MOLECULAR AND CELLULAR BIOLOGY, Mar. 1986, p. 768–778
0270-7306/86/030768-11$02.00/0

Vol. 6, No. 3

Structural Characterization of a Heterogeneous Family of Rat Brain mRNAs

ANN-PING TSOU,[1] CARY LAI,[1] PATRIA DANIELSON,[1] DANIEL J. NOONAN,[2] AND J. GREGOR SUTCLIFFE[1]*

Department of Molecular Biology[1] and Department of Immunology,[2] Scripps Clinic and Research Foundation, La Jolla, California 92037

Received 18 July 1985/Accepted 3 December 1985

A large heterogeneous family of RNAs derived from a single rat gene contains members that differ from each other at one or more of three positions. Their 5′ ends are nested and transcription can begin at 22 or more sites covering 265 nucleotides. Many of the 5′ ends are detectable only in brain RNAs, and even 5′ ends common with other tissues appear with different absolute and relative abundances in brain RNA. The central portions of the RNAs are of two forms, differing only by the presence or absence of 17 nucleotides; these forms are probably produced by alternative splicing. Polyadenylation occurrs at either of two sites. This complicated family of 88 RNAs encodes two novel putative proteins that differ at their C termini.

Understanding the development of higher eucaryotic organisms requires an appreciation of tissue-specific gene expression. Polymorphic families of mRNAs and proteins may be expected to provide special insights into the molecular properties that are important for cellular functions as well as the properties that prescribe particular molecular specificities. Such polymorphic molecules are likely to be involved in linking the processes that occur in every cell type with the properties that are exclusive to certain tissues. Because of the inherent diversity of neuronal phenotypes, the brain is an organ that one might expect to be particularly rich in polymorphic molecular species. Some of these may be involved in reception of chemical signals, others in making the myriad cell contacts, others in storing information, and some in altering molecular activities and half-lives. We have studied brain mRNAs, in part, to find members of polymorphic families. By analyzing a collection of almost 200 cDNA clones of rat brain poly(A)$^+$ RNA by RNA (Northern) blot analysis, we have arrived at some general properties of rat brain messages (24). Their most important property is that at least 40% and possibly as much as 56% of the mass of brain mRNA is not expressed detectably in liver or kidney and thus can be operationally defined as brain specific (class III mRNAs). The 1.4×10^8 nucleotides of mammalian brain poly(A)$^+$ RNA (5) encode approximately 30,000 brain message species averaging 5,000 nucleotides in length, most of which are rare in the whole brain but are probably enriched in particular cell types (24). More than 10% of brain transcripts under relatively high (95%) stringency conditions hybridize to multiple species of poly(A)$^+$ brain mRNAs. Such mRNAs might have functional polymorphisms and thus are worthy of in-depth studies.

One brain cDNA clone in our collection was especially interesting because of its pattern of hybridization on Northern blots to brain, liver, and kidney RNAs. While most clones in our initial study hybridized to one or more discrete RNA species, clone p0-44 (Fig. 1) hybridized to a diffuse target in brain, liver, and kidney cytoplasmic poly(A)$^+$ RNAs (24). As shown in the Northern blot (Fig. 1), the transcripts sharing sequence homology with p0-44 were distributed from 800 to 1,200 nucleotides. This group of transcripts appeared to be present at higher abundance (0.1% of mass) in liver and kidney than in the brain (0.02%), hence we originally classified these RNAs as class II (present at different levels in various tissues). More interestingly, the diffuse groups of targets had slower gel mobility in the brain RNA samples, suggesting the possibility of a tissue-specific mRNA polymorphism. In this report we define the structures and tissue distribution specificities of the mRNAs of the 0-44 family, and show that members of the family are products of a single gene but can differ from one another at three positions along a common backbone. This family is the most extreme example to date of diverse expression from a single gene, and the mechanism of expression demonstrates a way of generating molecular polymorphisms within the central nervous system.

MATERIALS AND METHODS

RNA isolation. Cytoplasmic RNA was isolated from brains, livers, or kidneys of both adult male Sprague-Dawley rats (Charles River Breeding Laboratories, Inc., Wilmington, Mass.) and adult BALB/c mice (Scripps Clinic) by a phenol-chloroform (1:1) extraction procedure as described previously (35) and enriched for poly(A)$^+$ RNA by passage over an oligo(dT)-cellulose column (4). For developmental studies, brain samples were obtained from 14-, 18-, and 20-day-old embryos and from 1-, 5-, 10-, 15-, 20-, 25-, and 30-day-old neonates and prepared as described above. Poly(A)$^+$ RNA was prepared from the whole heads of 14-day-old embryos because of difficulty in accurate dissection of the brain at that stage.

Blot analyses. Poly(A)$^+$ RNA samples (2 μg) were fractionated by electrophoresis on 1.2% agarose–1 M formaldehyde gels and transferred to nitrocellulose (39). The prehybridization and hybridization of RNA blots were done as described previously (24), except that 56% formamide and a 37°C incubation temperature were chosen for convenience instead of 50% formamide and 42°C. cDNA inserts (100 ng) were labeled with ^{32}P by nick translation (33) to specific activities of 5×10^8 to 10×10^8 cpm/μg. Blots were washed as described (24). Uncut φX174 and *Hae*III-digested φX174 DNA were used as size markers. Genomic DNA blotting (38) was carried out as described previously (24).

Cloning. Cloning into the Okayama and Berg cDNA expression vector was performed essentially as described previously (27). Briefly, rat brain cytoplasmic poly(A)$^+$

* Corresponding author.

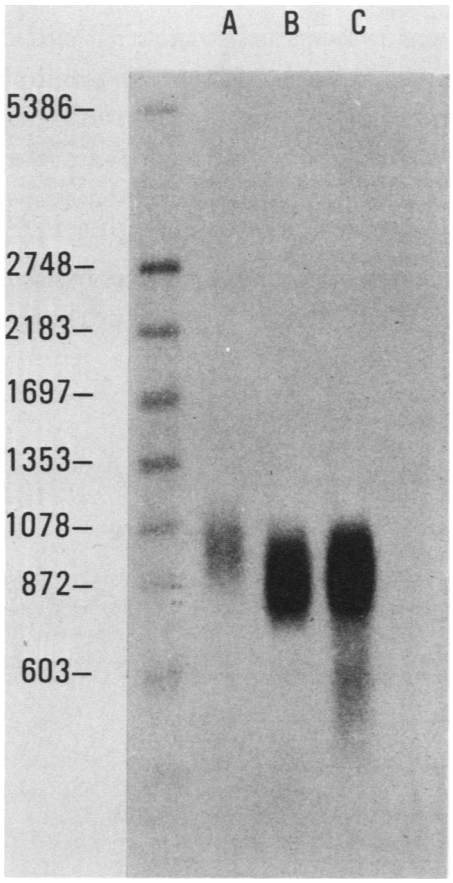

FIG. 1. Heterogeneity of 0-44 RNA family detected by Northern blot hybridization. Fractions (2 μg) of rat brain (A), liver (B), and kidney (C) poly(A)$^+$ RNA were separated by electrophoresis on 1.2% agarose-formaldehyde gels, blotted to nitrocellulose, and hybridized with the ^{32}P-labeled cDNA insert of p0-44. DNA size markers are shown on the left.

RNA was annealed to a deoxyribosylthymine-tailed pcDV1 vector primer fragment from which a complementary first strand was synthesized. Following deoxycytosine tailing and *Hin*dIII digestion the plasmid hybrid was cyclized with the deoxyguanidine-tailed pL1 linker fragment. Synthesis of the second strand was performed in the presence of RNase H, DNA polymerase I, and *Escherichia coli* ligase; and the resulting cDNA constructs were subsequently used in $CaCl_2$ transformation of *E. coli* MC1061. Genomic clones were isolated by plaque hybridization from the library of Sargent et al. (34).

DNA sequencing. Sequence analysis of 5'-end-labeled re-?striction fragments from the cDNA clones was accomplished by a modified version of the partial chemical degradation method (11, 17). Sequence analysis of the genomic fragment was performed by both the chain termination (9) and the chemical methods. All novel sequences were determined on both strands. Some redundant sequences were determined only once.

Primer extension analysis. The procedure described by Giorgi et al. (14) was used. Briefly, 5'-end-labeled primer fragments were combined with 8 μg of cytoplasmic poly(A)$^+$ RNAs from different tissues in a total of 4 μl of H_2O. The samples were heated for 2 min at 90°C, annealed for 10 min at 65°C in 200 mM NaCl, and then extended in a 30-μl reaction with reverse transcriptase. The reverse transcription was carried out in 50 mM Tris hydrochloride (pH 8.2)–10 mM $MgCl_2$–50 mM NaCl–5 mM dithiothreitol–500 μM to 1.2 mM each of the four dNTPs–50 μg of actinomycin D per ml–600 U of avian myeloblastosis virus reverse transcriptase (Life Sciences Inc., St. Petersburg, Fla.) per ml. The reactions were first incubated for 10 min at 37°C, and then for 1 h at 43°C. The products were analyzed by electrophoresis on polyacrylamide–7 M urea sequencing gels.

S1 nuclease analysis. Restriction fragments were isolated from clones p7-4 and p15-2 and 5' end labeled with [γ-^{32}P]ATP using polynucleotide kinase (P-L Biochemicals, Inc., Milwaukee, Wis.). The DNA probes were denatured and hybridized to poly(A)$^+$ RNA from different sources under the conditions described by Casey and Davidson (12). Approximately 1×10^5 to 2×10^5 cpm of labeled fragment were hybridized to 2 μg of cytoplasmic poly(A)$^+$ RNA in 30 μl of 80% formamide–0.4 M NaCl–0.04 M 1,4-piperazine-*N*,*N*'-bis(2-ethanesulfonic acid) (PIPES; pH 6.4)–0.001 M EDTA at 40 or 50°C for 3 to 5 h. The hybridized samples were diluted 10-fold with S1 nuclease buffer (0.28 M NaCl, 0.05 M sodium acetate [pH 4.5], 4.5 mM $ZnSO_4$) and digested with S1 nuclease (100 U/ml; P-L Biochemicals) at 37°C for 30 min by the method of Berk and Sharp (8). Samples were analyzed on urea-polyacrylamide sequencing gels.

RESULTS

To understand fully the structure of the 0-44 transcripts we isolated additional cDNA clones by constructing a cDNA library with rat brain poly(A)$^+$ RNA and the plasmid vector pcD and cloning techniques described by Okayama and Berg (27). This procedure produces a high proportion of full-length copies of mRNAs. Screening of this library with the cDNA insert of p0-44 yielded 68 positive clones from 2×10^5 transformants, which is a frequency (0.03%) that corresponds approximately to the previously estimated abundance of this message in the brain (0.02%).

Nucleotide sequences of 0-44 homologous cDNAs. Initially, we determined the nucleotide sequence of the p0-44 cDNA insert. Subsequently, eight cDNA clones with longer 5' sequences, as determined by restriction enzyme analysis, were sequenced. We constructed a composite nucleotide sequence spanning from clone p7-4, which has the most 5' end information, to clone p5-3, which has the most 3' end information. This composite and the extent of sequence overlap in the 9 cDNA clones is depicted in Fig. 2. We chose to number the sequences based on this composite information. In the regions in which the cDNA clones overlapped, all sequences were in complete agreement. The single long open reading frame (ORF) found in the nucleotide sequence is located between positions 329 and 709; the ORF codes for a putative protein of 127 amino acids (about 14,000 daltons) and is translated above the composite DNA sequence.

Three kinds of structural differences between the cDNA clones were identified. The most notable feature was that the putative 5' untranslated region of this group of clones varied in length. Results of the experiments discussed below show that each of these clones could correspond to a full-length copy of a different mRNA. Clone p7-4 contained the longest untranslated sequence of 328 nucleotides (Fig. 2B). Clones with shorter 5' sequences started at different positions, but the nested sequences were identical. The second characteristic showing a difference in the sequences was an insertion in the 3' region of the ORF which was observed in three

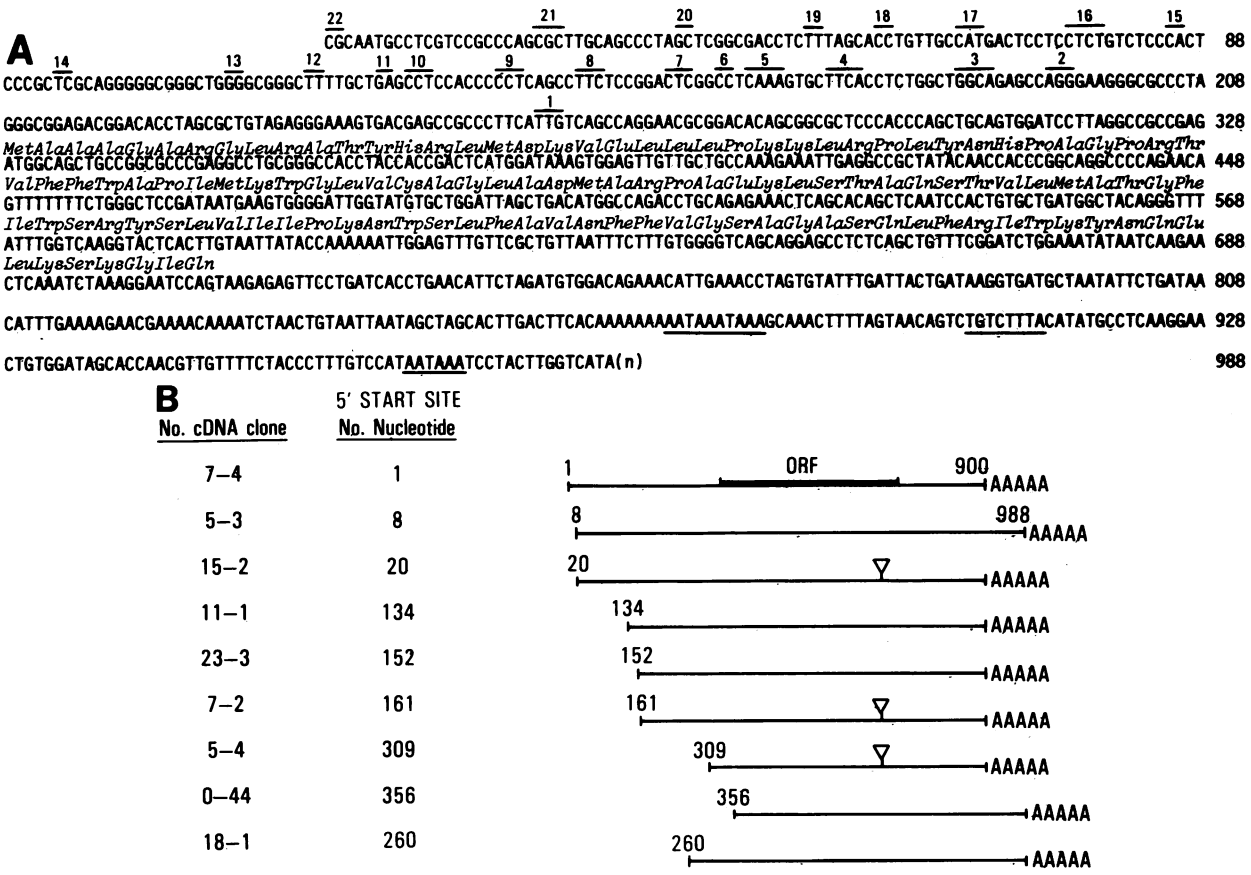

FIG. 2. Composite of the nucleotide sequences of 0-44-homologous cDNA inserts. (A): Nucleotide residue 1 represents the 5'-most nucleotide of the longest cDNA clone (p7-4). The multiple 5' ends from the S1 nuclease mapping study (Fig. 4) are indicated by numbered bars above the sequence. The amino acid sequence translated from the ORF (nucleotides 329 to 709) is presented above the nucleotide sequence. Two putative polyadenylation signals and a possible auxiliary sequence are underlined. The sequence is presented in the DNA form, but the Ts are Us in the RNAs. (B) The structures of 0-44-related cDNA clones are illustrated. The 5' nucleotides were determined from the DNA sequence. cDNA clones containing 17 extra nucleotides are shown with an inverted triangle near the 3' end of the ORF.

clones. Lastly, two different lengths of 3'-untranslated regions were seen.

Multiple, nested 5' ends. It was apparent from Northern blot hybridization (Fig. 1) that there was heterogeneity in the p0-44-related mRNAs, and the nucleotide sequence analysis (Fig. 2) suggested that the 5' ends of these mRNAs could be heterogeneous. We mapped the 5' ends of these mRNAs using two different techniques: primer extension (14) and S1 nuclease resistance analysis (8).

For primer extensions, we prepared two double-stranded DNA fragments from the 5' region of clone p0-44: the 330-base-pair (bp) *Hin*fI fragment corresponding to nucleotides 373 to 702 and the 59-bp *Hin*fI-*Hpa*II fragment corresponding to nucleotides 373 to 431. We 5' end-labeled these fragments with [γ-^{32}P]ATP and polynucleotide kinase; heat denatured the end-labeled fragments; and annealed them to poly(A)$^+$ RNA from rat brain, liver, or kidney. The paired antisense fragment then served as a primer which was elongated by reverse transcriptase in the presence of unlabeled deoxynucleotide triphosphates. The extended primer cDNA molecules were isolated, separated on gels, and detected autoradiographically. For most mRNAs, such an experiment would give one extended single-stranded cDNA fragment, the length of which, as deduced by gel mobility, would correspond to the distance of the 5' end of the primer to the 5' end of the mRNA.

When either of the two fragments from p0-44 was used for primer extensions, an equivalent, complex result was obtained (Fig. 3). The data revealed two major features. First, in all three tissues more than one extended cDNA was observed, indicating that more than one species of RNA is capable of serving as a template for the primer. The relative concentrations of the different template RNAs should correspond directly to the relative intensities of the extended primers. Some of the shorter cDNAs could possibly have corresponded to incompletely extended primers or "stutter" products, but results of the S1 nuclease experiments described below rule this out for most species. Second, both common and apparently brain-specific extended primers were observed, and the relative concentrations of the common species varied between the brain and the other test tissues. For example, fragments 1, 2, and 3 (Fig. 3A) were seen to be common to all three test tissues; their band intensities, however, varied greatly between brain and nonbrain tissues. Most of the longer products appeared to be unique to the brain. This is the first direct evidence that demonstrates that there is heterogeneity in the 5'-untranslated regions of p0-44 homologous transcripts. The result with the 330-bp primer (Fig. 3A) shows closely spaced gel bands, the distribution of which matches the Northern blot profile shown in Fig. 1.

Both the 330- and 59-bp fragments gave analogous results;

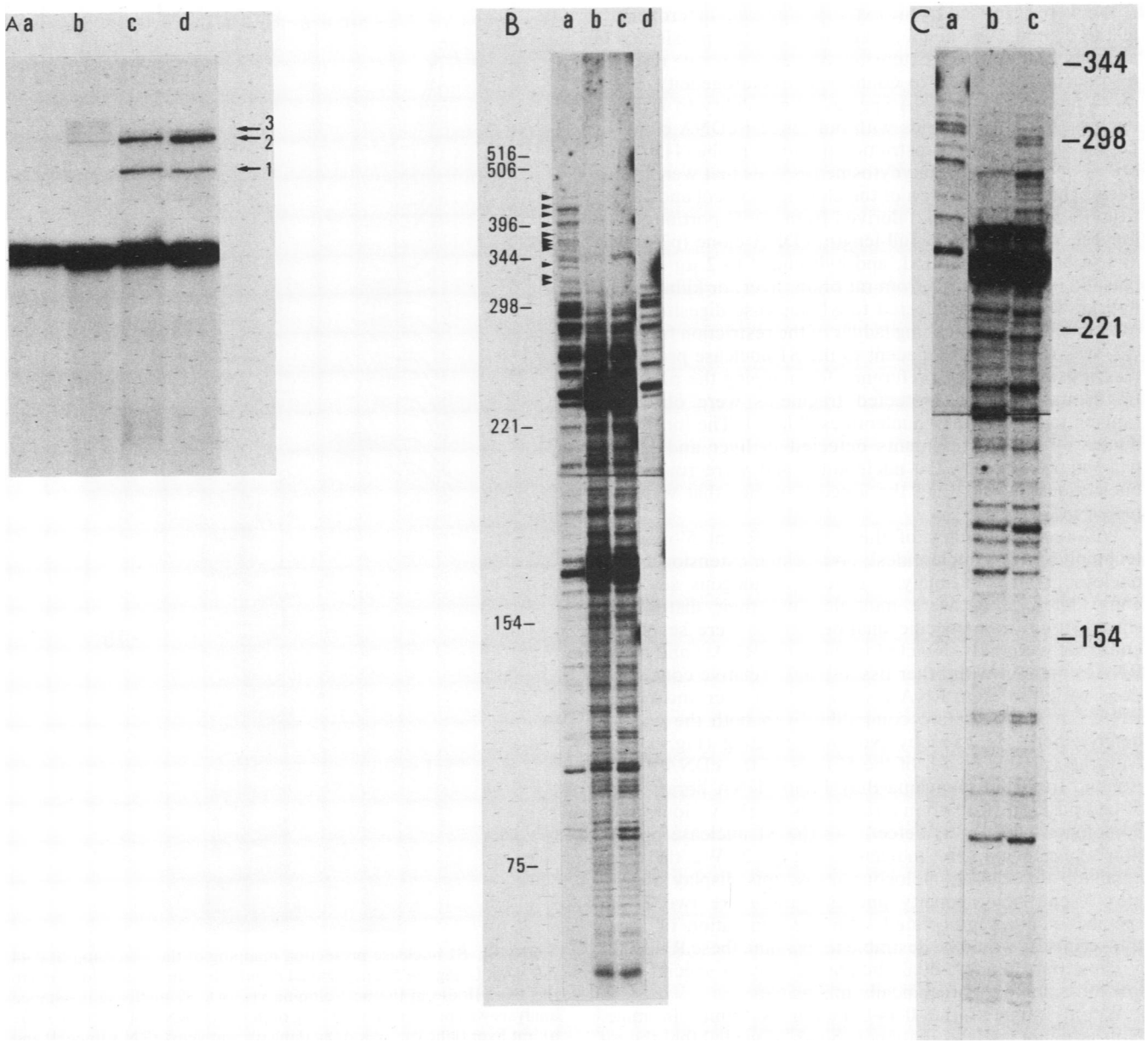

FIG. 3. Mapping the 5′ ends of the 0-44 mRNAs by primer extension analysis. The different poly(A)$^+$ RNA samples (8 μg) listed below were annealed to 50 kcpm of ^{32}P-labeled primers. Primer 1 (330-bp *Hin*fI fragment, positions 373 to 702) was used in panel A. Primer 2 (59-bp *Hin*fI-*Hpa*I fragment, positions 373 to 431) was used in panels B and C. The primer extension reaction was done as described in the text. The samples used in panel A were as follows: yeast tRNA (lane a), rat brain RNA (lane b), rat liver RNA (lane c), and rat kidney RNA (lane d). The samples used in panel B were as follows: rat brain RNA (lane a), rat liver RNA (lane b), rat kidney RNA (lane c), and mouse brain RNA (lane d). The samples used in panel C were as follows: RNA from mouse brain (lane a), liver (lane b), and kidney (lane c). The extended products were analyzed on a 5% polyacrylamide–7 M urea gel (panel A) or 6% polyacrylamide-7 M urea gel (panels B and C). In panel C lane a was separated from lanes b and c on the original gel, hence the slight misalignment of bands. The numbered arrows (panel A) indicate the extended products common to all three tissues, while the arrowheads (panel B) mark the brain-specific extended products. Numbers to the left in panel B and to the right in panel C are DNA size markers (pBR322 digested with *Hin*fI).

thus, primer length is not a factor. However, use of the 59-bp fragment allowed better resolution of the extended primer molecules in the resolving range of our gels (Fig. 3B). About 11 distinct bands were identified in liver and kidney, and at least 22 bands could be recognized in the brain, including the 11 bands common with liver and kidney. The 5′ ends of the two most predominant bands common to all three tissues correspond to extended primer lengths of 250 to 260 and 170 to 174 nucleotides and could be mapped to within a few nucleotides on the cDNA sequence. The 5′ ends detected only in the brain tissue RNAs extended further in the 5′ direction than the 5′ ends common with the liver and kidney mRNAs. The longest extended primer found when brain RNA was template (440 nucleotides) was 80 to 90 nucleotides longer than the longest extended molecules (350 nucleotides) that were found when either liver or kidney RNA was template. The longest brain-extended primer mapped to the 5′ end of our longest cDNA clone, p7-4.

We were surprised by the extent of apparent heterogeneity at the 5′ end and, hence, were wary of the possibility of artifacts in the primer extension studies. Thus, we resorted to a second approach to confirm this heterogeneity in the p0-44 homologous mRNA species: their ability to form S1 nuclease-resistant hybrids with our longest cDNA clone. A double-stranded DNA restriction fragment probe (445 nucleotides, including 32 deoxycytosine residues that were added during the tailing reaction of the cloning protocol and 103 bp from the cloning vehicle) derived from the 5′ portion of the longest, and apparently full-length, cDNA clone (p7-4) was kinase labeled, denatured, and hybridized to 2 μg of cytoplasmic poly(A)$^+$ RNA from rat brain, liver, or kidney. The hybrids were then subjected to S1 nuclease digestion. The Maxam-Gilbert sequencing ladder of the restriction fragment (17) was run in lanes adjacent to the S1 nuclease protection reactions to serve as a reference for mapping the protected fragments. Multiple protected fragments were observed, ranging from 45 to 310 nucleotides (Fig. 4). The longest S1 nuclease-resistant fragments detected in liver and kidney were approximately 230 nucleotides and were roughly 80 nucleotides shorter than the longest species that were detected in the brain.

Within the accuracy of these two different 5′-mapping techniques (a few nucleotides), both primer extension and S1 nuclease resistance analysis gave an analogous set of 5′ ends. These ends were multiple, and their distribution showed two brain-specific qualities: there were several 5′ ends that were only detected in the brain RNAs; and of the RNAs common with other tissues, their relative concentrations varied in brain RNA compared with liver and kidney RNA. These results are compatible with both the original heterogeneity detected by Northern blotting (Fig. 1) and with the variable length of the 5′ ends of the cDNA clones isolated from our Okayama-Berg brain cDNA library. The 5′-end positions (to within a few bases) of the various p0-44 homologous RNAs, as detected in the S1 nuclease protection experiments, are indicated in Fig. 2. We therefore conclude that 0-44-homologous transcripts display a complex 5′-end heterogeneity and that there are two brain-specific qualities associated with the generation of those transcripts. It would be desirable to examine these RNAs for 5′ caps, but the individual abundance of each species is so low as to make this technically unfeasible.

We investigated the 0-44-homologous family in mice. Northern blot analysis (data not shown) revealed that the rat p0-44 cDNA clone hybridized to heterogeneous targets in RNA samples from mouse liver and kidney and much less intensely to a heterogeneous smear in mouse brain RNA. Mobility differences between the targets in the RNAs from the different tissues were not obvious. Primer extensions with the rat probe revealed that the mouse brain RNA family contained members with many different 5′ ends, although not as many as we detected in rat brain RNA (Fig. 3B, lane d). The mouse brain RNA 5′ ends were all present in mouse liver and kidney RNAs (Fig. 3c), although again both the absolute and relative abundance of each 5′ end varied among RNAs from brain and the other two tissues. Brain contains relatively more of the larger RNAs. Thus, the mouse 0-44 RNAs show one of the two brain-specific qualities observed in the rat RNAs.

0-44 homologous transcripts encode putative proteins with different carboxyl termini. When 33 of the cDNA clones that hybridized to p0-44 were subjected to restriction analysis with the *Hin*fI enzyme, 3 clones (p5-4, p7-2, p15-2) were found to possess an internal fragment that migrated slightly more slowly than the corresponding fragment (nucleotides 373 to 702) of the other 30 clones, as though the altered fragment were about 15 bp longer. Sequence analysis of this altered *Hin*fI fragment indicated that there was a 17-bp insertion (relative to the composite sequence shown in Fig. 2) near its 3′ end (Fig. 5). The same insertion was found in all

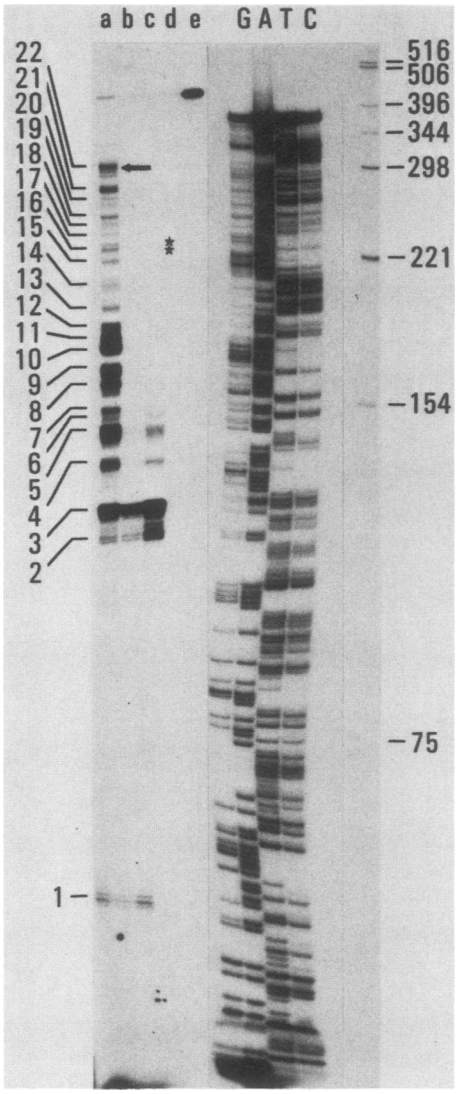

FIG. 4. S1 nuclease protection mapping of the 5′ termini of 0-44 mRNAs. Using the 5′-end-labeled *Bam*HI fragment (positions 310 to the *Bam*HI site in the pcD cloning vector), S1 nuclease resistance analysis was performed with 2 μg of poly(A)$^+$ RNA of rat brain (lane a), rat liver (lane b), rat kidney (lane c), and yeast tRNA (lane d) and was analyzed on a 6% polyacrylamide–7 M urea buffer gradient gel (9). The labeled probe is shown in lane e. To align the S1 nuclease-resistant fragments with the cDNA sequence, a parallel set of Maxam-Gilbert DNA sequencing reactions (17) was performed on the 379-bp restriction fragment generated by digesting the S1 nuclease probe with *Rsa*I endonuclease. The sequence obtained by this method is complementary to the sense strand. The products of the sequencing reaction and the S1 nuclease-resistant fragments on the sequencing ladder were assumed to be migrating 1.5 nucleotides faster than the S1 nuclease products because of the different 3′ structures produced by these techniques (18, 37). The arrow indicates the longest visible fragment protected by rat brain RNA, and the stars indicate the longest visible fragment protected by rat liver and kidney RNAs. DNA size markers are shown at the right.

```
                PheArgIleTrpLy                    sTyrAsnGlnGluLeuLysSerLysGlyIleGln
p0-44           TTTCGGATCTGGAA----------ATATAATCAAGAACTCAAATCTAAAGGAATCCAGTAA

                PheArgIleTrpLys
p15-2           TTTCGGATCTGGAAGTAAGCCTACTACCAAGATATAATCAAGAACTCAAATCTAAAGGAATCCAGTAA
```

FIG. 5. Nucleotide sequence of the 17-bp insertion. Comparison of the nucleotide sequences and the deduced amino acid sequences of an insertion-lacking clone (p0-44) and an insertion-containing clone (p15-2) is presented.

three clones. The insertion introduces an in-frame translational termination codon (TAA) at the 3′ end of the ORF (Fig. 5), resulting in reduction of the size of the putative protein from 127 to 116 amino acids. The amino acid sequences of the shorter and longer species are the same, except for the 11 amino acids unique to the C terminus of the longer protein.

Is this alteration in the putative protein-coding sequence unique to the mRNA expressed in the brain? To determine this we looked for the modified transcript in different tissues by S1 nuclease resistance analysis. The insertion-containing *Hin*fI fragment (nucleotides 373 to 719) was radiolabeled at its 5′ ends, heat denatured, and hybridized at 40°C with the poly(A)$^+$ RNA of rat brain or liver or yeast RNA. The hybrids were subsequently subjected to S1 nuclease digestion, and the product was analyzed on polyacrylamide gels (Fig. 6). Several fragments were protected by both brain and liver RNA, and the patterns were identical. Yeast RNA did not protect any fragments.

Two groups of fragments were observed. One large fragment comigrated with the denatured probe (347 nucleotides). Under the hybridization condition we used, the majority of the probe was maintained in a single-stranded state in the S1 nuclease assay in the absence of protection from specific RNAs, as shown in the control reaction with the yeast tRNA (Fig. 6, lane d). The fragment which migrated parallel with the probe therefore represents the probe that is fully protected by the specific RNA containing the 17-bp insertion (Fig. 6, lanes b and c, arrow). The second group of fragments migrated between 24 and 34 nucleotides. By aligning the protected fragments with the nucleotide sequence of the probe, we could identify each S1 nuclease product. The 34-nucleotide fragment represented the probe protected by RNAs lacking the 17-nucleotide insertion. The remaining smaller fragments probably arose from "nibbling" of the 34-mer by S1 nuclease at the AT-rich sequences near the 3′ end of the protected piece, as can be seen in the sequence in Fig. 5. Essentially equivalent results were obtained with a second probe, the 5′ end of which was 176 nucleotides from the insertion site (data not shown). In the experiment shown in Fig. 6 (lanes b and c), the cumulative intensities measured by densitometry (Table 1) of the partially protected 34- to 24-nucleotide fragments were approximately 10 times greater than that of the fully protected 347 nucleotide probe. The same ratio was observed in the experiment with the second probe. This ratio (1:10) reflects the frequency that clones with the 17-bp insertion were isolated (3 of 33). We therefore conclude that there is no tissue selectivity in the relative expression of the RNAs which differ by the 17-nucleotide insertion.

Two sites for polyadenylation. In the 3′-untranslated region (279 bp) of the composite sequence three AATAAA sequences appear. Such sequences precede poly(A) addition sites in many eucaryotic mRNAs (30). The first two signals are overlapped as AATAAATAAA. The third signal is 88 nucleotides downstream from the second (Fig. 2).

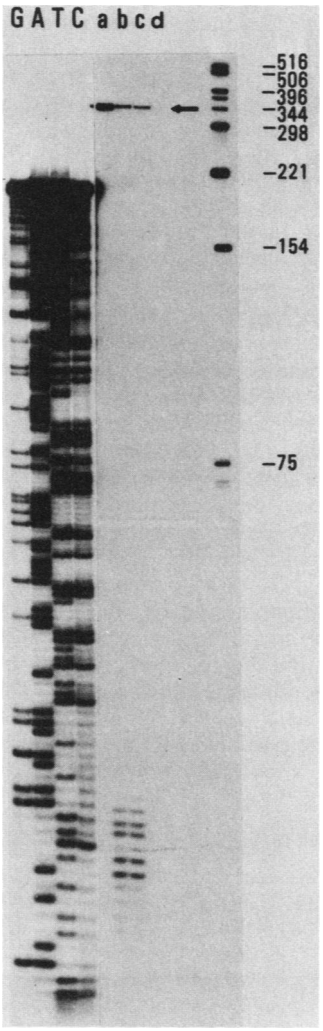

FIG. 6. Detection of the insertion-containing RNA species by S1 nuclease resistance analysis. The 347-bp *Hin*fI fragment (positions 373 to 719 from clone p15-2) was 5′ labeled and annealed to 2 μg of poly(A)$^+$ RNA of rat brain (lane b), rat liver (lane c), and yeast tRNA (lane d) at 40°C. The RNA-DNA hybrids were digested subsequently with S1 nuclease, and the resistant products were displayed on a 8% polyacrylamide–7 M urea gel. The S1 nuclease probe is presented in lane a. To align the protected products with the cDNA sequence, a parallel Maxam-Gilbert DNA sequencing reaction (17) was performed on the 219-bp *Hin*fI-*Pst*I restriction fragment generated by *Pst*I digestion of the S1 nuclease probe and the reaction products displayed on the same gel (two different exposure times were used and the photograph was reassembled). The sequence obtained represents the sequence complementary to the sense strand. DNA size markers are shown on the right.

Analysis with restriction enzymes and sequencing analyses revealed an *Nde*I restriction site (CATATG) 28 nucleotides downstream of the second poly(A) signal, providing a handle to investigate possible 3'-end differences. Thirty-three randomly selected clones from the p0-44 hybridization screen of the Okayama-Berg library were subjected to restriction analysis with the *Nde*I enzyme. Among them, 15 clones (46%) were found to contain this diagnostic *Nde*I site, while the remaining 18 clones (54%) did not have this site. This indicates that about 50% of the 0-44 transcripts chose the third poly(A) signal to dictate the site of poly(A) addition during the maturation of messages. The other half of the mRNAs selected the proximal sites for polyadenylation. This was further confirmed by sequencing eight of the independent cDNA clones. It was found that the poly(A) stretch starts either 14 nucleotides downstream from the third signal sequence (two clones) or 16 nucleotides after the second signal (six clones). Thus, the 3' ends may differ by 88 nucleotides. Three of the clones with the shorter 3'-untranslated region were the clones with the 17-nucleotide insertion at the 3' end of the ORF. We have not investigated any tissue differences in this alternative polyadenylation phenomenon for the 0-44 RNAs.

One gene gives rise to many mRNAs. One possible explanation for this complicated RNA family is that these related RNAs are transcribed from a multimember gene family. This possibility, however, is unlikely based on the evidence obtained from genomic Southern blot analysis. Digests of rat liver genomic DNA with each of five restriction endonucleases were size fractionated on an agarose gel, transferred to nitrocellulose paper, and subsequently hybridized with nick-translated p0-44 cDNA probes. For each of the five enzymes, one to three genomic fragments were detected (Fig. 7). Using a restriction map and partial sequence data from a genomic clone, we could account for each fragment. The 0-44 gene could be as small as 5 kilobases, as demonstrated by the single *Eco*RI fragment (4.5 kb), and the genomic data show that there is no tandem arrangement within this fragment. Therefore, it is likely that 0-44 is a single-copy gene rather than part of a gene family.

Sequence of the 5' region of the 0-44 gene. A clone was isolated from the partial *Eco*RI rat genomic library of Sargent et al. (34) using cDNA clone p0-44 as the hybridization probe, and a restriction map of the genomic clone was generated. An *Xba*I fragment that hybridized to fragments from the 5' end of the 0-44 clones was isolated, and its 1,328-nucleotide sequence was determined (Fig. 8). The sequence contains and is colinear with the 5' 437 nucleotides of the composite cDNA sequence (Fig. 8). An excellent match with the splice donor consensus sequence CAG↓GTAAGT (10) is observed at the 3' site of divergence. Thus, the region after that is an intron. Hybridization and restriction mapping studies indicate that the intron is large; hence, we did not attempt to determine the entire gene sequence.

Eucaryotic promoters typically are found to have TATA and CAAT sequences at positions 25 to 30 and 70 to 80 nucleotides, respectively, upstream from the sites of transcriptional initiation. Neither of these sequences or closely related sequences is observed at appropriate locations up-

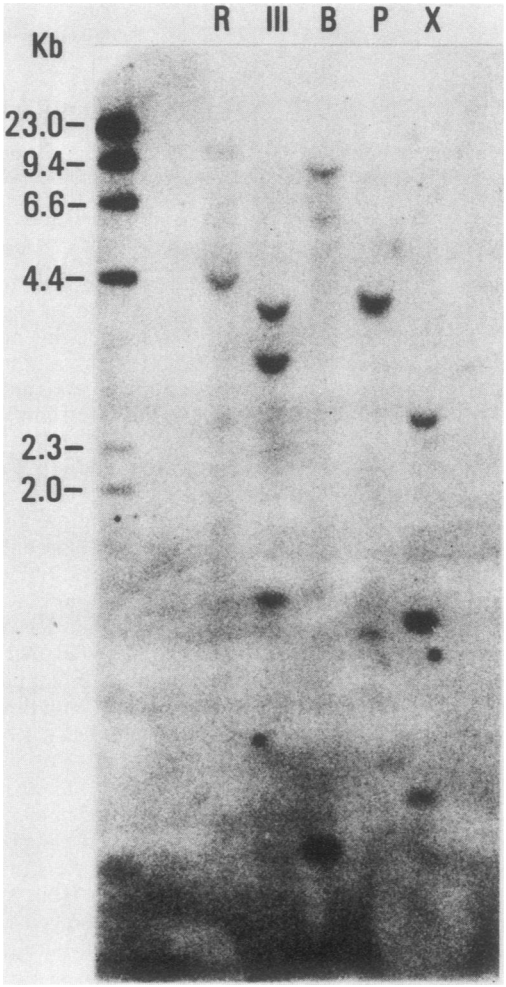

FIG. 7. Southern blot analysis of the 0-44 gene. Fractions (10 μg) of rat liver genomic DNA were digested with restriction enzymes *Eco*RI (R), *Hin*dIII (III), *Bam*HI (B), *Pst*I (P), and *Xba*I (X); size fractionated on 0.8% agarose gel; and transferred to nitrocellulose. The blot was then hybridized with nick-translated clone p7-4. The blot was washed in 0.5× SSC (1× SSC is 0.15 M NaCl plus 0.015 M sodium citrate) 0.5%–sodium dodecyl sulfate at 65°C for 2 h. *Hin*dIII-digested λ DNA was used as size markers, and sizes are given on the left in kilobases (Kb).

TABLE 1. Relative levels of S1 nuclease resistance determined by densitometry analysis[a]

Length of S1 nuclease-resistant fragments (nucleotides)	Relative (%) levels of S1 nuclease resistance in:	
	Brain	Liver
347	9.09	8.28
34	3.39	2.9
33	8.5	7.9
32	12.5	11.04
31	13.8	13.37
30	6.5	6.14
29	16.4	14.79
28	16.1	16.27
27	5.2	6.86
26	2.8	3.62
25	2.9	4.45
24	2.8	4.37

[a] The autoradiogram of S1-nuclease resistance analysis shown in Fig. 6 was subjected to densitometer scanning. The peaks were cut from the paper tracing and weighed. The relative resistance levels were determined by dividing the weight of the peak of each protected fragment by the total weight of the protected fragments.

FIG. 8. Genomic sequence of 5′ region of 0-44 gene. One strand of the 1,328-nucleotide sequence of the *Xba*I fragment containing the 437-nucleotide 5′ 0-44 exon (underlined) is shown. The 3′ extent of the exon was ambiguous because GG was in both genomic and cDNA sequences; thus, the best alignment with the 5′ splice donor consensus was used to identify the end of the exon. The position of the putative initiator methionine codon is shown (Met). No obvious homologies to consensus promoter elements were observed in either the 5′-upstream or 5′-untranslated regions.

stream from the 0-44 5′ exon. These elements are also absent from the heterogeneous putative 5′-untranslated region and the 600 nucleotides upstream from the 5′ exon.

Early developmental onset for the brain-specific 5′ ends. Because adult brain differs from other adult tissues in its 5′ end pattern for the 0-44 RNAs, we asked when this difference appears during development. S1 nuclease resistance analysis was performed with the cytoplasmic poly(A)$^+$ obtained from brains of both prenatal and postnatal rats. The same probe that was used previously to analyze 5′ heterogeneity (Fig. 4) was used in this assay. The pattern of the S1 nuclease probe protection (Fig. 9) by RNAs from rats of different ages appeared to be roughly identical to the pattern in adult rats, even at the earliest developmental stage (14 days) at which we could confidently dissect out reasonable amounts of tissue for analysis. Thus, the developmental switch which activates the expression of the brain-specific 5′ ends occurs prior to embryonic day 14.

We investigated this phenomenon further by examining RNAs derived from established rat cell lines: C6 glioma and NRK cells. The poly(A)$^+$ RNA from C6 glioma cells protected the probe in the same manner as the rat brain RNA (data not shown); thus, the brain-specific 5′ ends found in RNA samples from animals were also found in a cell line, and their existence is probably not an artifact of isolation. The NRK pattern looked like that of liver or kidney RNA. Thus, these cell lines contain 0-44 RNAs similar to those of their in vivo counterparts.

DISCUSSION

Structure of 0-44 family. We found that a member of the rat 0-44 RNA family may differ from other members of the family at one or more of three positions. Any of 22 or possibly more sites, may serve as the first nucleotide of the RNA; thus, the 5′ ends are nested and contain identical sequences where they overlap, but they vary in length. The central portion of the RNA may be of either of two forms that differ by the presence or absence of 17 nucleotides. The 3′ poly(A) tail may be added at either of two sites. A drawing representing these structures is shown in Fig. 10. This complicated family of RNAs encodes two putative proteins that differ only in the length of their C-terminal sequences. The evidence from Southern blotting and genomic cloning suggests that there is only one 0-44 gene. Thus, the three structural differences occur as a result of transcription or posttranscriptional processing of the same primary transcript. It is possible, assuming that the structure at each of the three positions is determined independently, that a single 0-44 gene gives rise to as many as 88 distinct RNA structures, which is the number of possible permutations ($22 \times 2 \times 2$) among the three positions. These multiple RNAs can account for the diffuse pattern of hybridization targets for p0-44 in Northern blots (Fig. 1).

There is no apparent tissue variation in the choice of structure at the middle position, because 10% of the RNAs in both brain and liver contain the extra 17 nucleotides. We did not examine RNA from tissues other than the brain at the 3′ position. However, at the 5′ position, two types of tissue variation were observed. First, some 5′ ends appeared in all tissues, whereas several were detectable only in the brain. Second, among those 5′ ends common to the three test tissues, the frequency that any particular 5′ site appeared, relative to that of any other sites or relative to that of the same site in another tissue, differed between brain and the other two tissues (which appeared to have frequencies that were the same). Thus some members of the 0-44 family are Class II mRNAs and others are class III mRNAs. We cannot rule out the possibility that the class III mRNAs are present in other tissues at a level below our detection sensitivity limit. The tissue-specific differences we documented can

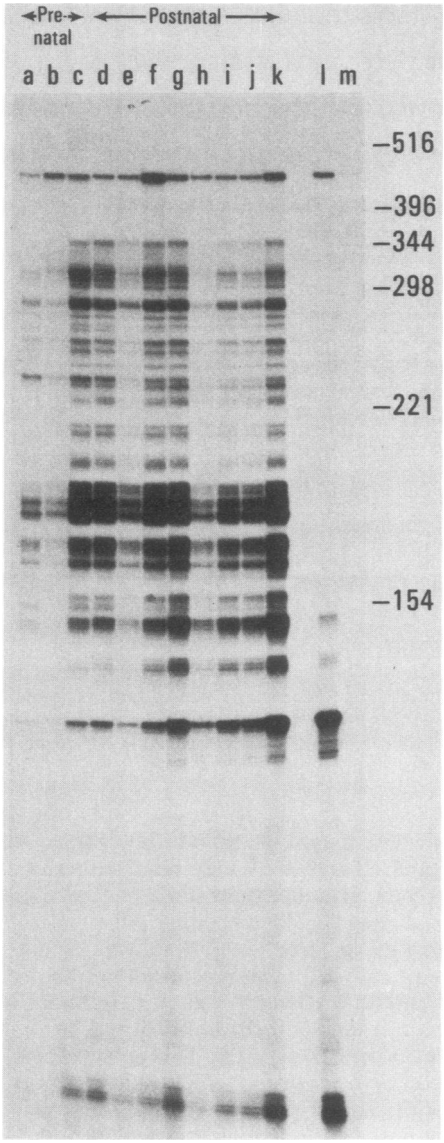

FIG. 9. Brain-specific 5′ ends detected early in development. S1 nuclease resistance analysis was performed with same S1 nuclease probe described in Fig. 4 and the poly(A)$^+$ RNA samples of rat brains from a series of developmental stages. RNA samples used in this assay were taken from prenatal rats that were 14, 16, and 20 days old (lanes a through c, respectively), postnatal rats that were 1, 5, 10, 15, 20, 25, and 30 days old (lanes d through j, respectively), and adult rat brain and liver (lanes k and l, respectively). A control reaction with yeast tRNA is included in lane m. DNA size markers are shown on the right.

account for the differences observed between brain RNA and liver or kidney RNA in the Northern blot experiment (Fig. 1).

Generation of 0-44 RNAs. How might these variations in RNA structure arise? At the middle position, the gene, and hence the primary transcript, must contain the 17-nucleotide sequence. Thus, these nucleotides must be removed from 90% of the transcripts at the level of RNA splicing. Ample precedents for alternative splicing exist (3, 6, 13, 21, 25, 26, 41). In some of those cases, the alternative splicing choices are made in a tissue-specific way; that is not the case for the 0-44 family. Nonetheless, for 0-44 the splicing choice appears to be important because its outcome determines the ratio between RNAs which do and do not carry the 17 nucleotides, and these two resulting RNA species encode putative proteins with different carboxy termini. Presumably, the 127-amino-acid protein is synthesized at 10 times the rate of the 116-amino-acid protein, assuming that the two types of RNA species are equally good templates for protein synthesis.

Small alterations in the carboxy-terminal sequences of proteins due to alternative RNA splicing events can affect their cellular localization as well as functions. A well-documented example of such is the different forms of immunoglobulin M (IgM), i.e., monomeric membrane-bound IgM receptors versus pentameric secreted IgM effectors (2). The amino acid sequences of the 0-44 putative proteins do not provide sufficient information to speculate the possible fates of each protein molecule. Computer search against protein sequence data bases has not revealed any related protein sequences. The identities of these two proteins and their functional analysis will be pursued with specific antipeptide sera.

The use of alternative 3′ polyadenylation sites also has ample precedence (23, 28, 36, 40). For 0-44, it must be that the initial poly(A) signal functions inefficiently either because of some suboptimal specific secondary signal or because the general structure of the nascent heterogeneous nuclear RNA transcript is inadequate for efficient cleavage. It has been suggested (20) that the consensus sequence YGTGTTYY 30 nucleotides downstream from the polyadenylation sequence AATAAA is required for efficient 3′-end processing. We find the sequence TGTCTTTA (Fig. 2) in the appropriate position following the AATAAATAAA sequence in 0-44.

Possible 5′ promoters. Precedents for 5′ mRNA heterogeneity exist: simian virus 40 early transcripts (31), human ε globin (1), c-*myc* (29), adenovirus E2 (15), mouse hypoxanthine phosphoribosyltransferase (22), 3-hydroxyl-3-methylglutaryl coenzyme A reductase (32), mouse dihydrofolate reductase (19). Thus far the only examples of tissue-specific 5′ polymorphism (to our knowledge) other than 0-44 are those for mouse α-amylase (41) and *Drosophila melanogaster* alcohol dehydrogenase (7). In these cases either of two alternate 5′ ends is spliced into a common protein-coding RNA body. The tissue-specific portions of the 0-44 family of RNAs also lie in the region upstream from

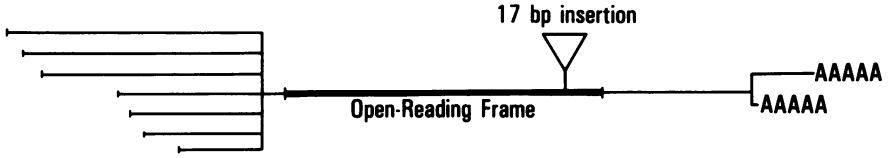

FIG. 10. Schematic presentation of the structures of 0-44 mRNA family.

the putative translational initiation site. However, as the genomic and cDNA sequences are colinear in the 5′ regions, the use of different exons does not account for the 0-44 heterogeneity at the 5′ end. Assuming that all of the 0-44 RNAs are 5′ capped, transcription begins at 22 or more distinct, closely spaced sites spanning 265 nucleotides in the 5′ region of the gene; some of these initiation sites are utilized detectably only in the brain lineage, and others are used differentially in that lineage. Thus, unlike α-amylase or alcohol dehydrogenase mRNAs, the two alternate 5′ ends of which have unrelated sequences, the 0-44 5′ sequences are nested.

The sequence of the 5′-untranslated region of clone 7-4 does not contain multiple CAAT or TATA-related sequences; in fact, this region is quite rich in GC (69%). In addition, such known control sequences are not located at appropriate positions upstream from the most 5′ initiation site. Thus, models in which classical RNA polymerase II initiation signals are used (10) are unlikely. Several of the genes listed above have GC-rich regions in their heterogeneous 5′ ends. Regardless of the mechanism for generating these related RNAs, the brain shows exceptional qualitative differences with other tissues in their production, and these differences could reflect fundamental alterations in how transcription is initiated.

The 5′ untranslated region of the 0-44 gene ranges from about 60 to 329 nucleotides in length. The region found in the largest brain transcripts is approximately 80 nucleotides longer than those found in the liver and kidney, as determined by both primer extension and S1 nuclease resistance assays. The length of the 0-44 5′-noncoding region is much longer than that of 95% of the eucaryotic cellular mRNAs: 160 nucleotides from a compilation study by Kozak (16) with 211 cellular RNAs. There are two AUG triplets in the upstream sequence (positions 5 and 65), but it is unlikely that these are utilized for translation because they are each followed by nonsense triplets. We note that the ORF beginning at position 5 extends for 40 triplets and is only found in the largest brain-specific RNA; thus, it is remotely possible, but very unlikely, that there is a brain-specific small protein that is encoded. No related proteins of known sequence were found by computer search.

Work in many laboratories over the past few years has shown that eucaryotic genes often give rise to more than one mRNA. The single 0-44 gene gives rise to possibly 88 or more mRNAs which cumulatively encode two putative proteins. Thus, the 0-44 family is perhaps the most extreme example to date of diverse expression from a single gene, including tissue-specific transcription patterns. It is difficult to understand the significance of the 0-44 5′ polymorphism. The mouse 0-44 family is simpler than that of the rat but resembles it in that the brain contains relatively more of the larger RNA species. It is unknown what the functional consequences of this 5′-end variability might be. However, this finding demonstrates a way of generating molecular polymorphism in the central nervous system, and it is most likely that for some transcription units (possibly including 0-44), this mechanism is critical.

ACKNOWLEDGMENTS

We thank Rob Milner, Sonja Forss-Petter, Gabriel Travis, Ian Brown, Mary Ann Brow, Mary Kiel, Dominique Lenoir, David Levy, Michael Wilson, and Randy McKinnon for frequent discussions and helpful suggestions and Linda Elder for manuscript assistance.

This work was supported in part by Public Health Service grant GM32355 from the National Institutes of Health. A.-P. T. thanks the Leukemia Society of America for fellowship support. This is publication no. 3944-MB from the Research Institute of Scripps Clinic.

LITERATURE CITED

1. **Allan, M., W. G. Lanyon, and J. Paul.** 1983. Multiple origins of transcription in the 4.5 kb upstream of the ϵ-globin gene. Cell **35**:187–197.
2. **Alt, F. W., A. L. M. Bothwell, M. Knapp, E. Siden, E. Mather, M. Koshland, and D. Baltimore.** 1980. Synthesis of secreted and membrane-bound immunoglobulin mu heavy chains is directed by mRNAs that differ at their 3′ ends. Cell **20**:293–301.
3. **Amara, S. G., V. Jonas, M. G. Rosenfeld, E. S. Ong, and R. M. Evans.** 1982. Alternative RNA processing in calcitonin gene expression generates mRNAs encoding different polypeptide products. Nature (London) **298**:240–244.
4. **Aviv, H., and P. Leder.** 1972. Purification of biologically active globin messenger RNA by chromography on oligothymidylic acid-cellulose. Proc. Natl. Acad. Sci. USA **69**:1408–1412.
5. **Bantle, J. A., and W. E. Hahn.** 1976. Complexity and characterization of polyadenylated RNA in the mouse brain. Cell **8**:139–150.
6. **Basi, G. S., M. Boardman, and R. V. Storti.** 1984. Alternative splicing of a *Drosophila* tropomyosin gene generates muscle tropomysin isoforms with different carboxy-terminal ends. Mol. Cell. Biol. **4**:2828–2836.
7. **Benyajati, C., N. Spoerel, H. Haymerle, and M. Ashburner.** 1983. The messenger RNA for alcohol dehydrogenase in *Drosophila melanogaster* differs in its 5′ end in different developmental stages. Cell **33**:125–133.
8. **Berk, A. J., and P. A. Sharp.** 1977. Sizing and mapping of early adenovirus mRNAs by gel electrophoresis of S1 endonuclease digested hybrids. Cell **12**:721–732.
9. **Biggin, M. D., T. J. Gibson, and G. F. Hong.** 1983. Buffer gradient gels and ^{35}S label as an aid to rapid DNA sequence determination. Proc. Natl. Acad. Sci. USA **80**:3963–3965.
10. **Breathnach, R., and P. Chambon.** 1981. Organization and expression of eukaryotic split genes coding for proteins. Annu. Rev. Biochem. **50**:349–383.
11. **Brow, M. A. D., R. Pesin, and J. G. Sutcliffe.** 1985. The tetracycline repressor of pSC101. Mol. Biol. Evol. **2**:1–12.
12. **Casey, J., and N. Davidson.** 1977. Rates of formation and thermal stabilities of RNA:DNA and DNA:DNA duplexes at high concentrations of formamide. Nucleic Acids Res. **4**:1539–1552.
13. **Early, P., J. Rogers, M. Davis, K. Calame, M. Bond, R. Wall, and L. Hood.** 1980. Two mRNAs can be produced from a single immunoglobulin μ gene by alternative RNA processing pathways. Cell **20**:313–319.
14. **Giorgi, C., B. M. Blumberg, and D. Kolakofsky.** 1983. Sendi virus contains overlapping genes expressed from a single mRNA. Cell **35**:829–836.
15. **Hashimoto, S., and M. Green.** 1984. Unusual heterogeneity of the 5′-termini of human adenovirus type 2 early region E2 mRNA. Nucleic Acids Res. **12**:1067–1082.
16. **Kozak, M.** 1984. Compilation and analysis of sequences upstream from the translational start sites in eukaryotic mRNAs. Nucleic Acids Res. **12**:857–872.
17. **Maxam, A. M., and W. Gilbert.** 1976. A new method of sequencing DNA. Proc. Natl. Acad. Sci. USA **74**:560–564.
18. **McConnell, D. J., D. G. Searcy, and J. G. Sutcliffe.** 1978. A restriction enzyme ThaI from the thermophilic mycoplasma Thermoplasma acidophilum. Nucleic Acids Res. **5**:1729–1739.
19. **McGrogan, M., C. C. Simonsen, D. T. Smouse, P. J. Farnham, and R. T. Schimke.** 1985. Heterogeneity at the 5′ termini of mouse dihydrofolate reductase mRNAs. J. Biol. Chem. **260**:2307–2314.
20. **McLauchlan, J., D. Gaffney, J. L. Whitton, and J. B. Clements.** 1985. The consensus sequence YGTGTTYY located downstream from the AATAAA signal is required for efficient formation of hnRNA 3′ termini. Nucleic Acids Res. **13**:1347–1368.

21. **Medford, R. M., H. T. Nguyen, A. T. Destree, E. Summers, and B. Nadal-Ginard.** 1984. A novel mechanism of alternative RNA splicing for the developmentally regulated generation of troponin T isoforms from a single gene. Cell **38**:409–421.
22. **Melton, D. P., D. S. Konecki, J. Brennand, and C. T. Caskey.** 1984. Structure, expression and mutation of the hypoxanthine phosphoribosyltransferase gene. Proc. Natl. Acad. Sci. USA **81**:2147–2151.
23. **Milner, R. J., C. Lai, K. A. Nave, J. Ogata, D. Lenoir, and J. G. Sutcliffe.** 1985. Nucleotide sequences of the two mRNAs for rat brain myelin proteolipid protein. Cell **42**:931–939.
24. **Milner, R. J., and J. G. Sutcliffe.** 1983. Gene expression in rat brain. Nucleic Acids Res. **11**:5497–5520.
25. **Nabeshima, Y.-I., Y. Fujii-Kuriyama, M. Muramatsu, and K. Ogata.** 1984. Alternative transcription and two modes of splicing result in two myosin light chains from one gene. Nature (London) **308**:333–338.
26. **Nawa, H., H. Kotani, and S. Nakanishi.** 1984. Tissue-specific generation of two preprotachykinin mRNAs from one gene by alternative RNA splicing. Nature (London) **312**:729–734.
27. **Okayama, H., and P. Berg.** 1983. A cDNA cloning vector that permits expression of cDNA inserts in mammalian cells. Mol. Cell. Biol. **3**:280–289.
28. **Parnes, J. R., and R. R. Robinson.** 1983. Multiple mRNA species with distinct 3′ termini are transcribed from the β_2-microglobulin gene. Nature (London) **302**:449–452.
29. **Prehn, J., M. Mercola, and K. Calame.** 1984. Translocation affects normal c-myc promoter usage and activates fifteen cryptic c-myc transcription starts in plasmacytoma M603. Nucleic Acids Res. **12**:8987–9007.
30. **Proudfoot, N. J., and G. G. Brownlee.** 1976. 3′ non-coding region sequences in eukaryotic messenger RNA. Nature (London) **263**:211–214.
31. **Reddy, V. B., P. K. Ghosh, P. Lebowitz, M. Piatak, and S. M. Weissman.** 1979. Simian virus 40 early mRNAs. J. Virol. **30**:279–287.
32. **Reynolds, G. A., S. K. Basu, T. F. Osborne, D. J. Chin, G. Gil, M. S. Brown, J. L. Goldstein, and K. L. Luskey.** 1984. HMG CoA Reductase: a negatively regulated gene with unusual promoter and 5′ untranslated regions. Cell **38**:275–285.
33. **Rigby, P. W. J., M. Dieckmann, C. Rhodes, and P. Berg.** 1977. Labeling deoxyribonucleic acid to high specific activity *in vitro* by nick translation with DNA polymerase I. J. Mol. Biol. **113**:237–251.
34. **Sargent, T. D, J.-R. Wu, J. M. Salatrepat, R. B. Wallace, A. A. Reyes, and J. Bonner.** 1979. The rat serum albumin gene: analysis of cloned sequences. Proc. Natl. Acad. Sci. USA **76**:3256–3260.
35. **Schibler, K., M. Tosi, A.-C. Pittet, L. Fabiani, and P. K. Wellauer.** 1980. Tissue-specific expression of mouse α-amylase genes. J. Mol. Biol. **142**:93–116.
36. **Setzer, D. R., M. McGrogan, J. H. Nunberg, and R. T. Schimke.** 1982. Nucleotide sequence surrounding multiple polyadenylation sites in the mouse dihydrofolate reductase gene. J. Biol. Chem. **257**:5143–5147.
37. **Solner-Webb, B., and R. H. Reeder.** 1979. The nucleotide sequence of the initiation and termination sites for ribosomal RNA transcription in *Xenopus laevis*. Cell **18**:485–499.
38. **Southern, E. M.** 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. **98**:503–517.
39. **Thomas, P. S.** 1980. Hybridization of denatured RNA and small DNA fragments transferred to nitrocellulose. Proc. Natl. Acad. Sci. USA **77**:5201–5205.
40. **Tosi, M., R. A. Young, O. Hagenbuchle, and U. Schibler.** 1981. Multiple polyadenylation sites in a mouse α-amylase gene. Nucleic Acids. Res. **9**:2313–2323.
41. **Young, R. A., O. Hagenbuchle, and U. Schibler.** 1981. A single mouse α-amylase gene specificies two different tissue-specific mRNAs. Cell **23**:451–458.