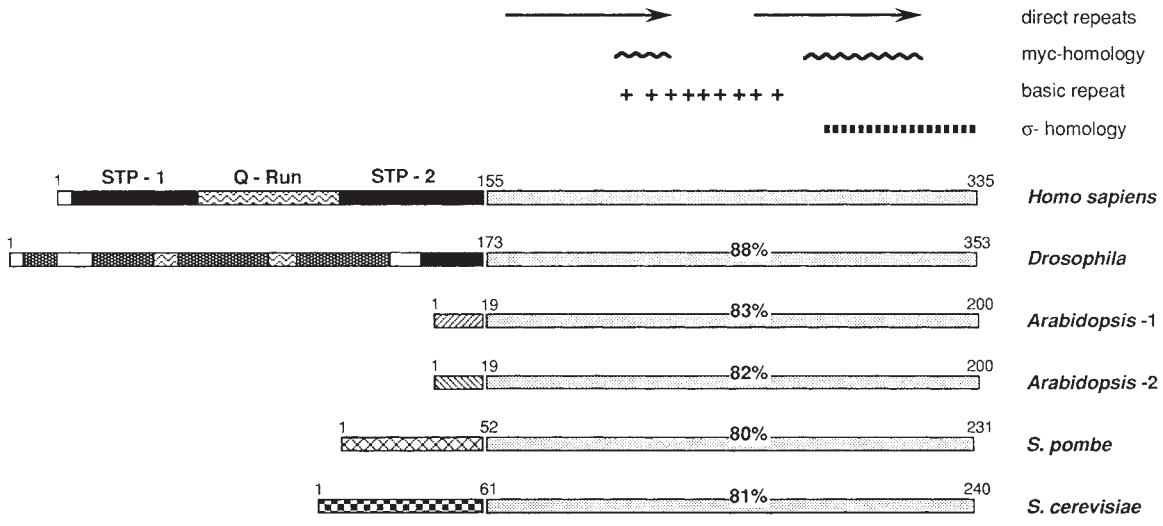

FIG. 4 Evolutionary conservation of the TATA factor primary structure. The degree of sequence identity (relative to human) within the conserved C-terminal core domains of the TFIIDs from human, *Drosophila* (M. Muhich, C. Iida, C. Parker, unpublished results), *Arabidopsis*[22], *Schiz. pombe*[21] and *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.

a

```
    1                                                          CG CAG ATC ACA AAT     14

   15 CTC ATC CCC AGA GAG AGA ACC CAG AGA GCG ATA TTG AAA ATC AAA ACT CTC TCC TTT ATA     74

   75 TAT AAT CCC AAT TTA CAA ATC TCT TTC CCT CTC TAA AAA TTT CTT ATC TTT GTA TAA AAA    134

  135 GCC TTC TCC TTT TTC AAA TCA TTC ACC TTC CTT TCG CCT TTT CGG GGG AAT TTC TTC CAA    194
                                                                                  ▾ (1)
  195 TCT GTT GGG TTT ACG AAA CCC TAG ATT TCG AAA TTC CGG ATT GCT AAC TAG CGA AAG AGA    254

  255 ATG GCT GAT CAA GGA ACG GAA GGG AGC CAG CCA GTT GAC CTT ACT AAA CAT CCT TCA GGG    314
    1  M   A   D   Q   G   T   E   G   S   Q   P   V   D   L   T   K   H   P   S   G      20
                          ▾ (2)
  315 ATT GTT CCT ACT CTT CAA AAT ATT GTC TCA ACA GTG AAC TTA GAC TGC AAG CTT GAT CTT    374
   21  I   V   P   T   L   Q   N   I   V   S   T   V   N   L   D   C   K   L   D   L      40
                                                                  ▾ (3)
  375 AAA GCC ATA GCT TTG CAA GCT AGG AAT GCT GAA TAT AAC CCC AAG CGT TTC GCT GCT GTA    434
   41  K   A   I   A   L   Q   A   R   N   A   E   Y   N   P   K   R   F   A   A   V      60
                                                                                  ▾ (4)
  435 ATC ATG AGG ATC AGA GAG CCA AAG ACC ACA GCG TTA ATT TTT GCT TCT GGG AAA ATG GTG    494
   61  I   M   R   I   R   E   P   K   T   T   A   L   I   F   A   S   G   K   M   V      80
                                                                      ▾ (5)
  495 TGT ACC GGA GCT AAA AGT GAA CAT CTT TCA AAG CTT GCT GCA AGA AAG TAT GCT CGG ATT    554
   81  C   T   G   A   K   S   E   H   L   S   K   L   A   A   R   K   Y   A   R   I     100
                                                  ▾ (6)
  555 GTT CAA AAG CTT GGC TTT CCT GCA AAG TTC AAG GAT TTT AAG ATA CAG AAC ATT GTA GGC    614
  101  V   Q   K   L   G   F   P   A   K   F   K   D   F   K   I   Q   N   I   V   G     120

  615 TCA TGT GAT GTC AAA TTC CCC ATT AGG CTT GAA GGT CTT GCA TAC TCT CAT AGT GCT TTC    674
  121  S   C   D   V   K   F   P   I   R   L   E   G   L   A   Y   S   H   S   A   F     140
              ▾ (7)
  675 TCA AGT TAC GAG CCT GAG CTA TTT CCA GGA TTG ATA TAT AGG ATG AAA CTT CCA AAG ATT    734
  141  S   S   Y   E   P   E   L   F   P   G   L   I   Y   R   M   K   L   P   K   I     160
                                                                      ▾ (8)
  735 GTA CTG CTT ATT TTT GTG TCA GGA AAG ATT GTT ATA ACT GGA GCC AAG ATG AGA GAA GAG    794
  161  V   L   L   I   F   V   S   G   K   I   V   I   T   G   A   K   M   R   E   E     180

  795 ACT TAT ACC GCA TTT GAG AAT ATC TAC CCA GTT CTT AGA GAA TTC AGG AAG GTC CAG CAA    854
  181  T   Y   T   A   F   E   N   I   Y   P   V   L   R   E   F   R   K   V   Q   Q     200
                  ▾ (9)
  855 TAA TAG TAT AGC CTA CCA TAT TGA AGG CAA AAA CCA TGT TCA CGT TTG GAT TGC CAT GAA    914
       *
  915 CTA GAG ATC TCG GGT GTT GGC AGG CGC TTG GTG GGG TTA TAT TCC TTG GAC AAG AAC TTG    974

  975 AAA TGA TTC AGA TTC AAA ACC AGA AAT TGT AAA TAG AGA ATG GAG ATG CCA ATA AAT CTC   1034

 1035 TTT TCT TTT TGT TCT AAA AAT GTC TTC CAA GTA ATC TTG TTT CAT GTT CTG TTT GTT TCA   1094

 1095 GTG AAC AAT ACC CAA TCA TTA AAG AGT TAC TTA GTT CTT TGC TAA CTG TTC TGT AAT AAG   1154

 1155 ACA CCA AGT CTG TTT GCC CTT TAA TAT ACT CTT TCA AGC                               1193
```

b

```
    1                  CA CAG ATC TAA AAA CCA GAA GCG AGA GAA ACA AAT TCA CTC TCC CCT     47

   48 TAT ATA AGC ACC GAT TTA TAA ATC TTT TTC CCT CTT CGA TTC TCA ATT CTT TGT ATA AAA    107

  108 GGC TTC TCC TTT TCT CAA TTC TTC GGC TTC CTT TCG CCC AAA CTC TTC CCT CGA ATC TTT    167

  168 CCT TCT CGT CTT AAA GCT ACG AAA CCC TAG ATT TCG GAT TTC TTC GCT ATC CAA AGA AGA    227

  228 ATG ACT GAT CAA GGA TTG GAA GGG AGT AAT CCA GTT GAT CTT AGC AAG CAT CCT TCA GGG    287
    1  M   T   D   Q   G   L   E   G   S   N   P   V   D   L   S   K   H   P   S   G      20

  288 ATT GTT CCT ACT CTT CAA AAC ATT GTC TCC ACG GTG AAC TTA GAC TGC AAG CTA GAT CTT    347
   21  I   V   P   T   L   Q   N   I   V   S   T   V   N   L   D   C   K   L   D   L      40

  348 AAA GCC ATA GCT TTG CAG GCT CGG AAT GCT GAA TAT AAT CCC AAG CGT TTT GCT GCG GTG    407
   41  K   A   I   A   L   Q   A   R   N   A   E   Y   N   P   K   R   F   A   A   V      60

  408 ATA ATG AGG ATC AGA GAA CCG AAG ACT ACA GCA TTA ATA TTC GCC TCA GGG AAA ATG GTC    467
   61  I   M   R   I   R   E   P   K   T   T   A   L   I   F   A   S   G   K   M   V      80

  468 TGT ACT GGA GCT AAG AGC GAG GAC TTT TCG AAG ATG GCT GCT AGA AAG TAT GCT AGG ATT    527
   81  C   T   G   A   K   S   E   D   F   S   K   M   A   A   R   K   Y   A   R   I     100

  528 GTG CAG AAA TTG GGA TTC CCT GCA AAA TTC AAG GAT TTC AAG ATT CAG AAT ATT GTA GGT    587
  101  V   Q   K   L   G   F   P   A   K   F   K   D   F   K   I   Q   N   I   V   G     120

  588 TCT TGT GAT GTC AAA TTC CCT ATA AGA CTT GAA GGT CTT GCT TAC TCT CAC GCT GCT TTC    647
  121  S   C   D   V   K   F   P   I   R   L   E   G   L   A   Y   S   H   A   A   F     140

  648 TCA AGT TAT GAG CCC GAG CTC TTC CCA GGG CTG ATT TAT AGG ATG AAA GTC CCA AAA ATC    707
  141  S   S   Y   E   P   E   L   F   P   G   L   I   Y   R   M   K   V   P   K   I     160

  708 GTC CTT CTA ATC TTT GTC TCT GGG AAG ATC GTA ATA ACA GGA GCC AAG ATG AGA GAT GAG    767
  161  V   L   L   I   F   V   S   G   K   I   V   I   T   G   A   K   M   R   D   E     180

  768 ACC TAC AAA GCC TTT GAG AAT ATA TAC CCC GTG CTC TCG GAA TTC AGA AAG ATA CAG CAA    827
  181  T   Y   K   A   F   E   N   I   Y   P   V   L   S   E   F   R   K   I   Q   Q     200

  828 TAG GTC ACG GAT TTG TTC CCT GCA AAA CTA GTT GTG CGG TCT TAG CCC CTT GGA GTT GCT    887
       *
  888 AAA GCT TGC TGA GAA TTT TGC CCT TGA ACA AAG GCT TTC ACA GTA GCT AGA CTC TCA CCT    947

  948 TGT GTT TTG GTT CAT AAT ATA ACA TTG TAT ATA CAC AAT GGA GAT TCT AAA GAC AAT TCT   1007

 1008 TCG GTT CTA TTT TTT TTC TTT TTC TTT CCA AGA TAT GTC TTA CAT GTA TGT TGC TGT AGT   1067

 1068 GTC CTC ACA TTT CCA CTT ATG TTA CAA GTA GAA CCT TAA CTC                           1109
```

FIG. 1 Nucleotide sequence of two cDNA classes that encode TFIID in *Arabidopsis thaliana*. *a*, DNA and deduced amino-acid sequence of At-1 cDNA clones. Solid arrowheads indicate the positions of introns in the sequence of the At-1 genomic clones. Translation start and stop codons are underlined. *b*, DNA and deduced amino-acid sequence of At-2 cDNA clones. Translation start and stop codons are underlined.

METHODS. A 680-base pair (bp) *Nde*I–*Hin*dIII fragment of the clone pGemIID8[17,23] containing the majority of the amino-acid coding region of the yeast TFIID gene was labelled by nick translation and used to screen a genomic library of *A. thaliana*. Hybridization was performed at 20% formamide, 5 × SSC, 50 mM Tris-HCl (pH 7.5), 1% SDS, 10 × Denhardt's solution and 50 μg ml^{-1} denatured herring sperm DNA for 16 h at 37 °C. After hybridization, filters were washed in 2 × SSC, 1% SDS, twice for 5 min at room temperature and twice for 30 min at 37 °C. This screen yielded four positive clones which were all derived from the same locus. Poly(A)$^+$ RNA of whole *A. thaliana* grown under continuous white light was used to construct an oligo(dT)-primed cDNA library in λZap (Stratagene) using the cDNA Synthesis System Plus (Amersham). 750,000 plaques of the primary library were screened under the same conditions as described above for the genomic clones, except that a fragment containing the At-1 TFIID gene was used as a probe. In total, 20 positive clones were obtained in this screen. After *in vivo* excision, nine of these clones, as well as subclones of the genomic fragments hybridizing with the yeast TFIID probe, were subjected to double-stranded sequencing using the Sequenase sequencing kit (USB). Sequence data were analysed with the DNASIS and PROSIS programs (Hitachi) on an IBM PS/2 computer.

screened a primary *A. thaliana* cDNA library under low stringency hybridization conditions using one of the genomic clones as a probe. We obtained 20 positive clones (out of 7.5×10^4) of which nine were analysed further by sequencing. Seven of these cDNA clones (*A. thaliana* type 1, At-1) correspond to the isolated *A. thaliana* genomic clones (Fig. 1*a*), whereas the remaining two cDNA clones (*A. thaliana* type 2, At-2) contain a very similar but clearly distinct sequence (Fig. 1*b*). Genomic Southern analysis using the At-1 and At-2 cDNA clones as probes confirms that there are at least two closely related genes present in the *Arabidopsis* genome (data not shown). Northern analysis using gene-specific probes reveals that the At-1 and At-2 messenger RNAs are 1.4 kilobases (kb) and 1.3 kb in size, respectively and that they are present in roughly equal amounts in whole light-grown *Arabidopsis* plants (data not shown). A sequence comparison between the At-1 cDNA and genomic clones reveals the presence of nine introns, the positions of which are indicated in Fig. 1*a*. The longest open reading frame in both types of cDNA clones encodes a protein of 200 amino acids, highly homologous to, but 40 amino acids smaller than, yeast TFIID. The calculated relative molecular mass (M_r) is 22,395 for *Arabidopsis* TFIID-1 and 22,367 for *Arabidopsis* TFIID-2.

To determine whether both cDNA clones encode a functional TFIID we tested TATA box-binding and basal transcription activities of *in vitro* expressed proteins. For At-1 as well as At-2 a protein of M_r 24,000 in size was obtained (Fig. 2*a*), matching the calculated size. Both proteins bind specifically to the adenovirus major late promoter in gel retardation assays (Fig. 2*b*). Both types of TFIID give three shifted complexes (termed 1 and 1′, 2, 3), suggesting that proteins in the reticulocyte lysate either modify or interact with the two proteins. Interestingly, complex 1 formed in the presence of the TFIID-1 is slower in mobility than complex 1′ formed in the presence of the TFIID-2, although both proteins have approximately the same molecular weight. This could indicate an interaction of TFIID-1 with proteins present in the reticulocyte lysate. Alternatively, the complexes of TFIID-1 and TFIID-2 bound to DNA could have different structures as a result of either a difference in the secondary structure of the two proteins or a difference in the way they bind to DNA.

To study further the function of both types of *Arabidopsis* TFIID, we expressed the two TFIID cDNAs in *Escherichia coli* (Fig. 2*c*). Crude extracts of cells expressing these proteins were then tested for TFIID activity in a reconstituted human *in vitro* transcription system containing RNA polymerase II and the basic transcription factors except TFIID. We used a template that contains the adenovirus major late promoter fused to a guanosine-free cassette[22]. Extracts of cells expressing TFIID-1 or TFIID-2 are both able to confer basal transcription in this assay whereas extracts of cells containing the expression vector alone have no activity (Fig. 2*d*). Taken together, these results clearly demonstrate that both cDNA clones encode functional TFIID as both proteins exhibit a specific binding activity to the TATA box and can substitute for the human TFIID to activate basal transcription *in vitro*.

Figure 3 compares the deduced amino-acid sequences of *Arabidopsis* TFIID-1 and TFIID-2. Out of 200 residues, 187 (93.5%) are identical between the two proteins. The N-terminal region (amino acids 1–18) is less conserved (75% identity) than the rest (19–200) of the protein (95% identity). Figure 3 also compares the sequences of both *Arabidopsis* TFIID proteins with that of the yeast TFIID. The N-terminal domain is completely divergent in sequence and smaller in size (18 compared with 60 residues) between *Arabidopsis* and yeast TFIID. Moreover, the N-terminal domain of yeast TFIID is dispensable for basal level transcription[23] suggesting that it is involved in contacting species-specific regulatory factors. The large C-terminal domains, however, are highly conserved between the *Arabidopsis* and yeast TFIID proteins (85% amino acid sequence identity). The high sequence conservation suggests that this domain is involved in the highly conserved TFIID properties, such as TATA box-binding and interaction with the general transcription machinery. Indeed, these properties have been mapped to the C-terminal residues 63–240 of yeast TFIID[23]. This TFIID domain harbours several previously described sequence motifs, all highly conserved between the *Arabidopsis* and yeast TFIID proteins: a domain rich in basic residues[17], a region of similarity to bacterial sigma factors[17] and a recently discovered direct repeat[21]. Possible functions of these sequence motifs, which are schematically shown in Fig. 4, are discussed in the accompanying manuscript[24]. Recently it was shown that

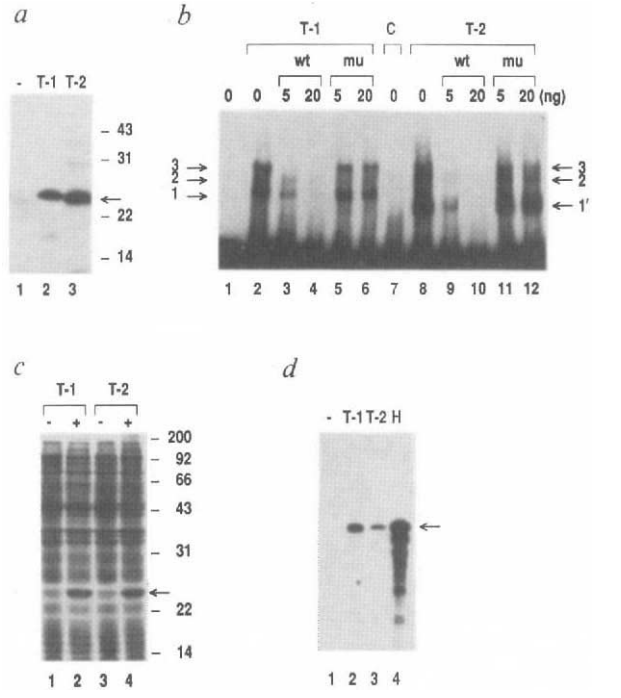

FIG. 2 Functional analysis of *Arabidopsis* TFIID-1 and TFIID-2. *a*, SDS-PAGE of [^{35}S]methionine-labelled TFIID-1 (T-1) and TFIID-2 (T-2) obtained by *in vitro* translation. Samples contained reticulocyte lysate programmed with no (−) RNA (lane 1), with At-1 RNA (lane 2), with At-2 RNA (lane 3). The arrow indicates the TFIID proteins. *b*, Site-specific DNA binding of TFIID-1 (T-1) and TFIID-2 (T-2). An adenovirus major late promoter fragment[17] was used for the gel shift assay[17]. The reactions included: no protein (lane 1), reticulocyte lysate programmed with At-1 RNA (lanes 2–6), with no (C) RNA (lane 7), with At-2 RNA (lanes 8–12). Binding was competed with 5 ng (lanes 3, 9), and 20 ng (lanes 4, 10) of wild-type (wt) TATA box sequence (TATAAA) competitor DNA[17] or with 5 ng (lanes 5, 11), and 20 ng (lanes 6,12) of mutant (mu) TATA box sequence (TAGAGAA) competitor DNA[17]. *c*, SDS-PAGE of TFIID-1 (T-1) and TFIID-2 (T-2) expressed in *E. coli*. The lanes contain *E. coli* extracts carrying pET3a[26] with At-1 cDNA (lanes 1, 2), or At-2 cDNA (lanes 3, 4), either before (−) or after (+) induction with IPTG. The arrow indicates the position of TFIID. *d*, Activation of *in vitro* transcription by TFIID-1 (T-1) and TFIID-2 (T-2) in a reconstituted HeLa *in vitro* transcription system. Added were extracts of induced cells containing pET3A[26] without (−) insert (lane 1), with At-1 cDNA (lane 2) or At-2 cDNA (lane 3). In lane 4 partially purified human (H) TFIID[1] was added. The arrow indicates transcripts of 380 nucleotides.

METHODS. The At-1 and At-2 cDNA clones were transcribed *in vitro* using T3 RNA polymerase and T7 RNA polymerase (Stratagene), respectively. *In vitro* translation, SDS-PAGE analysis and gel retardation assays were performed as described[17]. Expression in *E. coli*, preparation of crude extracts and transcription assays were performed as described[27].

yeast TFIID does not form homodimers[23]. As most DNA-binding proteins bind DNA as dimers, it is tempting to speculate that intra- or intermolecular interaction of the direct repeats present in TFIID is necessary for DNA binding. We therefore directly compared the sequence of these TFIID repeats with sequence motifs known to mediate intermolecular interaction of DNA-binding proteins. Surprisingly, we found a weak similarity of both repeats to the helix-loop-helix domain (Fig. 4). This similarity is mostly confined to the residues that are thought to form the hydrophobic face of the helix 2 region in these proteins[25]. Our suggestion that these two regions in TFIID (residues 70–85 and 161–176) form α-helices can offer an explanation of how the direct repeats within the same TFIID molecule might interact—through the hydrophobic faces of these two putative helices. The sequence homology between TFIID and helix-loop-helix proteins also raises the exciting possibility that these proteins may interact through their amphipathic helix to form heterodimers. Such protein–protein interaction could be one mechanism by which the regulatory effect of helix-loop-helix proteins is transduced.

Our most important result is the finding of two distinct TFIID proteins. Previous reports have shown a heterogeneity among eukaryotic TATA elements: in yeast, different TATA elements are required for constitutive and inducible expression of the *his3* gene[9–11]. Alteration of the TATA element of several mammalian promoters results in dramatic restriction in inducibility, with only a subset of factors being able to activate the promoter[12–15]. Taken together, these results suggested the existence of functionally different TATA factors. To execute its proper function *in vivo*, TFIID has to contact three important components: the TATA-box, the general transcription machinery and regulatory factors. Distinct forms of TFIID are therefore likely to differ in their contact to at least one of these structures. We believe that *Arabidopsis* TFIID-1 and TFIID-2

```
A.t. TFIID-1   1                                                    A   T   Q   T          18
A.t. TFIID-2   1                                                   MTDQGLEGSNPVDLSKHP      18
S.c. TFIID     1  MADEERLKEFKEANKIVFDPNTRQVWENQNRDGTKPATTFQSEEDIKRAAPESEKDTSAT      60

A.t. TFIID-1  19                                                                        78
A.t. TFIID-2  19  SGIVPTLQNIVSTVNLDCKLDLKAIALQARNAEYNPKRFAAVIMRIREPKTTALIFASGK      78
S.c. TFIID    61              A   T G R     TV   H                                  120

A.t. TFIID-1  79            HL   L                                              S     138
A.t. TFIID-2  79  MVCTGAKSEDFSKMAARKYARIVQKLGFPAKFKDFKIQNIVGSCDVKFPIRLEGLAYSHA     138
S.c. TFIID   121    V        D   L S       I   I   A   T                    F   G     180

A.t. TFIID-1 139                        L                  E   T          R     V  * 200
A.t. TFIID-2 139  AFSSYEPELFPGLIYRMKVPKIVLLIFVSGKIVITGAKMRDETYKAFENIYPVLSEFRKIQQ* 200
S.c. TFIID   181  T                     VK                 L     Q E I Q   A          M*   240
```

FIG. 3 Sequence alignment of *A. thaliana* TFIID-1, TFIID-2 and *S. cerevisiae* TFIID. The *S. cerevisiae* (S.c.) TFIID sequence is taken from Horikoshi *et al.*[17]. The amino-acid sequence of the *A. thaliana* (A.t.) TFIID-2 is used as the reference. Only non-conserved amino acids between *S. cerevisiae* TFIID and *A. thaliana* TFIID-2 and between the two types of *A. thaliana* TFIID are shown. The last amino acid residue of each protein is followed by an asterisk.

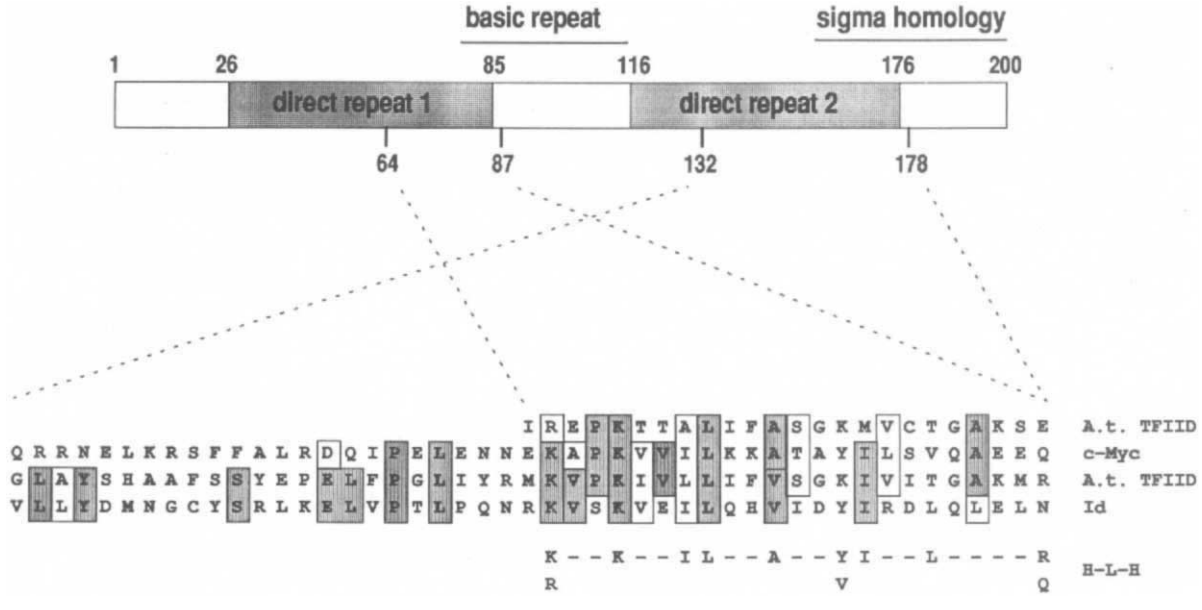

FIG. 4 Regions of similarity between TFIID and helix–loop–helix proteins. Upper panel: A schematic diagram of the structure of *Arabidopsis* TFIID-1 and TFIID-2. The direct repeat 1 encompasses residues 26–85 and the direct repeat 2 encompasses residues 116–176. Lower panel: row 1, amino acid sequence between residues 64–87 of *Arabidopsis* TFIID-1 and TFIID-2; row 2, amino-acid sequence of human c-Myc[28] (residues 357–403); row 3, amino-acid sequence of *Arabidopsis* TFIID-2 (residues 132–178). *Arabidopsis* TFIID-1 has the same sequence with the exception of a serine at position 138 and a leucine at position 157; row 4, amino-acid sequence of Id[29] (residues 85–131); row 5, consensus sequence of the corresponding region in helix–loop–helix (H-L-H) proteins[30]. Identical amino-acid residues in all four sequences are shaded. Conservative amino-acid exchanges are indicated by open boxes.

can be distinguished by their affinity to one or more of these components. The positions of amino-acid exchanges between TFIID-1 and TFIID-2 could provide some clue to where possible functional differences might be encoded. Further experiments are needed to determine exactly to what degree both factors differ in function and to pinpoint the residues that are important in these differences. □

Received 22 May; accepted 21 June 1990.

1. Nakajima, N., Horikoshi, M. & Roeder, R. G. *Molec. cell. Biol.* **8,** 4028-4040 (1988).
2. Van Dyke, M. W., Roeder, R. G. & Sawadogo, M. *Science* **241,** 1335-1338 (1988).
3. Buratowski, S., Hahn, S., Guarente, L. & Sharp, P. A. *Cell* **56,** 549-561 (1989).
4. Sawadogo, M. & Roeder, R. G. *Cell* **43,** 165-175 (1985).
5. Sawadogo, M. *J. biol. Chem.* **263,** 11994-12001 (1988).
6. Horikoshi, M., Carey, M. F., Kakidani, H. & Roeder, R. G. *Cell* **54,** 665-669 (1988).
7. Horikoshi, M., Hai, T., Lin, Y.-S., Green, M. R. & Roeder, R. G. *Cell* **54,** 1033-1042 (1988).
8. Hai, T., Horikoshi, M. Roeder, R. G. & Green, M. R. *Cell* **54,** 1043-1051 (1988).
9. Struhl, K. *Molec. cell. Biol.* **6,** 3487-3853 (1986).
10. Chen, W. & Struhl, K. *Proc. natn. Acad. Sci. U.S.A.* **85,** 2691-2695 (1988).
11. Harbury, P. A. B. & Struhl, K. *Molec. cell. Biol.* **9,** 5298-5304 (1989).
12. Simon, M. C., Fisch, T. M., Benecke, B. J., Nevins, J. R. & Heintz, N. *Cell* **52,** 723-729 (1988).
13. Simon, M. C., Rooney, R. J., Fisch, T. M., Heintz, N. & Nevins, J. R. *Proc. natn. Acad. Sci. U.S.A.* **87,** 513-517 (1990).
14. Taylor, I. C. A. & Kingston, R. E. *Molec. cell. Biol.* **10,** 165-175 (1990).
15. Wefald, C. W., Devlin, B. H. & Williams, S. *Nature* **344,** 260-262 (1990).
16. Kuhlemeier, C., Strittmatter, G., Ward, K. & Chua, N.-H. *Plant Cell* **1,** 471-478 (1989).
17. Horikoshi, M. *et al. Nature* **341,** 299-303 (1989).
18. Hahn, S., Buratowski, S., Sharp, P. A. & Guarente, L. *Cell* **58,** 1173-1181 (1989).
19. Eisenmann, D. M., Dollard, C. & Winston, F. *Cell* **58,** 1183-1191 (1989).
20. Schmidt, M. C., Kao, C. C., Pei, R. & Berk, A. J. *Proc. natn. Acad. Sci. U.S.A.* **86,** 7785-7789 (1989).
21. Cavallini, B. *et al. Proc. natn. Acad. Sci. U.S.A.* **86,** 9803-9807 (1989).
22. Sawadogo, M. & Roeder, R. G. *Proc. natn. Acad. Sci. U.S.A.* **82,** 4394-4398 (1985).
23. Horikoshi, M., Yamamoto, T., Ohkuma, Y., Weil, P. A. & Roeder, R. G. *Cell* **61,** 1171-1178 (1990).
24. Hoffmann, A. *et al. Nature* **346,** 387-390 (1990).
25. Murre, C., McCaw, P. S. & Baltimore, D. *Cell* **56,** 777-783 (1989).
26. Rosenberg, A. H. *et al. Gene* **56,** 125-135 (1987).
27. Hoffmann, A., Horikoshi, M., Wang, C. K., Schroeder, S. & Roeder, R. G. *Genes Dev.* **4,** 1141-1148 (1990).
28. Battey, J. C. *et al. Cell* **34,** 779-787 (1983).
29. Benezra, R., Davis, R. L., Lockshon, D., Turner, D. L. & Weintraub, H. *Cell* **61,** 49-59 (1990).
30. Beckmann, H., Su, L.-K. & Kadesch, T. *Genes Dev.* **4,** 167-179 (1990).

ACKNOWLEDGEMENTS. We thank T. Yamamoto for advice on gel shift assays and *in vitro* transcription, F. William Studier for the T7 expression system, E. Meyerowitz for the *A. thaliana* genomic library and H. Williams for photography. We thank A. van der Krol and P. Benfey for critical reading of the manuscript. A.G. is supported by a postdoctoral fellowship from the Deutsche Forschungsgemeinschaft. A.H. is an Arnold and Mabel Beckmann Graduate Fellow. M.H. is an Alexandrine and Alexander L. Sinsheimer Scholar. This work was also supported by the NIH, Monsanto, and the Pew Trust.

Spontaneous shuffling of domains between introns of phage T4

Mary Bryk*† & Marlene Belfort*‡§

* Wadsworth Center for Laboratories & Research,
New York State Department of Health, Empire State Plaza, PO Box 509,
Albany, New York 12201-0509, USA
† Department of Microbiology & Immunology, Albany Medical College,
47 New Scotland Avenue, Albany, New York 12208, USA
‡ School of Public Health, State University of New York at Albany,
Empire State Plaza, Albany, New York 12201, USA

THE three self-splicing introns in phage T4 (in the *td*, *sun*Y and *nrd*B genes) (Fig. 1*a*) each have the conserved group I catalytic RNA core structure (Fig. 1*b*), out of which is looped an open reading frame[1]. Although the core sequences are very similar (~60% identity), the open reading frames seem to be unrelated. Single crossover recombination events between homologous core sequences in the closely linked *td* and *nrd*B introns have led to 'exon shuffling'[2]. Here we describe spontaneous double crossovers between the unlinked *td* and *sun*Y introns that result in shuffling of an intron structure element, P7.1 (refs 3 and 4). The intron domain-switch variants were isolated as genetic suppressors of a splicing-defective P7.1 deletion in the *td* intron. This unprecedented example of suppression through inter-intron sequence substitution indicates that the introns are in a state of genetic flux and implies the functional interchangeability of the two analogous but non-identical P7.1 elements. The implications of such recombination events are discussed in the light of the evolution of the introns themselves as well as that of their host genomes.

To address questions of intron function and evolution we have isolated Td^+ pseudorevertants from Td^- phage containing stable splicing-defective intron mutations. The ΔP7.1 deletion (Fig. 1*c*), which was introduced into a functional 265-nucleotide (nt) *td* mini-intron[5] (Fig. 1*b*) and shown to inhibit splicing (R. Schroeder, unpublished observations), was crossed into the phage, allowing selection for Td^- plaques[6,7]. Subsequently, Td^+ plaques[7,8], which arose spontaneously from these T4*td*ΔP7.1 phage at a frequency of $\sim 10^{-7}$, were isolated and characterized by plaque hybridization (not shown) and sequence analysis (Fig. 2). Remarkably, four independent Td^+ pseudorevertants had acquired the P7.1 element from the *sun*Y intron (Figs 1*c* and 2). Although the P7.1 elements of *td* and *sun*Y differ (stems of 4 versus 5 base pairs (bp) and loops of 4 versus 10 nt, respec-

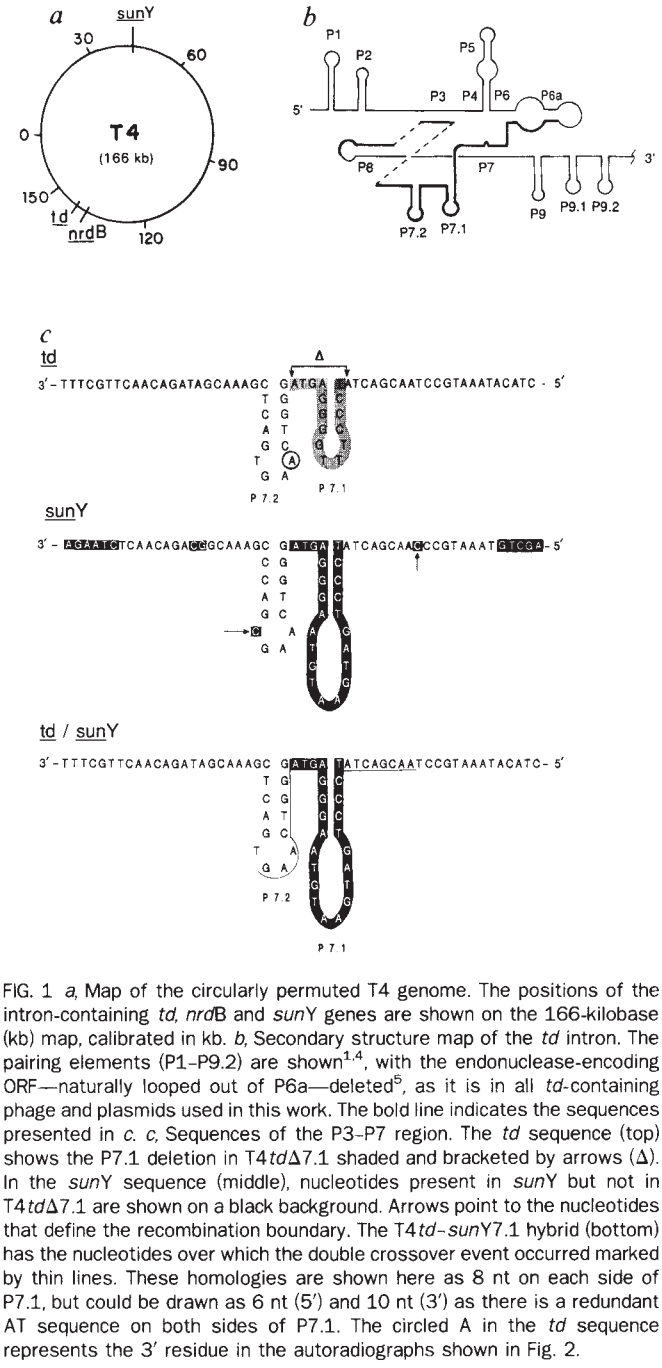

FIG. 1 *a*, Map of the circularly permuted T4 genome. The positions of the intron-containing *td*, *nrd*B and *sun*Y genes are shown on the 166-kilobase (kb) map, calibrated in kb. *b*, Secondary structure map of the *td* intron. The pairing elements (P1–P9.2) are shown[1,4], with the endonuclease-encoding ORF—naturally looped out of P6a—deleted[5], as it is in all *td*-containing phage and plasmids used in this work. The bold line indicates the sequences presented in *c*. *c*, Sequences of the P3–P7 region. The *td* sequence (top) shows the P7.1 deletion in T4*td*Δ7.1 shaded and bracketed by arrows (Δ). In the *sun*Y sequence (middle), nucleotides present in *sun*Y but not in T4*td*Δ7.1 are shown on a black background. Arrows point to the nucleotides that define the recombination boundary. The T4*td*–*sun*Y7.1 hybrid (bottom) has the nucleotides over which the double crossover event occurred marked by thin lines. These homologies are shown here as 8 nt on each side of P7.1, but could be drawn as 6 nt (5′) and 10 nt (3′) as there is a redundant AT sequence on both sides of P7.1. The circled A in the *td* sequence represents the 3′ residue in the autoradiographs shown in Fig. 2.

§ To whom correspondence should be addressed.