Proc. Natl. Acad. Sci. USA
Vol. 89, pp. 3280–3284, April 1992
Neurobiology

Structure and evolution of four POU domain genes expressed in mouse brain

(POU *Brain-1*/POU *Brain-2*/POU *Brain-4*/POU *Scip*/homeobox)

YOSHINOBU HARA*, ALESSANDRA C. ROVESCALLI, YONGSOK KIM, AND MARSHALL NIRENBERG

Laboratory of Biochemical Genetics, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD 20892

Contributed by Marshall Nirenberg, January 6, 1992

ABSTRACT **Four mouse POU domain genomic DNA clones—*Brain-1*, *Brain-2*, *Brain-4*, and *Scip*—and *Brain-2* cDNA, which are expressed in adult brain, were cloned and the coding and noncoding regions of the genes were sequenced. The amino acid sequences of the four POU domains are highly conserved; sequences in other regions of the proteins also are conserved but to a lesser extent. The absence of introns from the coding regions of the four POU domain genes and the similarity of amino acid sequences of the corresponding proteins suggest that the coding region of the ancestral class III POU domain gene lacked introns and therefore may have originated by reverse transcription of a molecule of POU domain mRNA followed by insertion of the cDNA into germ cell genomic DNA. Additional duplications of the ancestral class III POU domain gene (or mRNA) would create the *Brain-1*, *Brain-2*, *Brain-4*, and *Scip* genes.**

POU domain proteins bind to specific nucleotide sequences in DNA and regulate gene expression (for reviews, see refs. 1 and 2). The POU domain is a conserved amino acid sequence ≈150 amino acid residues long. The initial region of 69–72 amino acid residues is termed the POU-specific domain, which is followed by a 15- to 25-amino acid residue linker region and a 60-amino acid residue POU homeodomain. Both the POU-specific domain and the homeodomain are required for specific high-affinity binding to DNA (3, 4).

Rosenfeld and his colleagues have sorted POU domains into different groups on the basis of POU domain amino acid sequence similarity (2). Three mammalian class III POU domain cDNAs have been described—human (5) and rat (2) *Brain-1* and *Brain-2* and rat (5–8) and mouse (9–11) *Scip* (also termed *Oct-6* and *Tst-1*)—that have closely related POU domains and are expressed in embryonic and adult brain. Only the POU domain regions of human and rat *Brain-1* and *Brain-2* cDNA have been sequenced thus far (2, 5), whereas the complete coding sequences of rat (6–8) and mouse (9–11) *Scip* cDNA have been reported. Scip RNA is expressed in a subset of neurons, oligodendroglia, Schwann cells, and in the testis (5–11). The expression of the *Scip* gene is promoted by cAMP (6, 7).

In this report, a fourth class III mouse POU domain gene, *Brain-4*, is described, which is also expressed in adult mouse brain. *Brain-4* is similar to the recently reported *XLPOU 2* POU domain partial cDNA of *Xenopus laevis* (12). Three additional mouse class III POU domain genomic clones—*Brain-1*, *Brain-2*, and *Scip*—and *Brain-2* cDNA were obtained and the nucleotide sequences of the coding and noncoding regions were determined. Comparison of the deduced amino acid sequences of the four POU domain proteins shows that the structure of the genes and the amino acid sequences of the proteins are related to one another and suggests that the genes originated by duplication of an ancestral class III POU domain gene.†

MATERIALS AND METHODS

DNA Cloning and Sequencing. Our objective was to clone POU domain genes expressed in mouse brain. Mixtures of oligodeoxynucleotides that correspond to highly conserved amino acid sequences in POU domains were used as primers for amplification of adult mouse brain POU domain cDNA by PCR, and the amplified DNA fragments were cloned. Sequence analysis revealed 10 POU domain cDNA clones. Nick-translation of clone 38 DNA [228 base pairs (bp)] yielded a ^{32}P-labeled DNA probe that was used to screen an adult mouse brain cDNA library. One positive *Brain-2* POU domain cDNA clone (P5) was obtained (1380 bp). One million recombinants from a mouse BALB/c genomic DNA library in EMBL4, kindly provided by Konrad Huppi (National Institutes of Health) and 10^6 recombinants from an adult mouse brain cDNA library were screened with a ^{32}P-labeled *Brain-2* cDNA probe (nucleotides 1002–1609 in Fig. 4). DNA inserts were subcloned in pBluescript II SK+ and KS+. The nucleotide sequences of both strands of DNA were determined with universal or specific oligodeoxynucleotide primers and single-stranded DNA templates by the dideoxynucleotide chain-termination method (13).

Oligodeoxynucleotide Probes. Four oligodeoxynucleotides (48 bases) complementary to different sequences that encode part of the C-terminal regions of Brain-1, Brain-2, Brain-4, or Scip POU domain proteins were synthesized and purified. Each oligodeoxynucleotide was used as a specific probe for one POU domain gene; either *Brain-4* (nucleotides 1521–1568; see Fig. 2), *Brain-1* (nucleotides 1890–1937; see Fig. 3), *Brain-2* (nucleotides 1755–1802; see Fig. 4), or *Scip* (nucleotides 1117–1164 in figure 1 of ref. 9). Each oligodeoxynucleotide probe was labeled by adding ≈10 residues of [^{32}P]dATP (3000 Ci/mmol; 1 Ci = 37 GBq) to the 3′ terminus catalyzed by terminal deoxynucleotidyl transferase.

Genomic DNA Blot Analysis. Mouse genomic DNA was digested with *Eco*RI and/or *Bam*HI, subjected to electrophoresis (0.7% agarose gel, 5 μg per lane), and transferred to nitrocellulose filters. The filters were hybridized with a nick-translated ^{32}P-labeled *Brain-2* DNA probe (nucleotides 1002–1609; see Fig. 4) or an oligodeoxynucleotide probe (10^6 cpm/ml) in 4× standard saline citrate (SSC)/40% formamide/0.1% SDS/1× Denhardt's solution/25 μg of sheared salmon sperm DNA per ml at 42°C overnight. Filters were

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

*To whom reprint requests should be addressed at: Laboratory of Biochemical Genetics, National Heart, Lung, and Blood Institute, National Institutes of Health, Building 36, Room 1C06, 9000 Rockville Pike, Bethesda, MD 20892.

†The sequences reported in this paper have been deposited in the GenBank data base [accession nos. M88299 (Brain-1), M88300 (Brain-2), M88301 (Brain-4), and M88302 (Scip)].

washed three times in 1× SSC/0.1% SDS at 25°C and four times in 0.1× SSC/0.1% SDS at 42°C, 52°C, 62°C, or 72°C with a ^{32}P-labeled *Brain-2* DNA probe or at 52°C with a labeled oligodeoxynucleotide probe and were subjected to autoradiography for several days.

RNA Blot Analysis. Total RNA was prepared from various organs of 7-week-old mice by the guanidine isothiocyanate lysis method (14). Total RNA was fractionated by electrophoresis (10 μg per lane) on 1.2% agarose formaldehyde gels and RNA was transferred to nitrocellulose filters. The filters were hybridized with an oligodeoxynucleotide probe (10^6 cpm/ml) specific for either Brain-1, Brain-2, Brain-4, or Scip RNA in 4× SSC/40% formamide/0.1% SDS/1× Denhardt's solution/100 μg of yeast tRNA per ml/25 μg of sheared denatured salmon sperm DNA per ml at 42°C overnight. The filters were washed three times with 1× SSC/0.1% SDS at 25°C and four times with 0.1× SSC/0.1% SDS at 52°C and then were subjected to autoradiography for several days.

RESULTS AND DISCUSSION

Southern Analysis and POU Domain Probe Specificity. A *Brain-2* POU domain cDNA clone was isolated from an adult mouse brain cDNA library and the POU domain region of the DNA was used as a probe to detect POU domain genes. The specificity of the *Brain-2* DNA probe for hybridization to POU domain genes in mouse genomic DNA digested with restriction enzymes and subjected to Southern analysis is shown in Fig. 1*A* as a function of the temperature used for stringent washes of filters. Only one DNA fragment was detected when the filters were washed at 72°C; however, at least four DNA fragments were detected at 42°C–62°C. The specificity of four oligodeoxynucleotide probes complementary to different sequences in *Brain-1*, *Brain-2*, *Brain-4*, or *Scip* POU domain genes was examined by Southern analysis using a stringent wash temperature of 52°C (Fig. 1*B*). Each oligodeoxynucleotide probe hybridized to a different, single DNA fragment.

Expression of Four POU Domain Genes. The expression of *Brain-1*, *Brain-2*, *Brain-4*, and *Scip* POU domain genes was determined by Northern analysis with total RNA from various adult mouse tissues and oligodeoxynucleotide probes specific for Brain-1, Brain-2, Brain-4, or Scip RNA (Fig. 1*C*). Two major species of Brain-1 RNA [4.8 and 3.5 kilobases (kb)] were detected in RNA from brain and kidney and two minor species (8.0 and 1.8 kb) were detected in brain. One major species of Brain-2 RNA (4.8 kb) and 2 minor species of RNA (3.5 and 7.0 kb) were detected only in brain RNA. One species of Brain-4 RNA (4.8 kb) was found only in RNA from brain. The Scip probe revealed one major and one minor species of RNA from brain (3.5 and 2.5 kb, respectively).

Isolation of Four Mouse POU Domain Genes. One million recombinants from a mouse genomic DNA library and one million from an adult mouse brain cDNA library were screened with a *Brain-2* cDNA probe at low stringent washes (0.1× SSC at 47°C). Sixty positive genomic DNA and 50 cDNA clones were obtained. Restriction site analysis of the 60 genomic DNA clones (data not shown) revealed 11 kinds of clones. Thus far, *Brain-4*, *Brain-1*, *Brain-2*, and *Scip* POU domain genomic DNA clones and *Brain-2* cDNA clones have been identified by nucleotide sequence analysis.

***Brain-4* POU Domain Gene.** The nucleotide sequence (2500 bases) of cloned mouse *Brain-4* POU domain genomic DNA and the deduced amino acid sequence of the protein are shown in Fig. 2. An open reading frame was found for a POU domain protein consisting of 361 amino acid residues with a calculated M_r of 39,417. No intron or typical RNA splice site was found in the coding sequence. Since only 2.5 kb of *Brain-4* genomic DNA has been sequenced, the possibility of introns in the noncoding regions of the gene is not excluded. The 3′ noncoding region of the *Brain-4* gene contains repetitive AC and GC nucleotide sequences (Fig. 2, underlined regions), which under appropriate conditions might adopt the conformation of Z-DNA. The amino acid sequence of the Brain-4 POU domain is similar to the POU domain sequence recently reported (12) for XLPOU 2 of *X. laevis* (98% similarity); 79% similarity was found for 90 amino acid residues outside the POU domain. The 3′ untranslated nucleotide sequence of *XLPOU 2* cDNA has a repetitive AC nucleotide sequence similar to that of *Brain-4* and an AT repeat. No other obvious sequence similarity was found in the 3′ noncoding regions of *Brain-4* and *XLPOU 2* cDNA compared. A rat cDNA clone (*RHS2*) that is the equivalent of *Brain-4* is described in the accompanying paper (15).

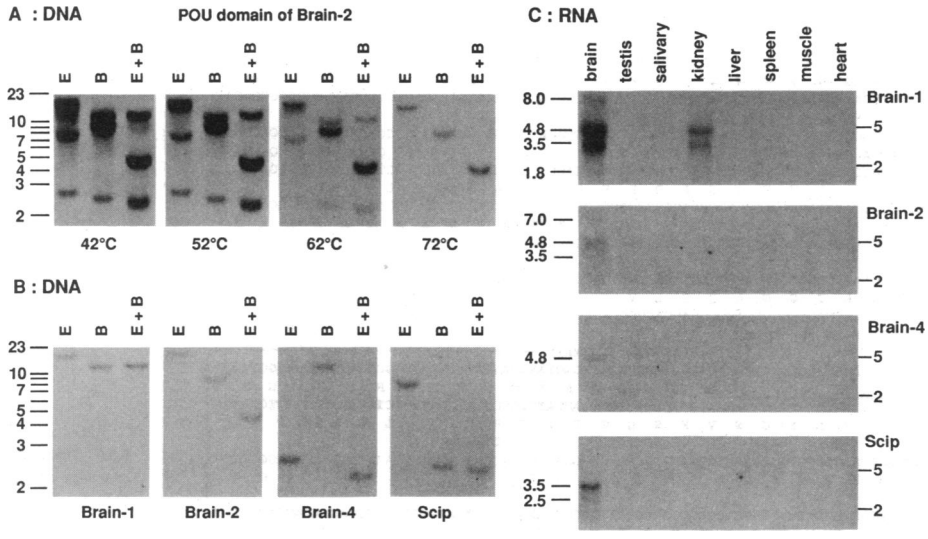

FIG. 1. Probe specificity and Southern and Northern analyses of *Brain-1*, *Brain-2*, *Brain-4*, and *Scip* DNA or RNA. (*A*) Southern analysis of mouse genomic DNA cleaved with *Eco*RI (E), *Bam*HI (B), or *Eco*RI and *Bam*HI (E + B). POU domain DNA fragments were detected by hybridization with a ^{32}P-labeled *Brain-2* POU domain cDNA probe. For the stringent washes of the filters, 0.1× SSC was used at 42°C, 52°C, 62°C, or 72°C. (*B*) The specificity of ^{32}P-labeled oligodeoxynucleotide probes for *Brain-1*, *Brain-2*, *Brain-4*, or *Scrip* genes in the genomic Southern analysis. (*C*) Northern analysis of total RNA from adult mouse tissues. Brain-1, Brain-2, Brain-4, or Scip RNA was detected with ^{32}P-labeled oligodeoxynucleotide probes.

FIG. 2. Nucleotide sequence and predicted amino acid sequence of mouse *Brain-4* genomic DNA. The POU domain is enclosed within a box and the corresponding nucleotide and amino acid sequences of the POU-specific domain and POU homeodomain are shown in boldface type. Underlined nucleotides are described in the text.

***Brain-1* POU Domain Gene.** A 15.5-kb mouse *Brain-1* genomic DNA clone was obtained and 4000 nucleotide residues were sequenced. The nucleotide sequence (2500 bases) of *Brain-1* genomic DNA and the deduced amino acid sequence of Brain-1 POU domain protein are shown in Fig. 3. The *Brain-1* gene contains an open reading frame for 495 amino acid residues, which corresponds to a POU domain protein with a calculated M_r of 50,012. Mouse Brain-1 POU domain protein contains long amino acid repeats consisting of glycine, alanine, proline, or histidine. In most cases, multiple copies of a single codon determine the repetitive amino acid sequence. The 5′ noncoding region of the *Brain-1* gene contains polypyrimidine and polypurine regions and repetitive GA nucleotide sequences (not shown in Fig. 3) as well as repetitive GGC nucleotide sequences. The 3′ noncoding region contains repetitive GCC nucleotide sequences and a polyadenylylation signal starting at nucleotide 2323. No intron or typical RNA splice site was detected in the coding sequence of Brain-1.

***Brain-2* POU Domain cDNA and Genomic DNA Clones.** An 18.5-kb clone of *Brain-2* genomic DNA and multiple *Brain-2* cDNA clones were obtained; 3864 nucleotides of genomic DNA and 1461 residues of cDNA were sequenced. The nucleotide sequence of *Brain-2* cDNA clone C4 corresponds to nucleotides 170–1631 of *Brain-2* genomic DNA shown in Fig. 4. An open reading frame (1335 nucleotides) was found that encodes a 445-amino acid protein with a calculated M_r of

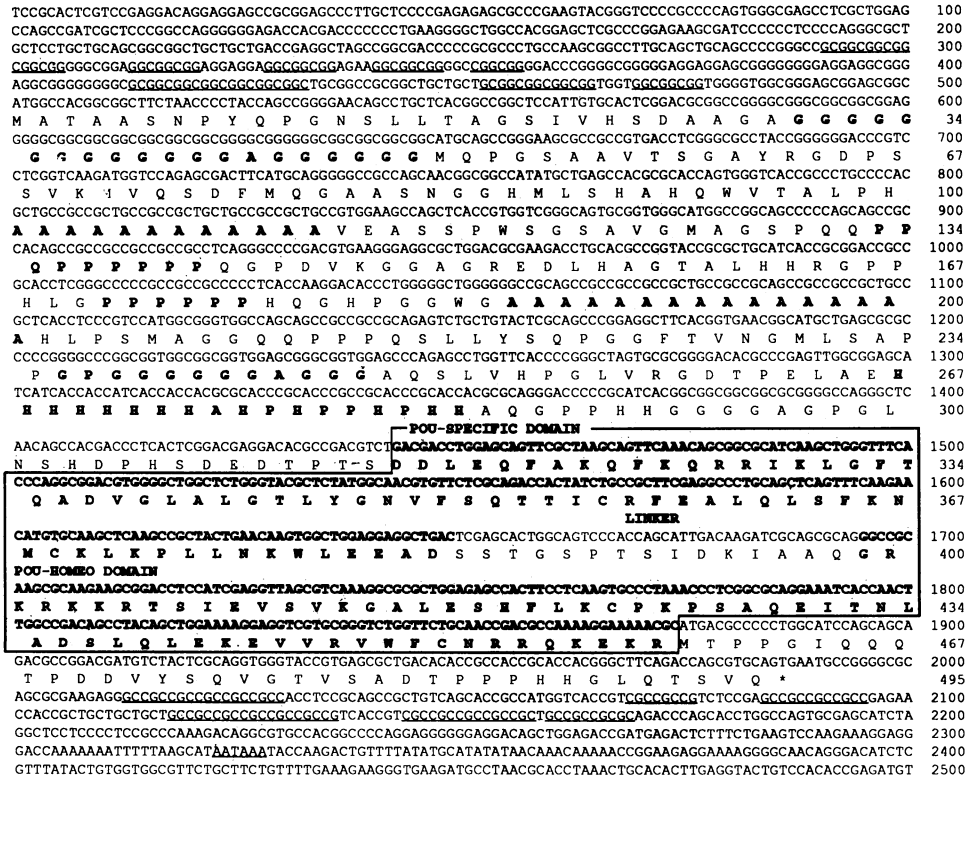

FIG. 3. Nucleotide sequence and predicted amino acid sequence of mouse *Brain-1* genomic DNA. The POU domain is enclosed in a box. The amino acid repeats are shown in boldface type. The nucleotide and amino acid sequences of the POU-specific domain and POU homeodomain are also shown in boldface type. Underlined nucleotides are described in the text.

47,148. Identical nucleotide sequences were found with *Brain-2* cDNA and genomic DNA; hence, the portion of *Brain-2* genomic DNA that corresponds to the *Brain-2* cDNA does not contain an intron. Brain-2 protein contains repetitive residues of glycine, glutamic acid, and proline. The first in-frame codon for methionine in the open reading frame of the cDNA is shown in Fig. 4 as the putative codon for initiation of protein synthesis. Two overlapping polyadenylation signals are present starting at nucleotide 2262. The 5′ noncoding region of the gene contains 32 consecutive GT repeats (not shown in Fig. 4) and repetitive GA nucleotide sequences. The 3′ noncoding region of the *Brain-2* gene contains repetitive GT, GA, and AC nucleotide sequences.

***Scip* POU Domain Gene.** A fourth mouse POU domain gene, the *Scip* gene, was cloned and 2766 nucleotides were sequenced (not shown here). The nucleotide sequence found for mouse *Scip* genomic DNA confirms the sequence that was reported for mouse *Scip* cDNA (9–11). No intron was detected in the coding sequence of the *Scip* gene.

Sequence Similarity. A comparison of the amino acid sequences of Brain-1, Brain-2, Brain-4, and Scip POU domain proteins is shown in Fig. 5. The four proteins clearly are related to one another. The POU domain is the most highly conserved region of each protein; however, many other regions of similarity are present. Brain-1, Brain-2, and Scip, but not Brain-4, proteins contain amino acid repeats 5–27 amino acids long that are unique, rather than conserved, parts of the proteins. Similar di- and trinucleotide repeats are present in the 5′ and 3′ noncoding regions of these genes but no other obvious sequence similarity was found in the noncoding regions compared.

Putative phosphorylation sites for different kinds of protein kinases also are shown in Fig. 5. Many highly conserved putative phosphorylation sites are present in or near the POU domains of the four proteins and in the N-terminal regions. The regions immediately before and after the POU-specific domain contain many highly conserved acidic amino acid residues and some serine or threonine residues that are putative sites for phosphorylation. If fully phosphorylated, 7–9 of the 14 or 15 amino acid residues before and after the POU-specific domain would be acidic. These POU domain proteins contain highly conserved putative phosphorylation sites for protein kinase C, cGMP-dependent protein kinase, and cAMP-dependent protein kinase known to be regulated by intracellular levels of calcium ions, cGMP, or cAMP, respectively. The possibility that the rate of transsynaptic communication and the rate of expression of certain genes may be coupled by phosphorylation of POU domain proteins, which may alter the ability of the protein to regulate genes, is a problem for future study.

Protein–protein interactions between POU domain genes have been reported in some cases (18, 19). Homo- and heterodimer formation by *Brain-1*, *Brain-2*, *Brain-4*, and *Scip* might generate proteins with different properties or specificities for regulating the expression of subsets of genes.

Class III POU domain genes have been found in nematodes (20), *Drosophila* (18, 21), amphibians (12), and mammals (5–11), which suggests that the ancestral class III POU domain gene originated at least 6×10^8 years ago. The absence of introns from the coding regions of the four mouse POU domain genes and the similarity of amino acid sequences of the corresponding proteins suggests that the coding sequence of the ancestral mouse class III POU domain gene lacked introns and therefore may have originated by reverse transcription of a molecule of POU domain mRNA, followed by insertion of the cDNA into germ cell genomic DNA. Thus, the coding sequence of the POU domain gene would be duplicated, but not the introns or the 5′ upstream regulatory region of the original gene. The DNA sequences that regulate expression of the original POU domain gene would be replaced by the regulatory sequences of another gene near the cDNA insertion site. The new and the original POU domain genes might well be expressed in different cell types and at different times during development. It is likely that expression of the newly created class III POU domain gene would, in a combinatorial fashion, create a new set of gene regulatory proteins that might interact in different ways compared to the original set. Additional duplications of the ancestral class III POU domain gene (or mRNA) would create the *Brain-1*, *Brain-2*, *Brain-4*, and *Scip* genes. We suggest that other sets of gene regulators may have originated during evolution by formation of chimeric genes by splicing enhancer and promoter DNA sequences from one gene to DNA from a second gene that encodes a protein that regu-

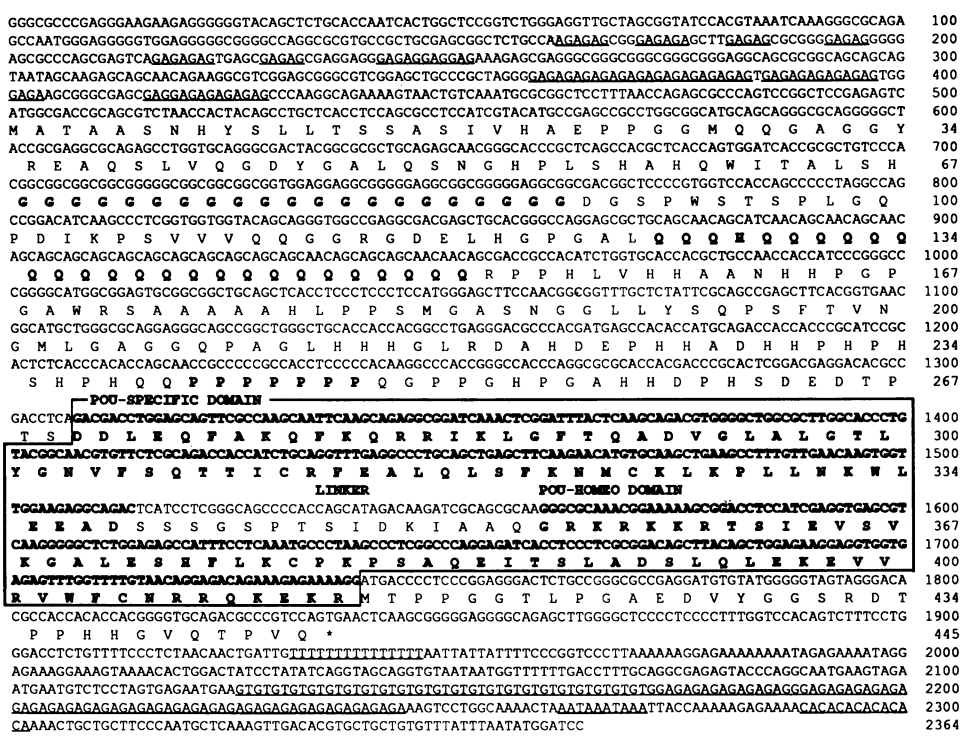

FIG. 4. Nucleotide sequence and predicted amino acid sequence of mouse *Brain-2* genomic DNA. The POU domain is enclosed in a box. The amino acid repeats are shown in boldface type. The nucleotide and amino acid sequences of the POU-specific domain and POU homeodomain are also shown in boldface type. Underlined nucleotides are described in the text.

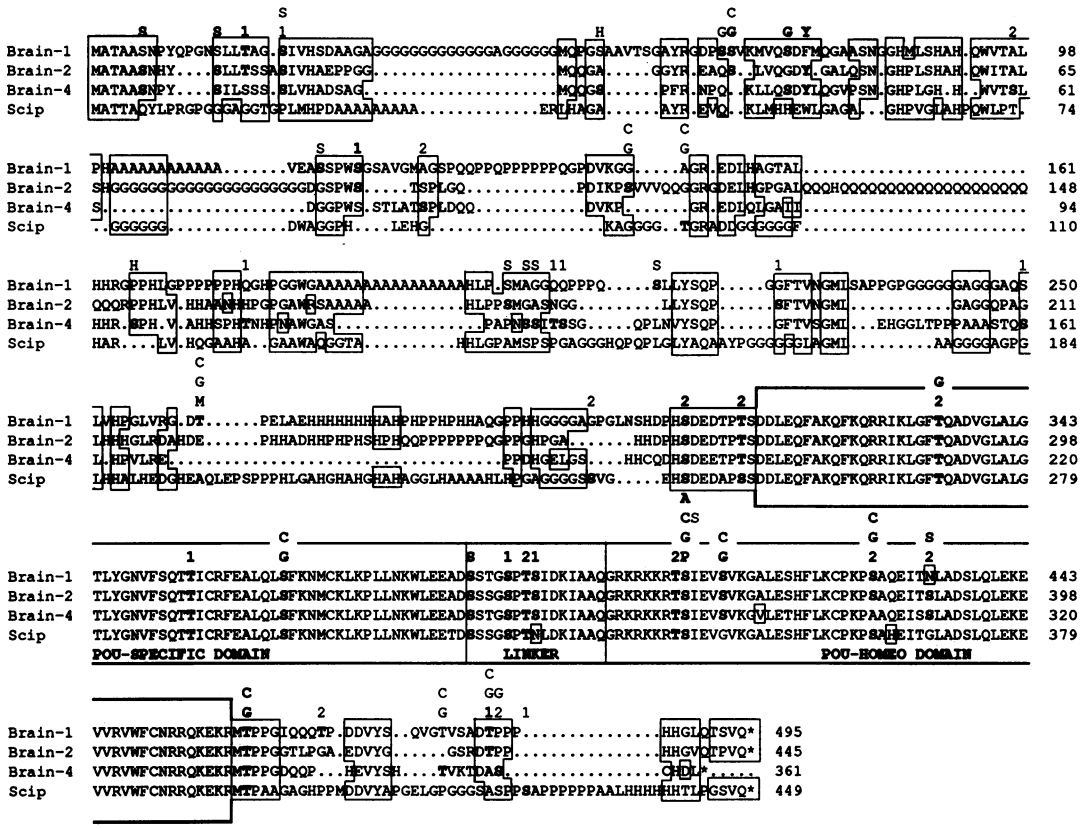

FIG. 5. Comparison of amino acid sequences of mouse Brain-1, Brain-2, Brain-4, and Scip POU domain proteins. Boxes represent amino acid sequence similarities. The following criteria for a box apply to two or more consecutive amino acid residues: (*i*) three or four proteins contain the same amino acid residue, (*ii*) at least two proteins contain the same amino acid residue and a third protein has a conservative amino acid replacement. Conservative amino acid replacement families defined by Dayhoff *et al.* (16) are as follows: (*i*) L, I, M, V; (*ii*) G, A, S, P, T; (*iii*) F, Y, W; (*iv*) E, D, Q, N; (*v*) R, K, H; (*vi*) C. Each dot represents a gap. S, T, and Y residues shown in boldface type correspond to putative consensus phosphorylation sites; each letter or number above the site corresponds to one of the following protein kinase abbreviations: A, cAMP-dependent protein kinase; G, cGMP-dependent protein kinase; C, protein kinase C; H, growth-associated histone H1 kinase; M, calmodulin-dependent protein kinase II; P, phosphorylase kinase; S, glycogen synthase kinase-3; Y, tyrosine protein kinase; 1, casein kinase I; 2, casein kinase II. Consensus phosphorylation sites are described by Pearson and Kemp (table II in ref. 17). Putative phosphorylation sites present in two or more proteins are shown in boldface type. The amino acid sequence of mouse Scip protein was deduced from the nucleotide sequence obtained for mouse genomic DNA; the data confirm the sequence reported for mouse *Scip* cDNA (9–11).

lates gene expression. The most effective sets of gene regulators would be retained by selection.

We thank K. Huppi for the gift of a mouse genomic DNA library, C. Le Moine and W. S. Young for exchanging sequence information prior to publication, and A. Peterkofsky for comments on the manuscript.

1. Herr, W., Sturm, R. A., Clerc, R. G., Corcoran, L. M., Baltimore, D., Sharp, P. A., Ingraham, H. A., Rosenfeld, M. G., Finney, M., Ruvkun, G. & Horvitz, H. R. (1988) *Genes Dev.* **2,** 1513–1516.
2. Rosenfeld, M. G. (1991) *Genes Dev.* **5,** 897–907.
3. Ingraham, H. A., Flynn, S. E., Voss, J. W., Albert, V. R., Kapiloff, M. S., Wilson, L. & Rosenfeld, M. G. (1990) *Cell* **61,** 1021–1033.
4. Verrijzer, C. P., Kal, A. J. & van der Vliet, P. C. (1990) *Genes Dev.* **4,** 1964–1974.
5. He, X., Treacy, M. N., Simmons, D. M., Ingraham, H. A., Swanson, L. W. & Rosenfeld, M. G. (1989) *Nature (London)* **340,** 35–42.
6. Monuki, E. S., Weinmaster, G., Kuhn, R. & Lemke, G. (1989) *Neuron* **3,** 783–793.
7. Monuki, E. S., Kuhn, R., Weinmaster, G., Trapp, B. & Lemke, G. (1990) *Science* **249,** 1300–1303.
8. He, X., Gerrero, R., Simmons, D. M., Park, R. E., Lin, C. R., Swanson, L. W. & Rosenfeld, M. G. (1991) *Mol. Cell. Biol.* **11,** 1739–1744.
9. Susuki, N., Rohdewohld, H., Neuman, T., Gruss, P. & Schöler, H. R. (1990) *EMBO J.* **9,** 3723–3732.
10. Meijer, D., Graus, A., Kraay, R., Langeveld, A., Mulder, M. P. & Grosveld, G. (1990) *Nucleic Acids Res.* **19,** 7357–7365.
11. Zimmerman, E. C., Jones, C. M., Fet, M., Hogan, B. L. M. & Magnuson, M. A. (1991) *Nucleic Acids Res.* **19,** 956.
12. Agarwal, V. R. & Sato, S. M. (1991) *Dev. Biol.* **147,** 363–373.
13. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5463–5467.
14. Davis, L. G., Dibner, M. D., & Battey, J. F. (1986) *Basic Methods in Molecular Biology* (Elsevier, New York), p. 130.
15. Le Moine, C. & Young, S. W., III (1992) *Proc. Natl. Acad. Sci. USA* **89,** 3285–3289.
16. Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. (1979) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. O. (Natl. Biomed. Res. Found., Washington), pp. 345–352.
17. Pearson, R. & Kemp, B. E. (1991) *Methods Enzymol.* **200,** 62–81.
18. Treacy, M. N., He, X. & Rosenfeld, M. G. (1991) *Nature (London)* **350,** 577–584.
19. Voss, J. W., Wilsom, L. & Rosenfeld, M. G. (1991) *Genes Dev.* **5,** 1309–1320.
20. Bürglin, T. R., Finney, M., Coulson, A. & Ruvkun, G. (1989) *Nature (London)* **341,** 239–243.
21. Johnson, W. A. & Hirsh, J. (1990) *Nature (London)* **343,** 467–470.