Striking homology of the ‘variable’ N-terminal as well as the ‘conserved core’ domains of the mouse and human TATA-factors (TFIID)

Taka-aki Tamura1*, Kohsuke Sumita1,2, Ichiro Fujino1,2, Atsuo Aoyama1+, Masami Horikoshi3, Alexander Hoffmann3, Robert G. Roeder3, Masami Muramatsu4 and Katsuhiko Mikoshiba1,3
1Division of Behavior and Neurobiology, National Institute for Basic Biology, Myodaiji-cho, Okazaki 444, 2Division of Macromolecular Function, Institute for Protein Research, Osaka University, 3-2 Yamada-oka, Suita 565, Japan, 3Laboratory of Biochemistry and Molecular Biology, The Rockefeller University, New York, NY 10021, USA and 4Department of Biochemistry, Faculty of Medicine, The University of Tokyo, Hongo, Bunkyo-ku, Tokyo 113, Japan

Received May 12, 1991; Revised and Accepted July 1, 1991

ABSTRACT

A complementary DNA (cDNA) encoding a mouse TFIID (mlID) was isolated from mouse brain cDNA libraries. The 316 amino acid sequence deduced from cDNA sequences revealed the presence of an amino-terminal region enriched in serine, threonine, and proline (STP-cluster), an uninterrupted stretch of 13 glutamine residues (Q-run), a second STP-cluster, and a conserved carboxy-terminal region. Amino acid sequences of the first STP-cluster and the conserved carboxy-terminal region were identical to those of the human TFIID (hIIID). However, the Q-run was considerably shorter than that in hIIID and sequences in the second STP-cluster diverged from those of the hIIID. The murine TFIID transcript is expressed as a 2 kilobase poly(A)+ RNA in the mouse brain. Southern blot analysis identified a single gene copy per haploid mouse genome.

INTRODUCTION

The general transcription factors for RNA polymerase II identified in human cells have been designated TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIG (1, and references therein). Human TFIID (hIIID) binds to the TATA box element (2, 3) and initiates the ordered assembly of RNA polymerase II and the other general factors into a functional preinitiation complex (4, 5). The demonstration of a yeast TFIID (yIIID) that was functionally interchangeable with the hIIID (6—8) led to the total purification of this factor (9) and the cloning of corresponding cDNA (10—14). The cloning of TFIID cDNAs from fission yeast (15, 16), plants (17), Drosophila (18, 19), and human (20—22). Sequence comparisons revealed a highly conserved 180 residue C-terminal domain, which earlier mutational studies had shown to be necessary and sufficient for basal level transcription by the yIIID (23). In contrast, these analyses failed to reveal significant sequence homologies in the N-terminal domains.

Given previous indications that hIIID is a target for the action of at least some regulatory proteins (24—27), as well as functional comparisons of natural and recombinant TFIID species, these results led to the speculation that the variable N-terminal regions might be involved in species-specific interactions with regulatory factors (20, 21, 28). However, the ability of several mammalian activators to enhance promoter activity via yeast TFIID in HeLa cell-derived systems (22, 29) may indicate that at least some interactions with regulatory factors might proceed via the conserved core domain and suggest alternate or accessory roles for the unique N-termini of hIIID. To further investigate structure-function and evolutionary relationships in mammalian TFIID species, we have isolated murine TFIID cDNA and demonstrated striking sequence conservations between human and murine TFIID.

MATERIALS AND METHODS

Cloning and Sequencing of the Mouse TFIID cDNA

Two cDNA libraries in lambda-gt11 were made from cerebellum poly(A)+ RNA of ICR mice by priming with oligo(dT) and with a random hexamer (30). Libraries were screened as previously described (31) using a 32P-labeled DNA probe corresponding to nucleotide positions 576—971 (encoding amino acid residues 160—291) in a hIIID cDNA (20). Nucleotide sequences were determined by the dideoxy chain termination method in each direction.

* To whom correspondence should be addressed at Department of Biochemistry, Saitama Medical School, Moroyama-machi, Saitama 350-04, Japan

+ Present address: Biotechnology Research Laboratory, Tosoh Co. Ltd, 2743-1 Hayakawa, Ayasa, Kanagawa, Japan
Southern and Northern Blot Analyses

Southern blot analysis under stringent hybridization conditions was carried out by the standard method with minor modifications (32). Briefly, 12 μg amounts of the mouse DNA were digested with BanHI and HindIII, electrophoresed on a 0.7% agarose gel, blotted onto a nylon membrane, and probed with a random sodium phosphate (pH 6.5), 0.1% SDS and 50% deionized solution contained 10°Cpml probe, 5xDenhardt's solution, 6°C covered the conserved C-terminal region. The hybridization

RESULTS

The TFII D species encoded by the several cDNAs reported thus far show strong sequence conservations over 180 residues in their respective C-termini (conserved C-terminal core). Assuming sequence conservation at both the protein and DNA level, we employed a hII D cDNA-derived probe encoding 132 residues of the conserved core to screen an oligo(dT)- and random hexamer-primed mouse brain cDNA libraries. The first screens (1×10°Cphage each) yielded 2 positive clones from the oligo(dT)-primed library and 6 from the random hexamer-primed library. These clones (including 23S-1 and H3, Figure 1a) contained sequences related to the conserved C-terminal core of hII D, but lacked the expected N-terminal sequences. The mII D sequence from nucleotide positions 369 to 675 (below) and the hII D sequence from nucleotide positions 1 to 183 were then used as probes to screen the random hexamer-primed library, which yielded one positive clone (23S-2b) carrying a part of the conserved C-terminal core and preceding N-terminal sequences. Overlapping clones 23S-2b, 23S-1 and H3 (Figure 1a) were then used to construct a 1652 base pair cDNA which contained one long open reading frame (ORF) begining at position 65 (Figure 1b).

Nucleotide and predicted amino acid sequences of the mII D Shown is the combined sequence identity in the coding regions. Sequence similarities far show strong sequence conservations over 180 residues in their respective C-termini (conserved C-terminal core). Assuming sequence conservation at both the protein and DNA level, we employed a hII D cDNA-derived probe encoding 132 residues of the conserved core to screen an oligo(dT)- and random hexamer-primed mouse brain cDNA libraries. The first screens (1×10°Cphage each) yielded 2 positive clones from the oligo(dT)-primed library and 6 from the random hexamer-primed library. These clones (including 23S-1 and H3, Figure 1a) contained sequences related to the conserved C-terminal core of hII D, but lacked the expected N-terminal sequences. The mII D sequence from nucleotide positions 369 to 675 (below) and the hII D sequence from nucleotide positions 1 to 183 were then used as probes to screen the random hexamer-primed library, which yielded one positive clone (23S-2b) carrying a part of the conserved C-terminal core and preceding N-terminal sequences. Overlapping clones 23S-2b, 23S-1 and H3 (Figure 1a) were then used to construct a 1652 base pair cDNA which contained one long open reading frame (ORF) beginning at position 65 (Figure 1b). The ORF encoded a 316 amino acid polypeptide which was similar in size to the 335 residue hII D (20). The overall amino acid sequence showed 92% identity with the hII D sequence, while the C-terminal 221 amino acid sequence was identical with the hII D sequence from nucleotide positions 1 to 183 were then used as probes to screen the random hexamer-primed library, which yielded one positive clone (23S-2b) carrying a part of the conserved C-terminal core and preceding N-terminal sequences. Overlapping clones 23S-2b, 23S-1 and H3 (Figure 1a) were then used to construct a 1652 base pair cDNA which contained one long open reading frame (ORF) beginning at position 65 (Figure 1b). The ORF encoded a 316 amino acid polypeptide which was similar in size to the 335 residue hII D (20). The overall amino acid sequence showed 92% identity with the hII D sequence, while the C-terminal 221 amino acid sequence was identical with the hII D sequence. The ORF encoded a 316 amino acid polypeptide which was similar in size to the 335 residue hII D (20). The overall amino acid sequence showed 92% identity with the hII D sequence, while the C-terminal 221 amino acid sequence was identical with the hII D sequence.
and 2b). A stretch of 13 uninterrupted glutamine residues (Q-run) was found between residues 58–70. Regions enriched in ser(S), thr(T), and pro(P) were detected on each side of the Q-run and designated STP-1 and STP-2. The overall contents of ser, thr, and pro residues in STP-1 and STP-2 were 35% and 50%, respectively. The region from residues 119–133 contained an imperfect tripeptide repeat, either pro-ser-thr or a derivative motif (20). In comparing these regions with those of hTFID, the mouse Myelin Basic Protein (MBP) probe detects major signals of 3.7 and 15 kilobase pairs as demonstrated in Fig. 4a. Specific activities of the two probes were the same (8×10^6 cpm/μg). The MBP probe detects major signals of 3.7 and 15 kilobase pairs (kb) as demonstrated by Okano et al. (34). Lane M denotes the positions (in kb) of the molecular size markers. In a Southern analysis, mouse DNA was digested with BamHI or HindIII and probed with the conserved region (nucleotide residue 486–1006) of the mTFID cDNA. This analysis (Fig. 4a) revealed single strong signals of 3.9 (BamHI) or 5.5 (HindIII) kilobase pairs, respectively. A control with a mouse myelin basic protein (MBP) probe of similar specific activity also identified single strong bands whose intensities were comparable to those observed with the mTFID gene analyzed here exists as a single copy in the haploid mouse genome. It may be noteworthy that the Southern

Fig. 3. Comparison of nucleotide sequences between mouse and human TFID. a; Harr plot analysis of the mouse and human (20) TFID cDNA sequences. One dot represents an identical 9 nucleotides out of 10. Start points of the coding region (65), Q-run (236), STP-2 (275), conserved C-terminal core (470) and the terminus of the coding region (1013) are indicated by arrowheads. b; Nucleotide sequence comparison of the 3'-flanking region between mouse (upper) and human (lower) TFID DNAs. Sequences were aligned to have a maximum homology between two cDNAs, including deletion and insertions. The conserved C-terminal core and matching nucleotides are indicated by a shaded area and asterisks, respectively.

Fig. 4. Analysis of the TFID DNA in the mouse chromosome and its expression. a; Southern analysis. Mouse DNA digested with BamHI (B) and HindIII (H) was probed with the mTFID (TFID) or the mouse myelin basic protein (MBP) DNA. Specific activities of the two probes were the same (8×10^6 cpm/μg). The MBP probe detects major signals of 3.7 and 15 kilobase pairs as demonstrated by Okano et al. (34). Lane M denotes the positions (in kb) of the molecular size markers. b; Northern blot analysis of the mouse poly(A)^+ RNA. Positions of the 18s (2 kilobases) and 28s (3.5 kilobases) rRNA are shown.
analysis with the mlID probe also revealed a few minor bands at 8.5 and 10 (BamHI) and 1.9 and 4.4 (HindIII) kilobase pairs (Fig. 4), whose possible significance is discussed below.

Expression of the TFIIID sequence in the mouse brain was studied by Northern analysis (Fig. 4b), which revealed a 2 kilobase signal. The size of the mlID RNA was quite similar to that reported for human (HeLa) cells (22). Thus, we estimate that the predominant mlID mRNA contains an additional 200–300 nucleotides (in the upstream non-coding region) beyond that present on the cDNA of Fig. 1.

**DISCUSSION**

A cDNA from mouse brain libraries was isolated by cross-hybridization with a cDNA encoding hIlD. The mouse sequence contained an ORF encoding a 316 amino acid protein with a molecular mass (35 kDa) similar to that of hIlD (37–38 kDa) (Fig. 1). The evolutionarily conserved C-terminal core of hIlD was totally conserved (below) in the mouse cDNA-encoded protein. Since this region is sufficient for TFIIID binding (to the TATA-box) and for function in basal level transcription (22, 23), we conclude that we have obtained a bona fide mouse TFIIID cDNA.

Overall amino acid sequence similarities between the mouse and human TFIIIDs were striking (Fig. 2a). Each protein had two STP-rich domains separated by Q-run in the N-terminus and a long conserved C-terminal region. We suggest that this structure is a common feature of the mammalian TFIIIDs. Most strikingly, the amino acid sequences in the functionally important C-terminal region were identical, indicating that the essential portion of the TFIIID remained unchanged during mammalian evolution. Nucleotide sequence comparisons between the mlID and hIlD cDNAs (Fig. 3) indicated a high (90%) sequence identity in the coding regions and a lower sequence identity (70%) in the non-coding regions. These data imply that the non-coding flanking sequences of the TFIIID varied more rapidly during evolution than did those in the coding region. The amino acid sequence differences observed in two TFIIIDs are potentially interesting, notably the short Q-run (13 residues) in mlID versus the longer (34–38 residues) Q-run in hIlD. Neither yeast nor plant TFIIIDs contain such uninterrupted Q-runs. However, the *Drosophila* TFIIID has two short (6 and 8 residue) Q stretches (18, 19). The intermediate length of the mlID Q-run indicates a possible expansion during animal evolution, although present data suggest that the length might vary even between different human cell lines (20–22). We suggest that sequences in the Q-run may mutate more frequently than other portions.

The hIlD clones isolated so far were derived from cultured HeLa cells while the present mlID clone was obtained from brain cDNA libraries. Therefore, we cannot eliminate the possibility that the differences found in mlID and hIlD may due to the cell or tissue source. Thus far, TFIIID heterogeneity has been documented in plants (17) and humans; microheterogeneity in the Q-run (20–22). If there were tissue-specific variants of TFIIID, the present study might have selected a TFIIID cDNA enriched in the brain. Since several studies (20, 22, 23, 28, 35) have led to speculation of a possible role for the variable N-terminal region of the TFIIID molecules in transcriptional activation by sequence-specific DNA-binding proteins, the N-terminal sequence difference between mlID and hIlD might possibly be related to tissue-specific transcription.

The size and copy number of mlID transcripts appear to be the same as those reported for hIlD (21, 22). As suggested for hIlD (22), the present analysis with a conserved C-terminal probe indicated a single TFIIID gene in the haploid mouse genome. However, the Southern analysis revealed not only one major signal but several minor signals that most likely represent unique DNA sequences. These minor bands may reflect intron-containing (interrupted) genomic DNA fragments which hybridize only weakly with the probe. Alternatively, the minor bands may reflect TFIIID-related sequences in the mouse genome. Consistent with the latter possibility, the variant TFIIID species in plants arise from at least two TFIIID genes with closely related but distinct nucleotide sequences (17). The clarification of this issue will require isolation of the corresponding genomic DNA fragments.

**ACKNOWLEDGEMENTS**

We thank T. Furuichi for providing mouse brain cDNA libraries. We also thank Y. Suzuki for encouraging us to continue the work. M.H. is an Alexandra and Alexander Sinseheimer Scholar. A.H. is an Arnold and Mabel Beckman graduate fellow. This work was supported by Grant-in-Aid for Scientific Research on Priority Areas (T.T. and M.M.) from the Japanese Ministry of Education, Science and Culture, by grants from the National Institute of Health (M.H. and R.G.R.) and by grants from the Pew Trusts to The Rockefeller University.

**REFERENCES**