and by redigestion with a second enzyme, of isolated DNA fragments (details available upon request). The map of restriction endonuclease sites for the 3' half of functional DHFR locus was established by analysis of DNA from λDHFR-1 (15), a clone derived from a cell line in which the DHFR gene was not amplified. We used the Southern blot technique to complete the map shown in Fig. 1B. A 350-bp EcoRI-XbaI fragment containing exon 2 and 300 bp of intron II was used as a probe against blots of restricted DNA from cells with a single functional DHFR locus ("single copy") and from cells in which the gene was amplified to approximately 80 copies ("multiple copy"). This probe annealed to a 6.2-kb EcoRI fragment, an approximately 22-kb HindIII fragment, an approximately 32-kb BamHI fragment, and an 8.3-kb PsI fragment in DNA from both cell lines (Fig. 2A). Identically sized fragments were detected with a 200-bp EcoRI-XbaI fragment from λDHFR-1 that contains most of exon 2, all of exon 3 sequence, and 37 bp of exon 4. From these results we concluded that exons 2 and 3 are both in these DNA fragments. Finally, with the exon 2-6 probe, the EcoRI, BamHI, HindIII, and PsI bands present in a blot of multiple copy DNA were either identical in size to those displayed by the other two probes or were those predicted from the map of λDHFR-1 (Figs. 1B and 2B). The same bands were shown in blot of single copy DNA but the several intronless pseudogenes hybridized preferentially to this coding sequence probe giving a very complex pattern of fragments (data not shown). Thus, as summarized in Fig. 1B, the functional human DHFR gene is more than 30 kb in length and has 5 introns varying in length from 362 (Intron I) to as much as 12,000 bp (Intron III).

**DNA Sequence of the Human DHFR Gene**—The DNA sequence of the entire 1.8-kb EcoRI fragment that contains the 5' end of the DHFR gene and that of the entire 2.9-kb untranslated region and 300 bp of the 3' flanking region was determined by the Maxam-Gilbert technique. The strategies used to obtain these sequences are shown in Figs. 3 and 4, respectively. The sequence of exon 2 and 50 bp at the 5' end of intron 2 was determined by sequencing both strands from the 5' end of the 8.5-kb EcoRI fragment from cos λDHFR-1 (Fig. 1A). The strategies used to determine the sequence of exons 4, 5, and 6 from the normal locus have previously been described (15). The presumed sequence of exon 3 was determined by sequencing the coding region of λDHFR-ψ1 (15). These DNA sequences are presented in Fig. 5.

The DNA sequences of the amino acid coding region of exons 1 through 6, as determined from these genomic clones, namely λDHFR-1, cos hDHFR-1, and λDHFR-ψ1, was identical to the DNA sequence of the coding region of a human DHFR cDNA clone (19). The coding region of the human DHFR gene contains 186 codons. The splice junction sequences at the boundaries of exons 1, 2, 4, 5, and 6 conform to the "Chambon" rule, namely there is a GT dinucleotide in each 5' junction sequence and an AG dinucleotide in each 3' junction sequence (40). Furthermore, the sequences at the boundaries are nearly identical to those at the exon-intron boundaries of the mouse DHFR gene and also match the consensus splice junction sequences defined by Mount (41) quite well.

Southern blot experiments with nick-translated genomic DNA showed that the 3' untranslated region of the DHFR gene contains moderately repetitive DNA sequences over most or all of its length (data not shown). A computer search identified homology to the consensus sequence of the Alu family of moderately repetitive DNA sequences (42) at several positions. Those regions at which there was at least an 80% match for more than 10 nucleotides are indicated in Fig. 4. One complete copy and several other small segments of Alu repetitive DNA sequence were identified.

The position of exon 1 in the 1.8-kb fragment was defined by DNA sequencing (Fig. 3). No promoter-like structure was obvious from study of the region upstream from exon 1 so that functional studies were performed to define the location and properties of the human DHFR gene promotor.

**Polymerase II Promoters Upstream from the DHFR Gene**—DNA from a plasmid containing the 1.5-kb EcoRI fragment was restricted with EcoRI and then used as a template for *in vitro* transcription (Fig. 6). Run-off RNA transcripts of 650 and 540 nucleotides, the synthesis of which was inhibited by 0.5 μg/ml of α-amanitin, were observed (Fig. 6C). Both transcripts were derived from the 1.8-kb EcoRI fragment rather than plasmid DNA, since both were observed when the purified 1.8-kb fragment was used as a template (Fig. 6C, lane 5). There is a unique AvaI site in this 1.8-kb EcoRI fragment, 60 bp from the end of the 1.5-kb EcoRI fragment that contains DHFR coding sequences (Fig. 6A). Template DNA that had been doubly restricted with EcoRI and AvaI yielded run-off transcripts of 650 and 480 nucleotides (Fig. 6, B and C) suggesting that these transcripts came from opposite ends of the 1.8-kb fragment as outlined in Fig. 6A. The shorter run-off transcript appeared to begin about 50 to 100 bp upstream from the exon 1 coding sequences in a position expected for a DHFR gene promotor.

**Properties of the DHFR Gene Promoter**—To characterize the DHFR gene promotor further, expression plasmids were constructed to allow its function to be studied in intact cells. Derivation of plasmids that contain the SV40 origin of DNA replication and lack plasmid sequences that may poison DNA replication in monkey kidney cells has previously been described (35). A human DHFR minigene was assembled by using the 1.5-kb EcoRI fragment from the functional locus that contains the 5' end of the gene and the 4.0-kb EcoRI
FIG. 2. Southern blot analysis of the DHFR locus. A, the probe used was a 340-bp fragment containing exon 2 and 300 bp from the 5' end of intron II in an analysis of DNA from single copy (left) and multiple copy cell lines (right), respectively (see text). B, the probe used was “human Ex2-6” in an analysis of DNA from a multiple copy cell line. The lengths (kb) on the left hand side of each panel are the migration position of λ HindIII fragments. E, EcoRI; B, BanHI; P, PstI; H, HindIII.

3.4-

FIG. 3. Restriction endonuclease map (A) and DNA sequencing strategy (B) of the 1.8-kb EcoRI fragment that contains the 5’ end of the DHFR gene. This fragment, derived from cos hDHFR-1, was subcloned into pBR322 (pDHFR-1.8) to provide DNA for sequencing.

FIG. 4. Restriction endonuclease map (A) and DNA sequencing strategy (B) of the 3’ untranslated region of the DHFR gene. DNA for sequencing was obtained from plasmid, pJC8-1, that contains the 8-kb EcoRI fragment from λdhFR-1 (15). The positions of the polyadenylation sites shown have been established by sequencing of processed intronless pseudogenes (15) and/or cDNA clones (16, 18). C, location of homology between the DHFR 3’ untranslated region and the Alu repetitive DNA consensus sequence. The regions of indicated homology were identified by a computer search for segments that match in at least 8 of 10 nucleotides.

A. EXON 2 PROBE

B. EXON 2-6 PROBE

DNA of vector pCN8-Enh+ and pCN9-Enh+ was introduced into monkey kidney cells (Cos cells) by calcium phosphate-mediated DNA transfer and RNA was harvested 48 h later. A synthetic oligonucleotide was used in a primer extension reaction to define the 5’ end of the DHFR gene transcript. The 5’ labeled end of this 17-nucleotide primer begins at the 26th nucleotide from the 5’ end of the exon 2; thus, the extended product contains the primer and the reverse transcript of the 9 nucleotides of exon 2 and the 86 nucleotides of the amino acid coding region of exon 1 (Fig. 8). A primer extended product 183 ± 2 nucleotides in length was obtained with RNA from monkey kidney cells transfected with either pCN8-Enh+ or pCN9-Enh+ DNA but not with RNA from control cells (Fig. 8A). These results indicated that the major RNA transcript from the DHFR minigene in monkey kidney cells has a 5’ untranslated region of 71 ± 2 nucleotides.

7). Analogous mouse DHFR minigenes have been used by others to characterize the mouse DHFR gene (36, 37). Note that the coding sequence of hDHFR-ϕ8 is identical to the coding sequence of the normal locus. Plasmids were studied with (pCN9), or without (pCN8) the 72-bp directly repeated DNA sequence from the SV40 genome (Fig. 9). This repeated DNA sequence acts as an enhancer and is required for the function of some, but not all, promoters.
The figure shows the DNA sequence of the 5' flanking, coding, and 3' untranslated regions of the functional DHFR gene as determined by analysis of DNA from genomic clones. The exon-intron boundaries, the initiator codon (ATG) and the terminator codon (TAA) are indicated. The arrows indicate the position of the transcriptional start site (above) and the three polyadenylation sites (below). Also enclosed is the A-U rich sequence upstream from the transcriptional start site that may be part of the promoter. 

Fig. 5. DNA sequence of the 5' flanking, coding, and 3' untranslated regions of the functional DHFR gene as determined by analysis of DNA from genomic clones. The exon-intron boundaries, the initiator codon (ATG) and the terminator codon (TAA) are indicated. The arrows indicate the position of the transcriptional start site (above) and the three polyadenylation sites (below). Also enclosed is the A-U rich sequence upstream from the transcriptional start site that may be part of the promoter.
FIG. 6. *In vitro* transcription of the EcoRI fragment containing the 5' end of the DHFR gene. A, map of the plasmid used as template DNA. RI, EcoRI; Pst, PstI; Pvu, PvuII; Ava, AvaI. R and P are the inferred positions of the DHFR promoter and of a promoter on the opposite strand, respectively. B, *in vitro* transcription of restricted plasmid DNA. Lane 1, template DNA restricted with EcoRI. Lane 2, template DNA restricted with EcoRI and AvaI. C, *in vitro* transcription of restricted plasmid DNA and of the isolated 1.8-kb EcoRI fragment. Lanes 1 and 2, template plasmid DNA restricted with EcoRI. Lanes 3 and 4, template DNA restricted with EcoRI and AvaI. Actinomycin (0.5 μg/ml) was added to the reaction for lanes 2 and 4. Lane 5, *in vitro* transcription of the isolated 1.8-kb EcoRI fragment.

Furthermore, the DHFR promoter does not require an exogenous enhancer element to function in this assay system, although there may be a quantitative effect of the enhancer on promoter function (Figs. 8A and 9A).

The 5' end of DHFR mRNA from human cells was found to be identical to that transcribed from the DHFR minigene as shown by an identical analysis of RNA isolated from methotrexate-resistant HeLa cells that contain approximately 80 copies of the DHFR gene (Fig. 8B). In other experiments, these results were confirmed using an EcoRI-XbaI fragment derived from the coding sequences of hDHFR-α. It was 5'-labeled at its XbaI site and used as primer. Furthermore, a very small amount of this primer extended product was identified in the reaction with control monkey kidney cell RNA (data not shown) suggesting that the normal, nonamplified primate DHFR gene is also transcribed beginning at this site.

5'-nuclease mapping was used to further define the end of DHFR mRNA molecules and to specifically eliminate the possibility of splicing within the 5' untranslated part of the mRNA. A 202-bp EcoRII-HpaII fragment, uniquely labeled at the 5' position of the EcoRII site, was used as a probe (Fig. 9). A cluster of bands between 141 and 144 nucleotides in length were protected from 5'-nuclease digestion by RNA from the HeLa cell line with the amplified DHFR gene, and by RNA from Cos cells transfected with pCN8-Enh- or pCn9-Enh+ DNA (Fig. 9). A precise match between these bands and the DNA sequence of the 5' end of the DHFR gene was made by running the products of the “G” sequencing reaction, derived from the probe, in the adjacent lane. These results show that most DHFR mRNA molecules are initiated within a 4-nucleotide segment of the gene. The sequence of this portion of the gene is shown at the bottom of Fig. 9. The start site(s) identified is within a cluster of adenines; each of the three adenines could serve as a Cap site. Approximately 30 bp upstream is an A-T rich 5-nucleotide segment that could be equivalent to the TATA box that apparently functions to phase the start of RNA transcription in many promoters.

FIG. 7. Map of plasmids used to study function of the DHFR promoter. Both plasmids contain the 1.8-kb EcoRI fragment that contains the 5' end of the functional DHFR gene and the 4.0-kb EcoRI fragment from the pseudogene in XDHFR. pCN8 was derived from pLTN1 (35). This plasmid contains the SV40 origin of replication but lacks a complete copy of the enhancer element. pCN9 is derived from the plasmid pLTN3 (15) and contains both copies of the 72-bp enhancer sequence from the SV40 genome. R, EcoRI; Pvu, PvuII. The solid bar indicates the position of the DHFR coding sequence and the hatched bar, the position of the long 3' untranslated region. Intron 1 remains in the plasmid so that exons 1 and 2 are separated as they are in genomic DNA.

Homology between the Human and Mouse DHFR Genes—Human and mouse DHFR differ in only 21 of 186 amino acids. This similarity of protein structure reflects an 89% homology in the DNA coding sequences of the genes. Our work has shown that the exon-intron boundaries of the two genes are precisely the same. Given this fairly strict conser-
the two genes, we decided to directly compare the homology in the coding segments and general organization of the two genes, which has been published (10, 12). The sequence of the upstream region is highly G-C rich in both mouse and human DNA and extends only 100 nucleotides beyond the terminator codon (Fig. 12). The entire region between the mouse and human 5′ flanking sequences extends beyond this repeat extends approximately 270 nucleotides upstream from the human ATG codon.

Fig. 9. S1 nuclease analysis of DHFR mRNA. The probe used was a 5′ labeled EcoRII-Hpa II fragment uniquely labeled on the EcoRII site. A, analysis of RNA isolated from Cos cells transfected with pCN9-enh+ (lanes 1 and 2), pCN8-enh− (lanes 3 and 4), or with a vector containing a globin gene as a control (lanes 5 and 6). The odd numbered lanes were derived from analysis of 30 μg of RNA and the even numbered lanes, 10 μg. B, comparison of the primer extension analysis of RNA extracted from Cos cells transfected with pCN9 (lane 1) and RNA from the HeLa cell line that contains approximately 80 copies of the DHFR gene (lanes 2 and 3). Lane 1, 30 μg; lane 2, 30 μg; lane 3, 10 μg. Marker lanes (M) contain the 5′ labeled HaeIII φ174 fragments. The diagram below shows an outline of the gene, mRNA molecule, and the primer extended product.

A matrix analysis was performed (43) at a stringency of 4 of 5 nucleotides, beginning at the initiation codon and extending upstream (Fig. 10, top) and from the terminator codon and extending downstream (Fig. 10, bottom). A direct comparison of the human and mouse 5′ flanking sequences and the 5′ end of the coding sequences is shown in Fig. 11. The 5′ untranslated regions have limited homology except for a 20-nucleotide segment near the human start site(s). Immediately upstream from this position there is a quadruply repeated 48-bp segment in the mouse genome that only occurs once in human DNA as shown by both the matrix analysis (Fig. 10) and the direct sequence comparison (Fig. 11). In Fig. 11, the single copy in human DNA is aligned to the copy in mouse DNA to which it has the best match. Between the single copy in human DNA and the start site(s) for transcription, there is an 11-bp sequence that is a perfect match for a portion of the mouse repeat. The homology between the mouse and human 5′ flanking sequences beyond this repeat extends upstream for another 270 nucleotides (Fig. 10). This entire region is highly G-C rich in both mouse and human DNA (Fig. 11). There is another 110 bp of homology, identified by the matrix analysis (Fig. 10), that begins approximately 700 nucleotides upstream from the human ATG codon.

The homology between the human and mouse 3′ untranslated region extends only 100 nucleotides beyond the terminator codon (Fig. 12). The entire 2900 bp of the human 3′ untranslated region was compared, by this matrix analysis, to the 950 bp of the mouse 3′ untranslated region and no other significant homology was found. The homology ends at a
Human DHFR Gene

Fig. 10. Matrix comparison of the human and mouse DHFR 5' flanking sequences (top) and the human and mouse 3' untranslated sequences (bottom). The entire 2900 bp of human 3' untranslated sequence was compared to the mouse 3' untranslated sequence; only the results with the first 650 bp of the human sequence are shown. No additional homology was found. A dot on the matrix represents a 4 for 5 nucleotide match.

point very close to the polyadenylation site for one of the several mouse DHFR mRNA species (Fig. 4) (12). A recently sequenced human intronless DHFR gene (8) and a human DHFR cDNA clone (18) end near this position indicating that some human DHFR mRNA molecules use this polyadenylation site. Although the sequences upstream are AT rich in the human gene, the ATAA sequence identified as a potential polyadenylation signal in the mouse gene (12) is not present in the human gene.

DISCUSSION

Our molecular cloning studies have led to characterization of the human DHFR gene and also provided the promoter region of the gene for detailed functional characterization. The human DHFR gene is about 30 kb in length in contrast to the estimated 31-kb size of the mouse DHFR gene (10) and the 26-kb size of the Chinese hamster gene (14). The general organization of the human gene was found to be identical in DNA from both normal cells and from cells in which the gene was amplified approximately 80-fold. Thus, the size of the amplification unit is considerably larger than the gene, a result also observed for the amplified DHFR gene in rodent cell lines (3–5, 14).

The start site for initiation of transcription of the human DHFR gene was uniquely identified by a combination of primer extension analysis and S1 nuclease mapping. The primer extension analysis provisionally localized the start site for transcription to a position 71 ± 2 nucleotides upstream from the initiator codon (Fig. 8). No evidence for additional start sites further upstream was found. The S1 nuclease analysis shown in Fig. 9 eliminate the possibility of splicing within the 5' untranslated region and more precisely localized the start site with respect to the DNA sequence. Our results thus show that the vast majority of mRNA DHFR molecules are initiated within the region of 3–4 nucleotides identified in Fig. 9. Each of the three adenines could serve as a Cap site. Minor bands consistent with start sites 10–30 nucleotides downstream from this position were seen with both the primer extension and S1 analysis; alternatively, these bands could reflect methodological artifacts. That most human DHFR mRNA molecules begin in a single short segment contrasts
Human DHFR Gene

3941

Fig. 11. Comparison of the human and mouse DNA sequences of the 5' flanking region of exon 1, a portion of intron 1, and the first part of exon 2. The boxed segments of the mouse sequence are the quadruply repeated 48-base pair segment (10). The corresponding part of the human 5' flanking sequence is aligned to the copy of the repeat with which it has the best match. The arrow on the human sequence indicates the position of the first A in the region where initiation of transcription occurs. The arrow on the mouse sequence indicates the most extreme 5' nucleotide included in a DHFR cDNA clone (26). The underlined 14-bp segment in intron I of the human gene is a 11/14 match to the consensus enhancer core sequence (42).

Fig. 12. A comparison of the region of homology in exon 6 and in the 3' untranslated region of the mouse and human DHFR genes. The arrows indicate the position of a polyadenylation site of a mouse DHFR gene (12) and the 3' end of a processed human DHFR pseudogene, respectively. Note that the poly(A) tract on the human cDNA clone sequenced by Masters et al. (18), begins 5 nucleotides upstream at the overlined nucleotide. The sequence we have determined differs in two positions from that obtained by Masters et al. (18) in sequencing a cDNA clone; the nucleotides found by these workers are given above our sequence.

to results obtained in study of the mouse DHFR gene transcriptional unit (13). Multiple transcriptional start sites, including one several hundred bp upstream from the initiation codon, have been identified for the mouse gene. As discussed in detail below, structural differences in the promoter regions of the human and mouse DHFR genes could account for these divergent results.

The human DHFR gene promoter was shown to function, when transfected into monkey kidney cells, in vectors that either contained or lacked the 72-bp directly repeated enhancer sequence from the SV40 genome (Figs. 7–9). Enhancers are short DNA sequences that function independent of orientation and, within limits, of position to increase the efficiency of promoter function (44). These elements, first identified in viral genomes, have recently been defined in animal cell DNA (45–48). Specifically, the first intron of the immunoglobulin genes contains a tissue specific enhancer that activates the rearranged variable region gene promoter. These results prompted us to search the DHFR gene region for homology to the enhancer core sequence (44). Several homologous segments were found; that with the best match is in intron I (Fig. 11). Functional studies are in progress to determine whether intron 1 has enhancer function when tested with other promoters or to determine whether autonomous
function in enhancer minus vectors is an intrinsic property of the DHFR gene promoter.

DHFR mRNA molecules of 800, 1000, and 3800 nucleotides have been identified by Northern blot analysis of RNA from human cell lines (16, 17). The 800-nucleotide species arises from initiation at the unique transcriptional start site and termination at the polyadenylation site shown in Fig. 12. The 1000-nucleotide species is derived by use of a second polyadenylation site as shown in Figs. 4A and 5. Two cDNA clones sequenced by Masters et al. (18) have been shown to have a poly(A) tract beginning at the position. The 3800-nucleotide species arises from initiation at this unique transcriptional start site and polyadenylation at a position 2834 nucleotides downstream from the terminator codon (Fig. 4), first identified as a polyadenylation site by our study of processed intronless DHFR genes (15). Thus, we believe that the heterogeneity in the size of DHFR mRNA can be explained by use of a single transcriptional start site and three polyadenylation sites. Only the last polyadenylation site is preceded by the signal AATAAA. The 3800-nucleotide mRNA is the most abundant of the three species (16).

A comparison of the mouse and human DHFR genes allows tentative identification of the functionally important regions of these genes. The coding sequence conservation presumably reflects constraints in the protein sequence for optimal enzyme function. Beyond these constraints, there has been very significant divergence between the noncoding regions of two genes. The intron lengths differ by 1.5-6.0-fold and intron sequences begin to diverge immediately beyond the splice signals (Fig. 11). Only one polyadenylation site appears to be common (Fig. 12). As noted by Setzer et al. (12), this site lacks the classic signal, AATAAA, perhaps accounting for the downstream regions are functionally irrelevant and perhaps not part of the DHFR ancestral gene.

A direct comparison of the 5′ flanking regions of the two genes suggests a complex evolutionary history. The human genome contains only one complete copy of the 48-bp sequence that is quadruply repeated in mouse DNA (10). Human DNA does contain an additional 11-bp stretch that is exactly homologous to part of the repeat suggesting that the human gene is derived from an ancestral gene that had at least two copies of the repeat. If so, deletion of a portion of one has placed the AT-rich block in the repeat, near a region that can function as a Cap site for human DHFR gene transcripts (Fig. 12). Beyond the repeat sequence, there is a stretch of substantial homology for approximately 270 bp, suggesting that this region may be important for transcriptional control.

The homology that occurs 700-800 nucleotides upstream from the human initiator codon is unexplained. One of the several mouse DHFR mRNA species is apparently initiated in this region (13) so that the homology could reflect an analogous functional region for the ancestral human gene. Note also that this homology occurs in the region in which there is polymerase II promoter activity in vitro on the DNA strand opposite the human DHFR gene (Fig. 6). An attempt to identify a transcriptional unit by a computer search for open reading frames bounded by RNA splice signals in this region was unsuccessful although these studies do not completely rule out the presence of an active gene.

Acknowledgments—We are grateful to R. E. Kellemes for providing the methotrexate-resistant human cell line, to M. McGrogan, C. Simonsen, and R. T. Schimke for allowing us to analyze their unpublished sequence data, to J. Owens for assistance in performing the matrix analysis, to Anne Baur for excellent assistance, and to Linda Garza for preparation of this manuscript.

REFERENCES

1. Chabner, B. (1982) in Pharmacological Principles of Cancer Treatment (Chabner B., ed) pp. 229-255, W. B. Saunders Co. Philadelphia
2. Alt, F. W., Kellemes, R. E., and Schimke, R. T. (1976) J. Biol. Chem. 251, 3063-3074
3. Nurnberg, J. H., Kaufman, R. J., Schimke, R. T., Uraluh, G., and Chasin, L. A. (1978) Proc. Natl. Acad. Sci. U. S. A. 75, 5553-5556
4. Kaufman, R. J., Brown, P. C., and Schimke, R. T. (1979) Proc. Natl. Acad. Sci. U. S. A. 76, 5689-5673
5. Kellemes, R. E., Morhenn, V. B., Pfendt, E. A., Alt, F. W., and Schimke, R. T. (1979) J. Biol. Chem. 254, 309-318
6. Schimke, R. T., Kaufman, R. J., Alt, F. W., and Kellemes, R. E. (1978) Science (Wash. D. C.) 202, 1051-1055
7. Cowan, K. H., Goldsmith, M. E., Levine, R. M., Atken, S. C., Douglass, E., Clendeninn, N., Nienhuis, A. W., and Lippman, M. E. (1982) J. Biol. Chem. 258, 15079-15086
8. Curt, G. A., Carney, D. N., Cowan, K. H., Jolivet, J., Bailey, B. D., Drake, J. C., Kao-shan, C. S., Minna, J. D., and Chabner, B. (1983) N. Engl. J. Med. 308, 192-202
9. Nurnberg, J. H., Kaufman, R. J., Chang, A. C. Y., Cohen, S. N., and Schimke, R. T. (1981) Cell 29, 355-384
10. Crouse, G. F., Simonsen, C. C., McEwan, R. N., and Schimke, R. T. (1982) J. Biol. Chem. 257, 7887-7897
11. Setzer, D. R., McGrogan, M., Nurnberg, J. H., and Schimke, R. T. (1980) Cell 22, 361-370
12. Setzer, D. R., McGrogan, M., and Schimke, R. T. (1982) J. Biol. Chem. 257, 5143-5147
13. McGrogan, M., Simonsen, C., and Schimke, R. T. (1982) J. Cell. Biochem. Suppl. 6, 862 (abstr.)
14. Carothers, A. M., Uraluh, G., Ellis, N., and Chasin, L. A. (1983) Nucleic Acids Res. 11, 1997-1992
15. Chen, M. J., Shimada, T., Moulton, A. D., Harrison, M., and Nienhuis, A. W. (1982) Proc. Natl. Acad. Sci. U. S. A. 79, 7435-7439
16. Morandi, C., Masters, J. N., Mottes, N., and Attardi, G. (1982) J. Mol. Biol. 156, 583-607
17. Medina, W. D., Srinatkaudada, S., Engel, D., Sheer, D. I., Moroson, B. A., Dybe, S., Cashmore, A. R., and Bertino, J. B. (1982) Clin. Res. 30, 534 A (abstr.)
18. Masters, J. N., Yank, J. K., Celini, A., and Attardi, G. (1983) J. Mol. Biol. 167, 51-66
19. Masters, J. N., and Attardi, G. (1983) Gene 21, 59-63
20. Maniatis, T., Fritsch, E. F., and Sambrook, J. (1982) Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory, New York
21. Kretschmer, P. J., Coon, H. C., Davis, A., Harrison, M., and Nienhuis, A. W. (1981) J. Biol. Chem. 256, 1975-1982
22. Southern, E. M. (1975) J. Mol. Biol. 98, 503-517
23. Smith, H. D., and Bersteil, M. L. (1976) Nucleic Acids Res. 3, 2387-2396
24. Maniatis, T., Jeffrey, A., and Kleid, D. G. (1975) Proc. Natl. Acad. Sci. U. S. A. 72, 1184-1188
25. Rigby, P. W. J., Dieckmann, M., Rhodes, C., and Berg, P. (1977) J. Mol. Biol. 239, 222-236
26. Chang, A. C. Y., Nurnberg, J. H., Kaufman, R. J., Erlich, H. A., Schimke, R. T., and Cohen, S. N. (1978) Nature (Lond.) 275, 617-624
27. Grosfeld, F. G., Dahl, H. H. M., de Boer, E., and Flavell, R. A. (1981) Gene 10, 227-237
28. Grosfeld, F. G., Lund, T., Murray, A. L., Dahl, H. H. M., and Flavell, R. A. (1982) Nucleic Acids Res. 10, 6715-6732
29. Hoffe, B. (1979) Methods Enzymol. 65, 488
30. Richardson, C. C. (1965) Proc. Natl. Acad. Sci. U. S. A. 54, 158-165
31. Donelson, J. E., and Wu, R. (1972) J. Biol. Chem. 247, 484-494
32. Maxam, A. M., and Gilbert, W. (1977) Proc. Natl. Acad. Sci. U. S. A. 74, 560-564
33. Maxam, A. M., and Gilbert, W. (1980) Methods Enzymol. 65, 499-559
34. Manley, J. L., Fire, A., Cano, A., Sharp, P. A., and Gefer, M. L.
Human DHFR Gene

(1980) Proc. Natl. Acad. Sci. U. S. A. 77, 3855-3859
35. Humphries, R. K., Levy, T., Turner, P., Moulton, A. D., and Nienhuis, A. W. (1982) Cell 30, 173-183
36. Gasser, G. S., Simonsen, C. C., Schilling, J. W., and Schimke, R. T. (1982) Proc. Natl. Acad. Sci. U. S. A. 79, 6522-6526
37. Crouse, G. F., McEwan, R. N., and Pearson, M. L. (1983) Mol. Cell. Biol. 3, 257-266
38. Treisman, R., Proudfect, N. J., Shander, M., and Maniatis, T. (1982) Cell 30, 903-911
39. Weaver, R. F., and Weissman, C. (1979) Nucleic Acids Res. 7, 1175-1193
40. Breathnach, R., Benoist, C., O'Hare, K., Gammon, F., and Cham-}

41. Mount, S. M. (1982) Nucleic Acids Res. 10, 459-471
42. Schmid, C. W., and Jelinek, W. R. (1982) Science (Wash. D. C.) 216, 1065-1070
43. Maizel, J. V., and Lenk, R. P. (1981) Proc. Natl. Acad. Sci. U. S. A. 78, 7665-7669
44. Khoury, G., and Gruss, P. (1983) Cell 33, 313-314
45. Gillies, S. D., Morrison, S. L., Oi, V. T., and Tonegawa, S. (1983) Cell 33, 717-728
46. Banerji, J., Olson, L., and Schaffner, W. (1983) Cell 33, 729-740
47. Queen, C., and Baltimore, D. (1983) Cell 33, 741-748
48. Eronmine, L., Kuehl, M., Weir, L., Leder, P., and Max, E. E. (1983) Nature (Lond.) 304, 447-452
The functional human dihydrofolate reductase gene.
M J Chen, T Shimada, A D Moulton, A Cline, R K Humphries, J Maizel and A W Nienhuis

J. Biol. Chem. 1984, 259:3933-3943.