JOURNAL OF BACTERIOLOGY, Nov. 1985, p. 957–959
0021-9193/85/110957-03$02.00/0
Copyright © 1985, American Society for Microbiology
Vol. 164, No. 2

IS*186*: an *Escherichia coli* Insertion Element Isolated from a cDNA Library

RASHMIKANT K. KOTHARY, DONALD JONES, AND E. P. M. CANDIDO*
Department of Biochemistry, University of British Columbia, Vancouver, British Columbia, Canada V6T 1W5

Received 7 May 1985/Accepted 20 August 1985

***Escherichia coli* IS*186* was isolated from cDNA libraries made from rainbow trout RNA and maintained in *E. coli* RR1. The element was 1,347 base pairs in length, had a perfect inverted repeat of 25 base pairs, and had an open reading frame of 375 amino acids. The hypothetical protein sequence of IS*186* had limited homology to the *E. coli* IS*4* hypothetical protein I sequence. There were three copies of IS*186* in *E. coli* RR1.**

We have been studying the heat shock-inducible proteins in a cultured cell line from rainbow trout, *Salmo gairdnerii*. The major stress-inducible protein in these cells has a molecular weight of 70,000 (4). This protein has been termed hsp70. As a first step in characterizing this protein, we constructed a cDNA library with RNA isolated from heat-shocked trout RTG-2 cells (5) by the method of Wickens et al. (12). The cDNA was annealed into the ampicillin resistance gene of pBR322 by the guanine-cytosine (G-C) tailing method of Deng and Wu (2). Recombinant plasmids were used to transform *Escherichia coli* RR1 cells. To probe this library, we used ^{32}P-labeled *Drosophila melanogaster* hsp70 DNA (9). One of the clones selected, pTHS70.7 (5), contained partial sequence information for trout hsp70 (designations without lowercase "p" refer to the inserts). The remaining 1.35 kilobases of the total 2.2-kilobase insert contained a sequence with no homology to hsp70. Owing to the unusual nature of this sequence, a second trout cDNA library was screened for similar sequences. The second library was made by the method of Land et al. (6) and inserted into the *Pst*I site of pBR322 by the G-C tailing method (2). Like the first library, it was used to transform *E. coli* RR1 cells by the method of Dagert and Ehrlich (1). The probe used to screen this library was a segment of pTHS70.7 that included 170 base pairs (bp) of the hsp70 coding sequence and 690 bp of the unassigned sequence. A total of 24 clones were selected from a library of 5,000. Partial restriction mapping of the selected clones proved them to contain sequences identical to the unassigned pTHS70.7 sequence. However, none of these clones contained additional sequence information from other RNAs. One of the new clones, pT31, along with pTHS70.7, was analyzed by nucleotide sequencing by both the method of Maxam and Gilbert (7) and that of Messing et al. (8).

The nucleotide sequences of the unassigned portions of THS70.7 and T31 were identical. The complete sequence of T31 is shown in Fig. 1. There was one long open reading frame within T31 (Fig. 1). This 375-amino-acid-long region contained both an ATG start site and a TAA stop codon. No polyadenylation signal (AAUAAA) could be identified downstream from the stop codon.

The T31 sequence had 25-bp-long terminal inverted repeats which lay within the guanine tail of pBR322 (Fig. 2A). We presume that pT31 lost its cDNA insert by recombination. In pTHS70.7 the inverted repeats were 27 bp long and lay in a cytosine tail which linked a heat shock cDNA to the plasmid (Fig. 2B). The variability in the length of the inverted repeats may reflect the insertion of different copies of the element (see below). Direct repeats flanking the sites of insertion in pT31 and pTHS70.7 could not be determined since they would presumably have been indistinguishable from the surrounding homopolymers.

To determine whether the T31 protein or its analogs had been previously reported, the hypothetical protein sequence was sent to the National Biomedical Research Foundation at the Georgetown University Medical Center in Washington, D.C. A search was conducted in their protein sequence data base for homology to the T31 hypothetical protein sequence. Although no extensive homologies were found in the 2,538 sequences searched, the best homology was to the IS*4* hypothetical protein I from *E. coli* (3). In particular, a region of the IS*4* hypothetical protein I which spans amino acid residues 274 to 338 (3) matched well with part of the pT31 hypothetical protein. The included residues were 41% homologous, although the pT31 protein had two insertions of 11 and 18 amino acid residues.

The protein homology and the physical characteristics of T31 suggested to us that this sequence originated in the *E. coli* genome. In addition, RNA dot blots (5) with both heat-induced and control RNAs from trout RTG-2 cells did not reveal hybridization to ^{32}P-labeled probes made from pT31. Southern transfers (11) of trout genomic DNA also did not reveal hybridization to T31 probes.

To test for the presence of T31 sequences in *E. coli*, we selected restriction sites within the sequence and carried out the corresponding digests on DNA isolated by phenol extraction (10) from *E. coli* RR1. After transfer to nitrocellulose (11), the restricted DNA was hybridized to ^{32}P-labeled *Pst*I fragments of T31 (which span the entire element). Hybridization conditions were standard and are described elsewhere (5). The result, shown in Fig. 3A, revealed all of the expected restriction fragments. A similar experiment was done with *E. coli* B DNA (Sigma Chemical Co., St. Louis, Mo.) with identical results (data not shown). Accordingly, the T31 sequence was submitted to the Plasmid Reference Center of the Stanford University School of Medicine, Stanford, Calif., and the designation IS*186* was assigned.

To determine the copy number of IS*186* in the *E. coli* genome, we chose restriction enzymes that did not cut within the IS*186* sequence. The results (data not shown) of *Hin*dIII and *Pvu*I digests of RR1 and B DNAs suggested that three copies of IS*186* exist in the *E. coli* genome. Figure 3B compares *Hin*dIII digests of various *E. coli* strains probed

* Corresponding author.

```
                                                                                 M  N  Y
CTGCAGGGGGGGGGGGGGGGGGGGCCCATAAGCGCTAACTTAAGGGTTGTGGTATTACGCCTGATATGATTTAACGTGCCGATGAATTAC
              15             30             45             60             75             90

 S  H  D  N  W  S  A  I  L  A  H  I  G  K  P  E  E  L  D  T  S  A  R  N  A  G  A  L  T  R
TCTCACGATAACTGGTCAGCAATTCTGGCCCATATTGGTAAGCCCGAAGAACTGGATACTTCGGCACGTAATGCCGGGGCTCTAACCCGC
             105            120            135            150            165            180

 R  R  E  I  R  D  A  A  T  L  L  R  L  G  L  A  Y  G  P  G  G  M  S  L  R  E  V  T  A  W
CGCCGCGAAATTCGTGATGCTGCAACTCTGCTACGTCTGGGGCTGGCTTACGGCCCCGGGGGATGTCATTACGTGAAGTCACTGCATGG
             195            210            225            240            255            270

 A  Q  L  H  D  V  A  T  L  S  D  V  A  L  L  K  R  L  R  N  A  A  D  W  F  G  I  L  A  A
GCTCAGCTCCATGACGTTGCAACATTATCTGACGTGGCTCTCCTGAAGCGGCTGCGGAATGCCGCCGACTGGTTTGGCATACTTGCCGCA
             285            300            315            330            345            360

 Q  T  L  A  V  R  A  A  V  T  G  C  T  S  G  K  R  L  R  L  V  D  G  T  A  I  S  G  P  G
CAAACACTTGCTGTACGCGCCGCAGTTACGGGTTGTACAAGCGGAAAGAGATTGCGTCTTGTCGATGGAACAGCAATCAGTGGCCCCGGG
             375            390            405            420            435            450

 G  G  T  A  E  W  R  L  H  M  G  Y  D  P  H  T  C  Q  F  T  D  F  E  L  T  D  S  R  D  A
GGCGGCACCGCTGAATGGCGACTACATATGGGATATGATCCTCATACCTGTCAGTTCACTGATTTTGAGCTAACCGACAGCAGAGACGCT
             465            480            495            510            525            540

 E  R  L  D  R  F  A  Q  T  A  D  E  I  R  I  A  D  R  G  F  G  S  R  P  E  C  I  R  S  L
GAACGGCTGGACCGATTTGCGCAAACGGCAGACGAGATACGCATTGCTGACCGGGGATTCGGTTCGCGTCCCGAATGTATCCGCTCACTT
             555            570            585            600            615            630

 A  F  G  E  A  D  Y  I  V  R  V  H  W  R  G  L  R  W  L  T  A  E  G  M  R  F  D  M  M  G
GCTTTTGGAGAAGCTGATTATATCGTCCGGGTTCACTGGCGAGGATTGCGCTGGTTAACTGCAGAAGGAATGCGCTTTGACATGATGGGT
             645            660            675            690            705            720

 F  L  R  G  L  D  C  G  K  N  G  E  T  T  V  M  I  G  N  S  G  N  K  K  A  G  A  P  F  P
TTTCTGCGCGGGCTGGATTGCGGTAAGAACGGTGAAACCACTGTAATGATAGGCAATTCAGGTAATAAAAAAGCCGGAGCTCCCTTTCCG
             735            750            765            780            795            810

 A  R  L  I  A  V  S  L  P  P  E  K  A  L  I  S  K  T  R  L  L  S  E  N  R  R  K  G  R  V
GCACGTCTCATTGCCGTATCACTTCCTCCCGAAAAAGCATTAATCAGTAAAACCCGACTGCTCAGCGAGAATCGTCGAAAAGGACGAGTA
             825            840            855            870            885            900

 V  Q  A  E  T  L  E  A  A  G  H  V  L  L  L  T  S  L  P  E  D  E  Y  S  A  E  Q  V  A  D
GTTCAGGCGGAAACGCTGGAAGCAGCGGGCCATGTGCTATTGCTAACATCATTACCGGAAGATGAATATTCAGCAGAGCAAGTGGCTGAT
             915            930            945            960            975            990

 C  Y  R  L  R  W  Q  I  E  L  A  F  K  R  L  K  S  L  L  H  L  D  A  L  R  A  K  E  P  E
TGTTACCGTCTGCGATGGCAAATTGAACTGGCTTTTAAGCGGCTCAAAAGTTTGCTGCACCTGGATGCTTTGCGTGCAAAGGAACCTGAA
            1005           1020           1035           1050           1065           1080

 L  A  K  A  W  I  F  A  N  L  L  A  A  F  L  I  D  D  I  I  S  H  R  W  I  S  P  P  E  V
CTCGCGAAAGCGTGGATATTTGCTAATCTACTCGCCGCATTTTTAATTGACGACATAATCAGCCATCGCTGGATTTCCCCCCCAGAAGTG
            1095           1110           1125           1140           1155           1170

 R  I  R  K  E  E  L  T  R  C  G  E
CGGATCCGAAAAGAAGAACTAACTCGTTGTGGAGAATAACAAAAATGGTCATCTGGAGCTTACAGGTGGCCATTCGTGGGACAGTATCCC
            1185           1200           1215           1230           1245           1260

TGACAGCCTACAAAACGCAATTGAAGAACGCGAGGCATCGTCTTAACGAGGCACCGAGGCGTCGCATTCTTCAGATGGTTCAACCCTTAA
            1275           1290           1305           1320           1335           1350

GTTAGCGCTTATGGGGGGGGGCCCC
     1360      1370
```

FIG. 1. Nucleotide sequence of T31 and its predicted amino acid sequence. The single-letter amino acid code is as follows: A, alanine; C, cysteine, D, aspartic acid; E, glutamic acid; F, phenylalanine; G, glycine; H, histidine; I, isoleucine; K, lysine; L, leucine, M, methionine; N, asparagine; P, proline; Q, glutamine; R, arginine; S, serine; T, threonine; V, valine; W, tryptophan; and Y, tyrosine.

pT31

CCCATAAGCGCTAACTTAAGGGTTG ··· 1291 bp ··· CAACCCTTAAGTTAGCGCTTATGGG

——— GGGGGGGGGGGGGGGGGG|GGGGGGCCCC ———

pTHS70.7

CCCCCATAAGCGCTAACTTAAGGGTTG ··· 1291 bp ··· CAACCCTTAAGTTAGCGCTTATGGGGG

—— GGGGGGGGGGGGGGG-[TROUT HS 70.7 cDNA]-CCCC|CCCCCCCCCCCCCC ———

857 bp

FIG. 2. Inverted repeats and sites of insertion in pT31 and pTHS70.7. Dark horizontal lines indicate the pBR322 sequence.

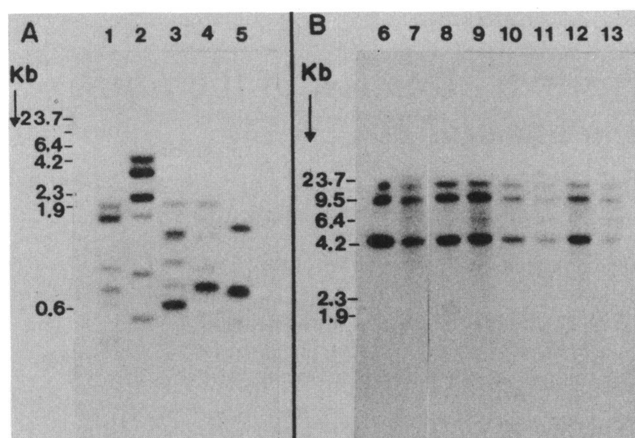

FIG. 3. Southern blot analysis of *E. coli* genomic DNA. DNA from *E. coli* was digested with various restriction enzymes, fractionated on a 0.7% agarose gel, and transferred to nitrocellulose. The probes used were ^{32}P-labeled *Pst*I fragments of T31. (A) *E. coli* RR1 digests were performed to liberate the following internal T31 fragments: lane 1, *Pst*I-*Rsa*I, 316 bp; lane 2, *Pst*I-*Bam*HI, 483 bp; lane 3, *Dde*I, 599 bp; and lane 4, *Bam*HI-*Rsa*I, 800 bp. Lane 5 was a control *Bam*HI-*Rsa*I digest of T31. (B) Various *E. coli* strains digested with *Hin*dIII, which has no site within T31, were probed with ^{32}P-labeled pT31 *Pst*I fragments. The strains were as follows: lane 6, RR1; lane 7, JM83; lane 8, JM103; lane 9, Q358; lane 10, Q359; lane 11, DP50 *supF*; lane 12, MC1000; and lane 13, MC4100. Kb, Kilobases.

with the T31 *Pst*I fragments. All tested strains showed a similar restriction pattern, and all had three copies of the element per genome. Some sequence variability in the three copies may be inferred from the various intensities of hybridization to individual *Hin*dIII fragments (Fig. 3B).

The specificity of IS*186* insertion may favor poly(G-C) stretches in pBR322, although only insertions into the ampicillin resistance gene were analyzed. In addition, since annealed cDNA was used with plasmids to transform *E. coli* cells, it is possible that the insertion of IS*186* may be associated with both G-C stretches and the in vivo ligation of the annealed regions of the plasmid. Further experiments will be required to resolve this point.

In summary, IS*186* appears to be a common element in our cDNA libraries. Its presence there may be related to an insertion specificity for G-C stretches. Its presence further creates two problems. First, it causes an overestimated number of bona fide cDNA clones, possibly by as much as 1% of the total clones. Second, it can be mistaken for cDNA, particularly if it lies tandem to a cDNA clone of interest, such as in the case of pTHS70.7. The ability to make specific probes for the IS*186* sequence should help avoid such problems.

This work was supported by grants from the Medical Research Council of Canada and the British Columbia Health Care Research Foundation.

LITERATURE CITED

1. **Dagert, M., and S. D. Ehrlich.** 1979. Prolonged incubation in calcium chloride improves the competence of Escherichia coli cells. Gene **6**:23–28.
2. **Deng, G., and R. Wu.** 1981. An improved procedure for utilizing terminal transferase to add homopolymers to the 3′ termini of DNA. Nucleic Acids Res. **9**:4173–4188.
3. **Klaer, R., S. Kuhn, E. Tillman, H.-J., Fritz, and P. Starlinger.** 1981. The sequence of IS4. Mol. Gen. Genet. **181**:169–175.
4. **Kothary, R. K., and E. P. M. Candido.** 1982. Induction of a novel set of polypeptides by heat shock or sodium arsenite in cultured cells of rainbow trout, Salmo gairdnerii. Can. J. Biochem. **60**:347–355.
5. **Kothary, R. K., D. Jones, and E. P. M. Candido.** 1984. 70-Kilodalton heat shock polypeptides from rainbow trout: characterization of cDNA sequences. Mol. Cell. Biol. **4**:1785–1791.
6. **Land, H., M. Grez, H. Hauser, W. Lindenmaier, and G. Schütz.** 1981. 5′-Terminal sequence of eukaryotic mRNA can be cloned with high efficiency. Nucleic Acids Res. **9**:2251–2266.
7. **Maxam, A., and W. Gilbert.** 1977. A new method for sequencing DNA. Proc. Natl. Acad. Sci. USA **74**:560–564.
8. **Messing, J., R. Crea, and P. H. Seeburg.** 1981. A system for shotgun DNA sequencing. Nucleic Acids Res. **9**:309–321.
9. **Moran, L. A., M.-E. Mirault, A. Tissiéres, J. Lis, P. Schedl, S. Artavanis-Tsakonas, and W. J. Gehring.** 1979. Physical map of two *Drosphila melanogaster* DNA segments containing sequences coding for the 70,000 dalton heat-shock protein. Cell **17**:1–8.
10. **Schleff, R. F., and P. C. Wensink (ed.).** 1981. Practical methods in molecular biology, p. 98–99. Springer-Verlag, New York.
11. **Southern, E. M.** 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. **98**:503–517.
12. **Wickens, M. P., G. N. Buell, and R. T. Schimke.** 1978. Synthesis of double-stranded DNA complementary to lysozyme, ovomucoid, and ovalbumin mRNAs. J. Biol. Chem. **253**:2483–2495.