Highly conserved core domain and unique N terminus with presumptive regulatory motifs in a human TATA factor (TFIID)

Alexander Hoffmann, Eric Sinn, Tohru Yamamoto, Josephine Wang, Ananda Roy, Masami Horikoshi & Robert G. Roeder

Laboratory of Biochemistry and Molecular Biology, The Rockefeller University, New York, New York 10021, USA

THE factor TFIID is one of several general factors that are necessary and sufficient for transcription initiation by mammalian RNA polymerase II[1,2]. Stable interactions with the common TATA element[3,4] lead both to template commitment and to the assembly of the other general factors into a functional preinitiation complex[5,6]. Consistent with its key role in the promoter activation pathway, human TFIID also seems to be a target for some regulatory factors, as evidenced both by physical[3,7,8] and functional[9–12] studies of interactions between these components. The evolutionary conservation of functional properties[13,14] led to the purification and cloning of yeast TFIID[15–19], the identification of presumptive structural motifs[15,19], and direct structure–function studies[20]. Here we report the cloning of a complementary DNA encoding a functional human TFIID. This reveals an evolutionarily conserved core which corresponds precisely to the 180-residue DNA binding/activation domain determined[20] for yeast TFIID, a near absolute conservation of component structural motifs (direct repeats, central basic core/lysine repeat, and sigma homology), providing further support for their functional importance, and a unique N-terminal structure that suggests involvement in species-specific regulatory factor interactions.

Overlapping human TFIID cDNAs (Fig. 1) were cloned following polymerase chain reaction (PCR) amplification with minimally degenerate primers that were chosen from the yeast TFIID core domain previously shown to be necessary and sufficient for DNA binding[20] and to be conserved in *Schizosac-*

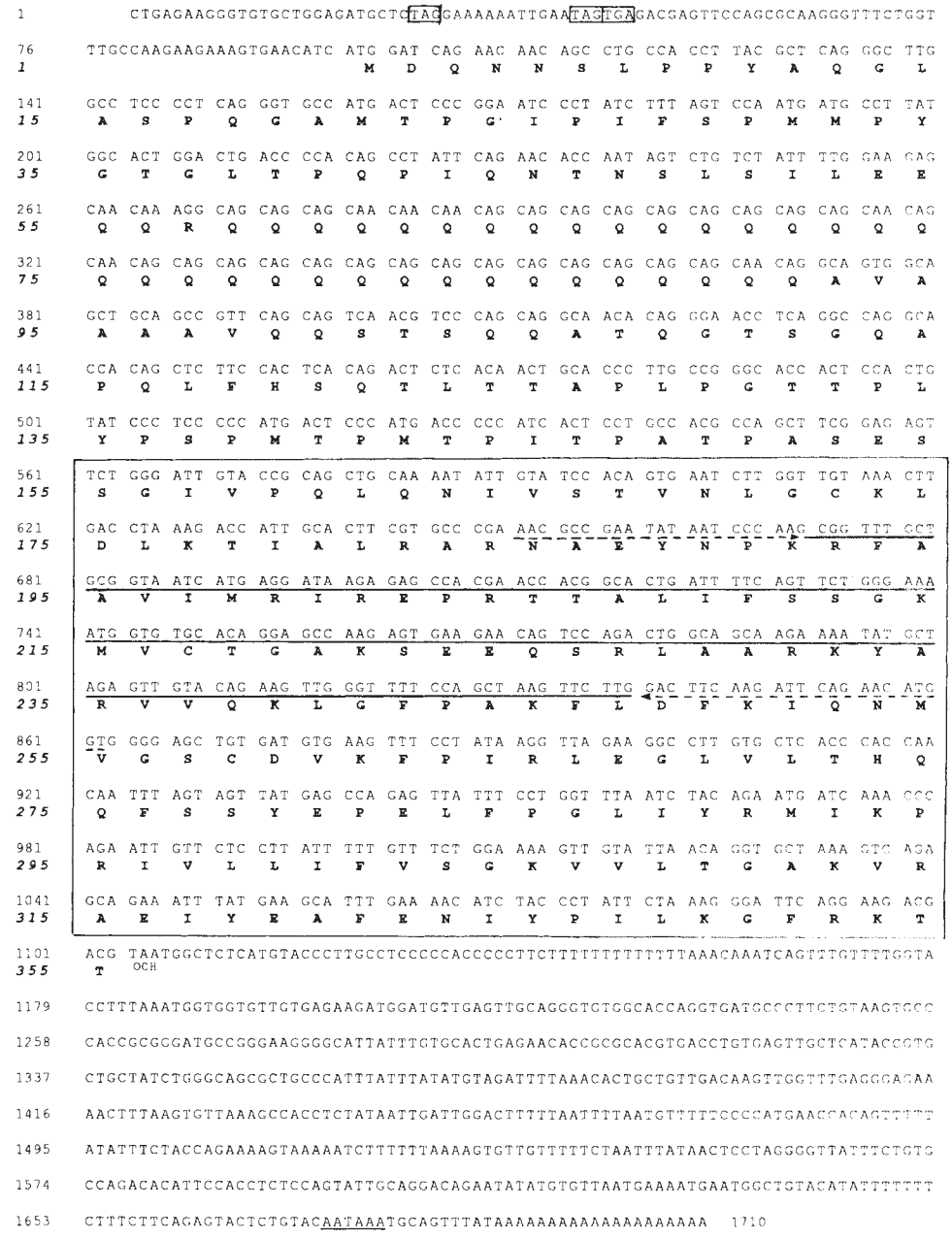

FIG. 1 Nucleotide and predicted amino-acid sequence of human TFIID cDNA. Shown is the combined sequence of three independent overlapping cDNA clones. The PCR product (nucleotide positions 651–862) contained in each of these is underlined with a solid line and sequences encoded by the primers with dashed arrows. The 1,005-base pair (bp) open reading frame is defined by translation start and stop codons (boxed in the nucleotide sequence) and encodes a polypeptide containing the characteristic TFIID homology core domain (amino acids 155–335, boxed) as well as an N-terminal region unique to this human TFIID polypeptide. A putative polyadenylation signal near the 3′ end is underlined.

METHODS. A novel computer program was used to calculate the minimum degeneracy of oligonucleotides necessary to specify five amino-acid stretches within the previously mapped TFIID core domain[21]. These were used as primers to amplify DNA fragments by PCR from a Namalwa cell cDNA library[28]. Five candidate products were selected on the basis of expected size and cross-hybridization to yeast TFIID probes at standard low stringency conditions[21]. Sequencing of the subcloned fragments revealed one that encoded a polypeptide with a high sequence homology to amino acids 91–161 of the *S. cerevisiae* TFIID. This insert was used to screen 10^6 recombinants from the same library under standard high stringency conditions[21]. Three independent clones of 1,258, 1,254 and 1,232 bp were sequenced in both directions using the chain termination method. One contained a poly(A) tail 685 bp 3′ of the TFIID ORF stop codon, whereas another contained the full ORF as evidenced by a start codon (ATG) in the correct reading frame preceded by stop codons (boxed) in the same reading frame.

charomyces pombe TFIID[21]. The combined sequence contains a 335-residue open reading frame encoding a polypeptide of calculated relative molecular mass (M_r) 37,160. Southern blots with short RNA probes showed hybridization with several high M_r placental DNA fragments, indicative either of several homologous genes or of a complex intron–exon structure (as observed for *Schiz. pombe*[21] and *Arabidopsis*[22]), whereas RNase protection assays indicated a low abundance of uniform length transcripts (data not shown).

Several assays showed that the cloned cDNA encodes a functional TFIID with TATA-box-binding and basal-level-transcription activities. A polypeptide (p38) of apparent M_r 38,000 was generated from expression of the cDNA in bacteria (data not shown) or translation of cDNA-derived transcripts in a reticulocyte lysate (Fig. 2*a*). Polypeptide p38 bound specifically to the TATA element of the adenovirus ML promoter as shown by the different sensitivities of two p38-dependent complexes to oligonucleotides with normal or mutant TATA sequences (Fig. 2*b*). One of these complexes was indistinguishable from the complex formed with natural human TFIID, whereas the second had a faster mobility; this suggests that the translated p38 protein may not be structurally equivalent to natural TFIID and/or that p38 may interact with or be modified by factors in the lysate.

When analysed in a cell-free transcription system reconstituted with general factors from HeLa cells, both the lysate-derived (Fig. 2*c*) and the bacterially expressed (Fig. 2*d*) human p38, as well as bacterially expressed yeast TFIID (Fig. 2*d*), could mediate basal level transcription as effectively as natural human TFIID. In contrast, although the stimulatory effect of the human transcriptional activator USF (which binds upstream of TATA on the ML promoter[3]) was readily apparent with native human TFIID, no detectable enhancement was observed with the bacterially expressed yeast or human TFIID polypeptides (Fig. 2*d*). Possibly related to this difference, the native human TFIID showed transcripts in the guanosine(G)-less cassette assay indicative of multiple initiations on the same template (Fig. 2*d* legend), whereas the bacterially expressed proteins did not.

Figure 3*b* details, and Fig. 4 summarizes schematically, a comparison of the human TFIID sequence with TFIID sequences from *Saccharomyces cerevisiae*, *Schiz. pombe*[21], *Arabidopsis thaliana*[22] and *Drosophila melanogaster* (M. Muhich, C. Iida and C. Parker, unpublished observations). There is a remarkable degree of sequence conservation (≥80% identity) and complete colinearity between the ~180-residue C-terminal regions of the factors. Moreover, this region correlates exactly with the region of *S. cerevisiae* TFIID previously shown[20] to be necessary and sufficient for TATA box binding and core promoter transcription. Altogether, these observations clearly argue for a conserved TFIID 'core' which explains the evolutionary conservation of TFIID function.

The TFIID core also contains the structural motifs (Fig. 3*b*, Fig. 4) noted previously[15,19] and in the accompanying paper[21]. The residues comprising these elements show an even more striking (near absolute) conservation from yeast through to humans, which argues strongly for a conservation of functional roles as follows. First, there are direct repeats[19] which allow an element of symmetry in the folded TFIID molecule. This has been proposed to be important for DNA binding, both by analogy to the dimeric structure of other site-specific DNA binding proteins and on the basis of mutational analysis[20] (T.Y.,

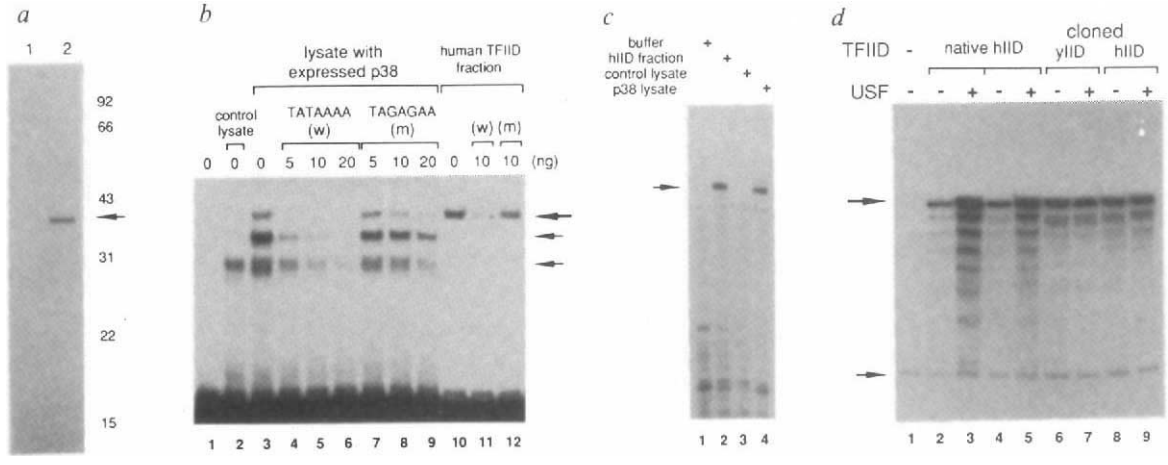

FIG. 2 *In vitro* expression and function of the human cDNA-encoded TFIID. *a*, SDS-PAGE analysis of [^{35}S]methionine-labelled TFIID produced in reticulocyte lysate. Lysates were programmed with no RNA (lane 1) or human TFIID cDNA-derived mRNA (lane 2). The size of the labelled TFIID polypeptide (arrow) based on size markers (not shown) is M_r 38,000. *b*, Site-specific binding of expressed and native human TFIID. Gel shift assays with a major late promoter (MLP) fragment contained: lane 1, no protein; lane 2, control lysate with no RNA; lanes 3–9, lysate programmed with human TFIID mRNA; lanes 10–12, amino-octyl fraction of native human TFIID[4]. Unlabelled oligonucleotides (MLP, positions −45 to −15) contained either a wild type (w) or a mutant (m) TATA box (as indicated) and were added as competitors in the nanogram amounts indicated above each lane. The bold arrow and the upper thin arrow indicate human TFIID specific complexes whereas the lower thin arrow indicates a nonspecific complex formed by a component in the lysate. *c*, Transcriptional activity of expressed human TFIID in a reconstituted system. Reactions containing an MLP template (pSmaF) and partially purified factors TFIIB, TFIIE/F and RNA polymerase II[29] were complemented with buffer (lane 1), DE52 fraction of human TFIID[4] (hIID) (lane 2), unprogrammed lysate (lane 3), and lysate programmed with human TFIID mRNA (lane 4). Specific transcripts measured by primer extension are indicated by the arrow. *d*, USF-stimulated transcription with human TFIID. Reactions containing an adenovirus MLP template with upstream sequences (−400 to +10) attached to the 380-bp G-less cassette pML(C_2AT)[29] and HeLa factors TFIIB, TFIIF/F, and RNA polymerase II were complemented with buffer (lane 1), DE52 fraction of native TFIID (lanes 2, 3), amino-octyl fraction of native TFIID (lanes 4, 5) and purified bacterially expressed yeast (lanes 6, 7) or human (lanes 8, 9) TFIID. A DE52 fraction of huamn activator USF[3] was present in lanes 3, 5, 7 and 9. Transcription from the G-free cassette was in the absence of GTP; the upper arrow indicates the full length cassette transcript resulting from specific initiation at +1 whereas the shorter ~30-bp spaced bands (lanes 2–5) are indicative of multiple initiations on the same template[29]. The lower arrow indicates the position of a labelled DNA fragment added as an internal standard.

METHODS. A human TFIID cDNA (containing nucleotides 1–1,258) in Bluescript SK was transcribed with T7 RNA polymerase and derived RNA was translated in rabbit reticulocyte lysates (Promega) after heat denaturation. Mobility shift assays[15] and primer extension[15] and G-free cassette[29] transcription assays were performed as previously described. *S. cerevisiae* and human TFIID cDNAs containing the coding regions were expressed in bacteria, and lysates were prepared as previously described[21]. The recombinant yeast (p27) and human (p38) TATA factors were purified to near homogeneity.

unpublished results). Second, there is a sigma homology region[15], which overlaps the C-terminal direct repeat and introduces an element of asymmetry into the direct repeat structure. This region of TFIID has been implicated in TATA box binding by deletion[20] and point mutagenesis studies (T.Y., unpublished results) and we have suggested that it could provide the primary determinants for the specificity of DNA binding. Third, there is a repeat of basic residues[15] (notably lysines) in the central basic core connecting the direct repeats that has the potential for formation of an α-helix with basic residues on one face. Mutagenesis studies indicate that individual basic residues are required neither for DNA binding nor for basal transcription (T.Y., unpublished results), consistent with our earlier suggestion[15] that they might interact with certain types of activators. Fourth, there are myc homologies, noted first by Gasch *et al.*[22] for the *Arabidopsis* and yeast proteins, between regions of TFIID and the helix-loop-helix domains of a *myc*-related family of regulatory proteins. As discussed by these authors, this raises the possibility that these regions are important either for homotypic interactions within TFIID or for heterotypic interactions with distinct regulatory factors.

In striking contrast to the conserved C-terminal regions, the N-terminal regions of TFIID differ markedly both in size and sequence (Fig. 4). The human TFIID N terminus lacks the high density of charged residues previously noted[21] in yeast TFIID, but contains obvious structural features that could be functionally related to similar structural motifs in other regulatory factors. These include (see Fig. 3*a*): a region containing 34 consecutive glutamines (the Q-run) flanked by two regions (STP-1 and STP-2) rich in serine (S), threonine (T) and proline (P) residues. The latter show a general similarity to Ser, Thr, (Pro)-rich regions in regulatory proteins like SP1[23] and GAL11[24], whereas the Ser-Pro, Thr-Pro and Tyr-Pro/Pro-Tyr repeats (usually preceded, followed or interspersed with an Ala, Ile, Leu or Met residue) are reminiscent of the reversibly phosphorylated heptad repeat (Tyr-Ser-Pro-Thr-Ser-Pro-Ser) in the largest subunit of RNA polymerase II[25]. Similarly the Q-run and adjacent Gln-rich regions resemble polyglutamine and Gln-rich regions present in other factors[23,24,26] and implicated, in some cases, as activation domains[23,27]. These regions might well serve as interfaces for intermolecular protein–protein interactions.

Yeast TFIID structure-function studies[20] and comparative sequence analyses (Fig. 3) suggest that the human TFIID N terminus is not essential for basal transcription functions. Unlike natural human TFIID[3,7,10-12], however, neither bacterially expressed yeast nor human TFIID seem capable of mediating the normal functions of certain activators (like USF) in reconstituted cell-free systems (Fig. 2*d*; M. Meisterernst, unpublished observations). This suggests that the human N terminus might be involved in mediating regulatory factor interactions, either with TFIID itself or with other general factors, and that these functions might be encoded in separate factors (such as GAL11

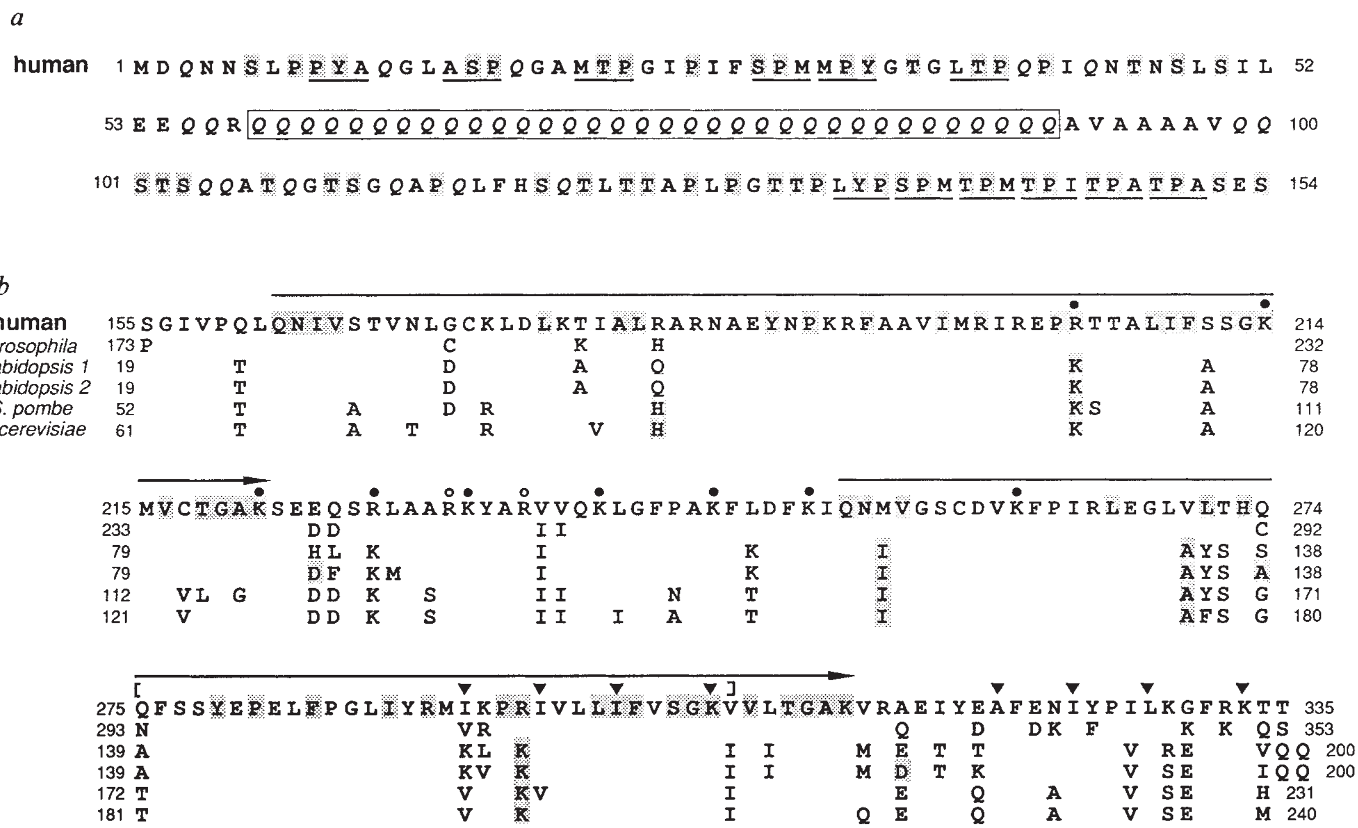

FIG. 3 The human TATA factor consists of a highly conserved TFIID core domain and a unique N-terminal structure. *a*, The amino-acid sequence of the nonconserved N-terminal region of the human TATA factor (single-letter amino-acid code). The potentially interesting primary structure motifs highlighted include a glutamine repeat (boxed) which is flanked by two regions (STP-1 and 2) that are rich in serine, threonine, and proline and characterized by triplets (underlined) containing Ser-Pro, Thr-Pro, and Pro-Tyr/Tyr-Pro pairs preceded or followed by a Met, Ile, Leu, or Ala residue. Alternative arrangements of repeating triplets (such as Pro-Met-Thr, Pro-Ile-Thr) are also possible. *b*, The sequence of the human TFIID core domain is displayed in the top line, whereas the lower lines show only amino acids that differ in the TFIID sequences from other species. Putative structural domains that are discussed in the text are indicated as follows: (1) direct repeats, long arrows indicate regions containing the two interrupted repeats whose constituent residues are shaded; (2) basic repeat, solid circles (lysines) and open circles (arginines) indicate residues in the central basic core which lie on one face of potential α-helical structures; (3) sigma homology, the region with an overall sequence similarity to the sigma 2.3 and 2.4 domains is indicated by brackets while solid inverted triangles denote residues, in two small regions, that correspond to those most highly conserved in the 2.4 regions of various sigma factors. The regions with sequence similarities to the helix-loop-helix proteins (Gasch *et al.*[22]) extend from residues 201 to 220 in direct repeat 1 and from 269 to 314 in direct repeat 2.

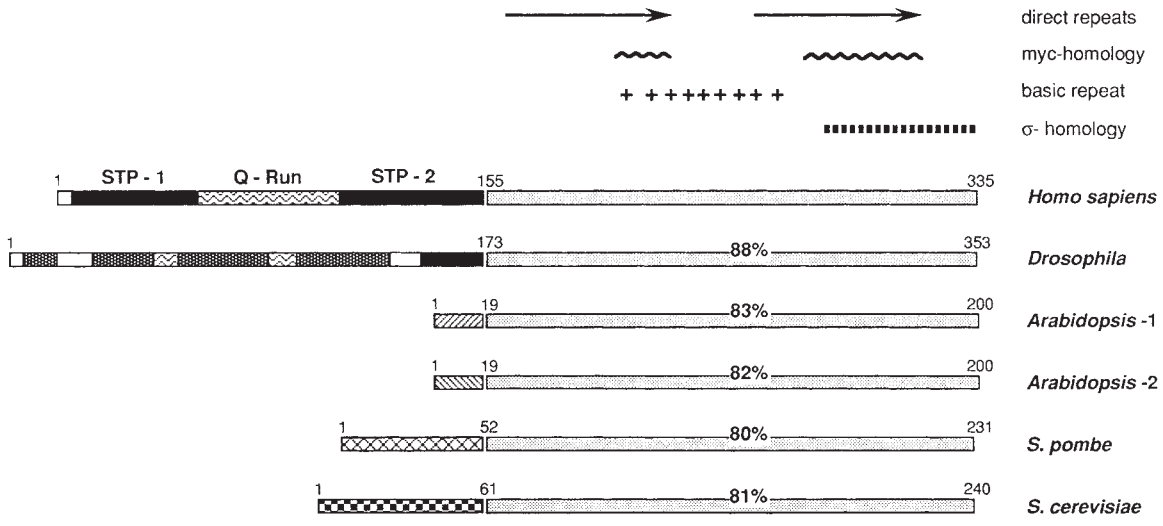

FIG. 4 Evolutionary conservation of the TATA factor primary structure. The degree of sequence identity (relative to human) within the conserved C-terminal core domains of the TFIIDs from human, *Drosophila* (M. Muhich, C. Iida, C. Parker, unpublished results), *Arabidopsis*[22], *Schiz. pombe*[21] and *S. cerevisiae*[15] are summarized. Also indicated are the positions in the core of the structural domains discussed in the text. The N-terminal regions of these proteins differ markedly both in length and amino-acid composition and sequence, except that the *Drosophila* N terminus shows some sequence similarity (including Gln-rich and Ser, Thr, Pro-rich regions) with the human N terminus. For the *Drosophila* N terminus the black and dark grey boxes represent, respectively, regions with strong and weaker sequence similarities to the human STP regions, and short runs of glutamine (Q-runs) are indicated as for the human N terminus.

in lower eukaryotes whereas in mammalian cells they might require either modifications of the N-terminal domains or interactions with other cofactors. The latter situation is also suggested by size differences between natural human TFIID (M_r 120,000)[2] and the cloned TATA-binding protein (M_r 37,160). Although it is tempting to invoke the N-terminal domains in regulatory factor interactions, it is important to bear in mind the great differences in the N-terminal domains from various species and their near absence in *Arabidopsis* TFIID. Thus, in some cases regulatory factor functions may be mediated totally or in part through direct (or cofactor-dependent) interactions with the conserved TFIID core or with other general factors.

In summary, we have described the cloning and functional characterization of a human TATA factor (TFIID). Sequence comparisons with TFIID from other species suggest both structural similarities that account for conserved functions and dissimilarities that may reflect distinct regulatory mechanisms. □

Received 22 May; accepted 21 June 1990.

1. Matsui, T., Segall, J., Weil, P. A. & Roeder, R. G. *J. biol. Chem.* **255,** 11992–11996 (1980).
2. Reinberg, D., Horikoshi, M. & Roeder, R. G. *J. biol. Chem.* **262,** 3322–3330 (1987).
3. Sawadogo, M. & Roeder, R. G. *Cell* **43,** 165–175 (1985).
4. Nakajima, N., Horikoshi, M. & Roeder, R. G. *Molec. cell. Biol.* **8,** 4028–4040 (1988).
5. Van Dyke, M. W., Roeder, R. G. & Sawadogo, M. *Science* **241,** 1335–1338 (1988).
6. Buratowski, S., Hahn, S., Guarente, L. & Sharp, P. A. *Cell* **56,** 549–561 (1989).
7. Horikoshi, M., Hai, T., Lin, Y.-S., Green, M. R. & Roeder, R. G. *Cell* **54,** 1033–1042 (1988).
8. Horikoshi, M., Carey, M. F., Kakidani, H. & Roeder, R. G. *Cell* **54,** 665–669 (1988).
9. Workman, J. L. & Roeder, R. G. *Cell* **51,** 613–622 (1987).
10. Abmayr, S. M., Workman, J. L. & Roeder, R. G. *Genes Dev.* **2,** 542–553 (1988).
11. Workman, J. L., Abmayr, S. M., Cromlish, W. A. & Roeder, R. G. *Cell* **55,** 211–219 (1988).
12. Workman, J. L., Roeder, R. G. & Kingston, R. E. *EMBO J.* **9,** 1299–1308 (1990).
13. Buratowski, S., Hahn, S., Sharp, P. A. & Guarente, L. *Nature* **334,** 37–42 (1988).
14. Cavallini, B. *et al. Nature* **334,** 77–80 (1988).
15. Horikoshi, M. *et al. Nature* **341,** 299–303 (1989).
16. Hahn, S., Buratowski, S., Sharp, P. A. & Guarente, L. *Cell* **58,** 1173–1181 (1989).
17. Eisenmann, D. M., Dollard, C. & Winston, F. *Cell* **58,** 1183–1191 (1989).
18. Schmidt, M. C., Kao, C. C., Pei, R. & Berk, A. J. *Proc. natn. Acad. Sci. U.S.A.* **86,** 7785–7789 (1989).
19. Cavallini, B. *et al. Proc. natn. Acad. Sci. U.S.A.* **86,** 9803–9807 (1989).
20. Horikoshi, M., Yamamoto, T., Ohkuma, Y., Weil, P. A. & Roeder, R. G. *Cell* **61,** 1171–1178 (1990).
21. Hoffmann, A., Horikoshi, M., Wang, C. K., Schroeder, S., Weil, P. A. & Roeder, R. G. *Genes Dev.* **4,** 1141–1148 (1990).
22. Gasch, A., Hoffmann, A., Horikoshi, M., Roeder, R. G. & Chua, N.-H. *Nature* **346,** 390–394 (1990).
23. Courey, A. J. & Tjian, R. *Cell* **55,** 887–898 (1988).
24. Suzuki, Y., Nogi, Y., Abe, A. & Fukasawa, T. *Molec. cell. Biol.* **8,** 4991–4999 (1988).
25. Payne, J. M., Laybourn, P. J. & Dahmus, M. E. *J. biol. Chem.* **264,** 19621–19629 (1989).
26. Mitchell, P. J. & Tjian, R. *Science* **245,** 371–378 (1989).
27. Gerster, T., Balmaceda, C. G. & Roeder, R. G. *EMBO J.* **9,** 1635–1643 (1990).
28. Scheidereit, C. *et al. Nature* **336,** 551–557 (1988).
29. Sawadogo, M. & Roeder, R. G. *Proc. natn. Acad. Sci. U.S.A.* **82,** 4394–4398 (1985).

ACKNOWLEDGEMENTS. We thank M. Muhich, C. Iida and C. Parker for communicating unpublished sequence data and A. Gasch for discussions throughout the project. A.H. is an Arnold and Mabel Beckman graduate fellow, E.S. is a fellow of the American Cancer Society, and M.H. is an Alexandrine and Alexander L. Sinsheimer Scholar. This study was supported by the NIH (R.G.R.) and the Pew Trusts to The Rockefellar University.

Arabidopsis thaliana contains two genes for TFIID

Alexander Gasch*, Alexander Hoffmann†, Masami Horikoshi†, Robert G. Roeder† & Nam-Hai Chua*‡

* Laboratory of Plant Molecular Biology, The Rockefeller University, 1230 York Avenue, New York, New York 10021-6399, USA
† Laboratory of Biochemistry and Molecular Biology, The Rockefeller University, 1230 York Avenue, New York, New York 10021-6399, USA

THE general transcription initiation factor TFIID plays a primary part in the activation of eukaryotic genes transcribed by RNA polymerase II. Binding of TFIID to the TATA box initiates the assembly of other general transcription factors as well as RNA polymerase II at the promoter resulting in a preinitiation complex capable of accurate transcription initiation *in vitro*[1–3]. Human TFIID has been shown to interact with various regulatory factors[4–8]. The observation that stimulation of transcription by different *trans*-acting factors is mediated through distinct TATA elements led to the suggestion that different types of TFIID may exist in yeast[9–11], humans[12–15] and plants[16]. Here we report the cloning and characterization of two distinct TFIID complementary DNA clones from *Arabidopsis thaliana*. Furthermore, we have found that TFIID from *Arabidopsis* and other organisms shows homology to helix-loop-helix proteins.

Several genomic clones from *Arabidopsis thaliana*, all derived from the same locus, were isolated using the recently isolated yeast TFIID gene[17–21] as a probe. To obtain related cDNAs we

‡ To whom correspondence should be addressed.