Reprint Series
15 January 1988, Volume 239, pp. 285–288

SCIENCE

The Primary Structure and Heterogeneity of Tau Protein from Mouse Brain

GLORIA LEE,* NICHOLAS COWAN, AND MARC KIRSCHNER

The Primary Structure and Heterogeneity of Tau Protein from Mouse Brain

GLORIA LEE,* NICHOLAS COWAN, MARC KIRSCHNER

Tau protein is a family of microtubule binding proteins, heterogeneous in molecular weight, that are induced during neurite outgrowth and are found prominently in neurofibrillary tangles in Alzheimer's disease. The predicted amino acid sequences of two forms of tau protein from mouse brain were determined from complementary DNA clones. These forms are identical in their amino-terminal sequences but differ in their carboxyl-terminal domains. Both proteins contain repeated sequences that may be tubulin binding sites. The sequence suggests that tau is an elongated molecule with no extensive α-helical or β-sheet domains. These complementary DNAs should enable the study of various functional domains of tau and the study of tau expression in normal and pathological states.

MICROTUBULES ARE ASSEMBLED from tubulin, which is a dimer of two polypeptides that are members of distinct multigene families (*1*). A high degree of conservation exists within these families and the various polypeptides form copolymers in vivo and in vitro (*2*). Despite this similarity of tubulin polypeptides, microtubules exhibit much diversity in structure and function, suggesting that other proteins must be present that determine the properties of different microtubules. Among the factors thought to regulate microtubule structure and function are the microtubule-associated proteins (MAPs) that copurify with microtubules (*3*). Two major classes of MAPs have been identified from vertebrate brain: high molecular weight MAPs and tau protein. Tau protein promotes microtubule assembly in vitro and limits the growing and shrinking phases of dynamic microtubules (*4*). Tau colocalizes with microtubules in cells (*5*) and is induced along with MAP1 during neurite outgrowth from rat pheochromocytoma cells (*6*). Microinjected tau protein increases tubulin polymerization and decreases the rate of microtubule depolymerization, suggesting that tau protein can regulate microtubule assembly in vivo (*7*).

A striking feature of tau protein is its extensive heterogeneity. In adult porcine brain, it is comprised of at least four related phosphoproteins, 55,000 to 62,000 daltons in size (*8*). Tau proteins were initially thought to be the result of artifactual proteolysis of a common precursor protein; however, translation of messenger RNA (mRNA) in vitro shows that this is not the case (*9*). Other proteins reacting with tau antibodies have been detected in brain, neuroblastoma cells, spinal ganglia, and coated vesicles (*6, 9, 10*). In addition, tau protein

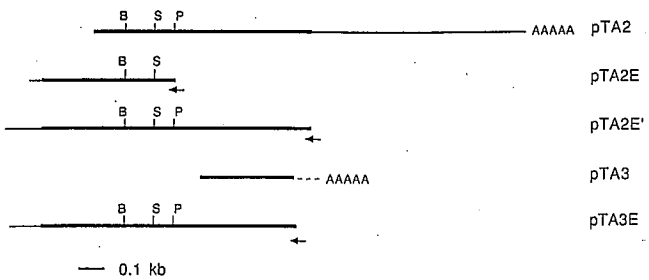

Fig. 1. Schematic representation of cDNA clones used to determine tau protein sequence. Heavy line indicates open reading frame regions, thinner line and dotted line indicate noncoding regions of pTA2 and pTA3, respectively. The pTA2 and pTA2E are pBR322 clones, pTA3 is a λgt11 clone, and pTA3E and pTA2E′ are pUC9 clones. Restriction sites are Bam HI (B), Sma I (S), and Pst I (P). AAAAA indicates a poly(A) stretch of 18 to 19 bases. Arrows indicate the location of specific oligonucleotide primers used to initiate cDNA synthesis for library construction. The isolation of pTA2 was as described (*9*). The pTA2E was isolated from a pBR322 primer extension library constructed with cDNA primed by a 22-base oligonucleotide (5′-GACATTCTTTAGGTCTGGCATG-3′) (*20*). The pTA3 was isolated as described (*21*); antibody employed was affinity-purified anti-tau (*10*). The pTA3E was isolated from a pUC9 primer extension library (*22*) with size-selected cDNA primed by a 21-base oligonucleotide (5′-TTGACTGCCCTGGGAGCCTGA-3′). Two additional libraries were constructed in the manner described for the pTA3E library: one primed with a 21-base oligonucleotide from the 3′ untranslated region of pTA2 (5′-GGCAGAGGTCCCCCAAGAGGC-3′), from which pTA2E′ clones were isolated, and the other primed with the 22-base oligonucleotide used for the pBR322 library above. In cDNA synthesis, primers were preincubated with mRNA prior to reverse transcriptase reaction. The cDNAs were dC-tailed for insertion into dG-tailed plasmid vectors.

G. Lee, Department of Neurology and Program in Neuroscience, Harvard Medical School, Boston, MA 02115.
N. Cowan, Department of Biochemistry, New York University Medical Center, New York, NY 10016.
M. Kirschner, Department of Biochemistry and Biophysics, University of California, San Francisco, CA 94143-0448.

*To whom correspondence should be addressed.

has been identified as a major antigenic determinant in the characteristic neurofibrillary tangles of Alzheimer's disease (*11*). In addition to the immunological cross-reactivity detected between tau protein and the paired helical filaments that comprise the tangles, at least two peptides common to the filaments and human tau protein have been found (*12*).

In this report we examine the complete primary sequence of two tau proteins from mouse for information about the structure of tau, the possible mode of interaction between tau and tubulin, and the source of tau heterogeneity. We also describe special difficulties encountered when determining the structures of members of a closely related class of mRNAs.

Earlier studies had identified two complementary DNA (cDNA) clones for tau protein in mouse, pTA1 and pTA2. Both hybridized to a 6-kb mRNA and selected mRNA that gave tau protein on translation, but did not hybridize to each other (*9*). From the nucleotide sequence, we found that pTA1 (1376 bp) contains a poly(A) tail and a polyadenylation signal, AATAAA, 15 bp upstream of the poly(A) tail, but contains no open reading frame. On the other hand, pTA2 (1840 bp) also contains a poly(A) tail but no polyadenylation signal, and has an open reading frame of 900 bases at one end of the clone. We conclude that pTA1 corresponds to the 3′ end of the mRNA while pTA2 originated from the interior of the mRNA, primed from an internal poly(A) sequence, and encodes the COOH-terminal end of a protein. To complete the tau sequence, we constructed additional libraries using a specific primer from the 5′ end of pTA2 sequence. A 600-bp clone, pTA2E (Fig. 1), contained 288 bases 5′ to pTA2, an ATG start codon, and an

Fig. 2. Nucleotide sequence determined from tau cDNA clones. Numbers at left designate nucleotide base number, with the first base of the initiation codon as reference point (base 1). The predicted amino acid sequence is numbered with position 1 being the NH_2-terminal methionine. The pTA2 sequence, shown in (**A**), is determined from pTA2, pTA2E, and pTA2E′ clones. The pTA3 sequence is determined from pTA3 and pTA3E clones and is identical to pTA2 up through base 1020; only bases 1007 to 1101 are shown in (**B**). The 21- to 22-base oligonucleotides used to specifically prime cDNA library constructions are underlined. The underlined 18–amino acid stretches indicate repeats. The pTA2 open reading frame is 364 residues with a calculated *p*I of 6.24; pTA3 is 341 residues with a calculated *p*I of 6.27. Nucleotide sequences were determined by dideoxy chain termination (*23*). Problematical stretches were further analyzed by either the use of 7-deaza-2′dGTP (Boehringer Mannheim) or by the substitution of AMV reverse transcriptase (Bio-Rad) for Klenow DNA polymerase fragment in dideoxy sequencing, with appropriate adjustments of nucleotide concentrations (*24*). Coding region sequences were confirmed by sequencing the opposite strand.

```
(1) Val Arg Ser Lys Ile Gly Ser Thr Glu Asn Leu Lys His Gln Pro Gly Gly Gly
(2) Val Thr Ser Lys Cys Gly Ser Leu Gly Asn Ile His His Lys Pro Gly Gly Gly
(3) Val Gln Ser Lys Ile Gly Ser Leu Asp Asn Ile Thr His Val Pro Gly Gly Gly
    *  -  *  *  +  *  *  +  -  *  +  -  *  -  *  *  *  *
```

Fig. 3. Three 18–amino acid repeats identified in the predicted tau protein sequence. Repeat (1) corresponds to amino acid residues 187 to 204 as numbered in Fig. 2; repeat (2) corresponds to residues 218 to 235; repeat (3) corresponds to residues 250 to 267. The asterisks indicate positions with identical residues in all three repeats; the pluses, conservative replacements; and the dashes, nonconservative replacements.

upstream in-frame stop codon. The sequence of pTA2 and pTA2E in Fig. 2A predicts a size for the encoded protein of 38,204 daltons.

To isolate additional clones for tau protein, we screened a λgt11 mouse brain library with affinity-purified antibody to tau protein and obtained an immunoreactive clone, pTA3, that contained an insert of 500 bp. The sequence of pTA3 contained a 19-base-long poly(A) stretch and an open reading frame of 400 bases that was identical to bases 621 to 1020 of pTA2 (Fig. 1). This clone, therefore, provided independent confirmation of pTA2 as a tau cDNA clone. However, the remaining sequences in pTA3 did not correspond to pTA2 sequences. The encoded protein differs from pTA2 in that it lacks 23 amino acid residues at the COOH-terminus; the 3′ untranslated regions also differ. To determine whether these differences resulted from a cloning artifact where pTA2 sequences had become joined to unrelated sequences and, also, whether additional regions of heterogeneity were present, an oligonucleotide specific for the 3′ untranslated region of pTA3 (Fig. 2B) was used to prime a cDNA library from which additional tau clones were isolated.

Out of a pUC9 primer extension library of approximately 1500 transformants, 54 colonies hybridized to pTA3. Since the abundance of tau mRNA has been estimated at 0.1% (*9*), this argued that the 3′ untranslated region in pTA3 used to prime the library was indeed part of tau mRNA and not a cloning artifact. Over half of the clones were longer than 1 kb and all had restriction maps similar to pTA2 (Fig. 1). The nucleotide sequence for the longest clone corresponded exactly to pTA2-pTA2E sequences, apart from the divergent COOH-terminus. Since this clone, pTA3E, included the start point of translation, the only difference in the two predicted tau protein sequences is the COOH-terminal end. The pTA3 sequence is shown in Fig. 2B; the encoded protein has a predicted size of 35,718 daltons.

At this point, it was important to show that the pTA2 COOH-terminus was in fact associated with the NH_2-terminal sequence identified by pTA2E. Because the synthesis of pTA2E cDNA had been specifically primed from the NH_2-terminal half of the protein sequence, there was the possibility that the pTA2 COOH-terminus was associated with another NH_2-terminus whose coding sequence diverged from the pTA2E sequence upstream of (or 5′ to) the oligonucleotide sequence used to prime pTA2E. The fact that the sequence of pTA3E coincided with pTA2E made it possible for the pTA2E clone to have been synthesized from pTA3-pTA3E mRNA. To resolve this issue, two additional primer extension libraries were constructed. One was specifically primed with an oligonucleotide copied from the pTA2 3′ untranslated region (Fig. 2A) while the other was specifically primed with the same oligonucleotide used to construct the library from which pTA2E was isolated. From these libraries, six clones containing the most 5′ sequences were analyzed by sequencing. These clones revealed no additional heterogeneity in the NH_2-terminal sequences.

The complete predicted amino acid sequence for the two mouse tau proteins is shown in Fig. 2. A distinctive feature is the presence of an 18-residue stretch that is repeated three times. The repeats are located in the COOH-terminal half of the molecule and are separated by 13- and 14-residue stretches (Figs. 2 and 3). The significance of the repeat is unclear, though it is tempting to speculate that since tau protein induces the tubulin monomer to assemble, it may interact with repeating sites in the microtubule lattice. We found no significant sequence homology to any other protein (*13*), although the primary structure of other MAPs is not available.

The tau sequence supports the biophysical data suggesting an elongated shape for tau protein. The amino acid composition, which agrees well with that reported for porcine brain tau protein (*8*), shows that tau protein has much higher proportions of lysine, glycine, proline, and serine and lower proportions of phenylalanine and leucine than the average vertebrate globular protein. This suggests that tau protein has less buried or interior volume, is more extended and hydrophilic, and might maintain this shape because of the rigidity introduced by the prolines. Lastly, no extensive α-helix or β-sheet structures are detected by secondary structure prediction programs (*13*), which is consistent with circular dichroism measurements (*8*). Two other facts about tau protein are consistent with its being an extended molecule with a large surface to volume ratio: (i) the protein is heat-stable (*14*), implying that, in its native form, many of the residues are on the surface interacting with solvent, and (ii) the protein migrates in SDS gel electrophoresis as a much larger protein. [The size of tau protein as predicted from the cDNA clones is 35,718 and 38,204 daltons; as determined by gel electrophoresis, mouse tau protein is 47,000 to 50,000 daltons (*9*).] While it is known that phosphorylation affects the mobility of tau protein (*15*), it is more likely that protein structure and SDS binding make the larger contributions to the anomalous electrophoretic mobility.

The identification of two distinct tau cDNA clones suggests that tau heterogeneity is already present, at least in part, at the mRNA level; this is an expected result since mRNA from mouse and rat brain has been shown to yield multiple tau proteins by in vitro translation experiments (*9, 16*). The predicted amino acid sequences from the two clones have revealed the COOH-terminal end of the protein as a site for heterogeneity in the protein; the function of this heterogeneity remains unknown. It is also not known whether the two mRNAs are transcribed from the same gene, but Southern hybridization with pTA2 has revealed only one copy of the tau gene in the mouse genome (*9*). It seems likely that the two mRNAs result from alternative splicing since the point at which pTA2 sequence diverges from pTA3 contains the consensus sequence for splicing junctions (5′CAG G) (*17*). Additional heterogeneity may also result from translation initiation at alternative sites (*18*); the tau sequence has two other methionines close to the NH_2-terminal end.

It is curious that both pTA2 and pTA3 mRNAs should be approximately 6 kb long when it is clear that much of the 3′ untranslated regions differ. However, an examination of 3′ untranslated sequences from actin cDNAs has revealed that mRNAs coding for isotypic proteins can have 3′ untranslated regions of similar size, though differing in sequence; this may suggest a functional importance for 3′ noncoding sequences in mRNA (*19*). Furthermore, both mRNAs have internal poly(A) sequences in the 3′ untranslated region, though each is located at different distances downstream of the stop codon.

Studies of tau protein at the primary sequence level could provide an explanation for the heterogeneity of microtubules and provide important probes for studying the function of various domains on the molecule. It will be interesting to test directly whether the 18–amino acid repeats are in

fact tubulin binding domains and whether there are any common structural motifs in microtubule-associated proteins. Studies of the expression and structure of tau proteins in Alzheimer's disease should also provide important clues to the etiology of the neurofibrillary tangles.

REFERENCES AND NOTES

1. D. W. Cleveland and K. F. Sullivan, *Annu. Rev. Biochem.* 54, 331 (1985).
2. S. A. Lewis, W. Gu, N. J. Cowan, *Cell* 49, 539 (1987).
3. J. B. Olmsted, *Annu. Rev. Cell Biol.* 2, 421 (1986).
4. T. Horio and H. Hotani, *Nature (London)* 321, 605 (1986).
5. J. Connolly, V. Kalnins, D. Cleveland, M. Kirschner, *Proc. Natl. Acad. Sci. U.S.A.* 74, 2437 (1977); J. A. Connolly, V. I. Kalnins, D. W. Cleveland, M. W. Kirschner, *J. Cell Biol.* 76, 781 (1978); J. A. Connolly and V. I. Kalnins, *Exp. Cell Res.* 127, 341 (1980).
6. D. G. Drubin, S. C. Feinstein, E. N. Shooter, M. W. Kirschner, *J. Cell Biol.* 101, 1799 (1985); D. Drubin, S. Kobayashi, M. Kirschner, *Ann. N.Y. Acad. Sci.* 466, 257 (1986).
7. D. G. Drubin and M. W. Kirschner, *J. Cell Biol.* 103, 2739 (1986).
8. D. W. Cleveland, S.-Y. Hwo, M. W. Kirschner, *J. Mol. Biol.* 116, 207 (1977); D. W. Cleveland, S.-Y. Hwo, M. W. Kirschner, *ibid.*, p. 227.
9. D. G. Drubin, D. Caput, M. W. Kirschner, *J. Cell Biol.* 98, 1090 (1984).
10. S. R. Pfeffer, D. G. Drubin, R. B. Kelly, *ibid.* 97, 40 (1983).
11. J. G. Wood, S. S. Mirra, N. J. Pollock, L. Binder, *Proc. Natl. Acad. Sci. U.S.A.* 83, 4040 (1986); K. S. Kosik, C. L. Joachim, D. J. Selkoe, *ibid.*, p. 4044; I. Grundke-Iqbal *et al.*, *J. Biol. Chem.* 261, 6084 (1986); N. Nukina and Y. Ihara, *J. Biochem.* 99, 1541 (1986).
12. Y. Ihara, *J. Neurochem.* 48, S14 (1987).
13. DNA sequence information was keyed into an online connection with Bionet (Intelligenetics, Mountain View, CA). The National Biomedical Research Foundation Protein Sequence Data Base was scanned for sequence homology and Chou Fasman algorithms were used for secondary structure prediction.
14. M. D. Weingarten, A. H. Lockwood, S.-Y. Hwo, M. W. Kirschner, *Proc. Natl. Acad. Sci. U.S.A.* 72, 1858 (1975); A. Fellous, J. Francon, A. M. Lennon, J. Nunez, *Eur. J. Biochem.* 78, 167 (1977); W. Herzog and K. Weber, *ibid.* 92, 1 (1978).
15. G. Lindwall and R. D. Cole, *J. Biol. Chem.* 259, 5301 (1984).
16. I. Ginzburg, T. Scherson, D. Giveon, L. Behar, U. Z. Littauer, *Proc. Natl. Acad. Sci. U.S.A.* 79, 4892 (1982).
17. S. M. Mount, *Nucleic Acids Res.* 10, 459 (1982).
18. M. Kozak, *Cell* 44, 283 (1986).
19. D. Yaffe, U. Nudel, Y. Mayer, S. Neuman, *Nucleic Acids Res.* 13, 3723 (1985).
20. U. Gubler and B. J. Hoffman, *Gene* 25, 263 (1983).
21. S. A. Lewis, A. Villasante, P. Sherline, N. Cowan, *J. Cell Biol.* 102, 2098 (1986).
22. K. E. Mostov, M. Friedlander, G. Blobel, *Nature (London)* 308, 37 (1984).
23. F. Sanger, S. Nicklen, A. R. Coulson, *Proc. Natl. Acad. Sci. U.S.A.* 74, 5463 (1977).
24. Bio-Rad Laboratories Bulletin 1205 (1985), Bio-Rad Chemical Division, 2200 Wright Avenue, Richmond, CA 94804.

25. We thank D. Drubin, R. Fletterick, C. Pabo, D. Martin, A. Himmler, D. Selkoe, and K. Kosik for helpful discussions and P. Seeburg for initial help on the sequencing. Supported by an NIH grant from NIGMS (to M.K.). G.L. received postdoctoral support from the NIH and support from the Center for Neurologic Diseases, Brigham and Women's Hospital, Boston, MA.

6 October 1987; accepted 2 December 1987