Volume 11 Number 12 1983 Nucleic Acids Research

Sequence of the *C. elegans* transposable element Tc1

Bradley Rosenzweig, Louise W.Liao and David Hirsh

Department of Molecular, Cellular and Developmental Biology, University of Colorado, Boulder, CO 80309, USA

Received 4 April 1983; Accepted 6 May 1983

ABSTRACT

The complete nucleotide sequence was determined for Tc1, a transposable element in the nematode Caenorhabditis elegans. The 1610-base-pair element terminates in 54-base-pair perfect inverted repeats and is flanked by a 2-base-pair duplication of the target sequence. The Tc1 sequence contains two long open reading frames on the same DNA strand but in different translational reading frames. The positions of transcriptional control sequences suggest that a single transcript is made, which could produce two polypeptides, 273 and 112 amino acids in length. These features, i.e. terminal repeats, target site duplication and open reading frames, make Tc1 similar to transposable elements from other species.

INTRODUCTION

DNA polymorphisms have frequently been found between the Bristol and Bergerac strains of the nematode Caenorhabditis elegans (1,2). Many of these DNA polymorphisms are due to 1.7-kilobase (kb) insertions in the Bergerac genome at locations where they are absent in the Bristol genome. One of these Bergerac insertions is near a cluster of three actin genes, and this polymorphism facilitated the mapping of these three actin genes to chromosome V (3). Both the 1.7-kb insert adjacent to the actin genes and another 1.7-kb insert located elsewhere in the genome were isolated from Bergerac and characterized as the transposable element Tc1 (4,5). Tc1 is present 25-30 times in the Bristol genome and several hundred times in the Bergerac genome. Several natural isolates of C. elegans have different genomic locations and copy numbers of Tc1, indicating that Tc1 is mobile in evolutionary time. Furthermore, Tc1 apparently excises at a high frequency from some genomic locations in laboratory stocks (4). Similar to transposable elements in other species, Tc1 has short, inverted terminal repeats, as revealed by heteroduplex analysis. The Tc1 family has a homogeneous restriction endonuclease cleavage pattern, implying that all, or nearly all, of the Tc1 elements have the same sequence. The sequence homogeneity and small size of Tc1 make it an

attractive choice for investigating the structure of a eukaryotic transposable element and for determining which portions of the sequence are essential for transposition.

We present here a detailed study of one member of the Tc1 family. EcoRI fragments adjacent to actin gene III from the Bristol and Bergerac strains were subcloned and found to be homologous by restriction mapping and heteroduplex analysis except for the Tc1 element inserted in Bergerac (5). The Bergerac segment containing Tc1 is referred to as the "filled site", the Bristol segment lacking Tc1 is referred to as the "empty site", and the sequence surrounding the point where Tc1 has inserted is referred to as the "target sequence". We have analyzed the sequence of this Bergerac filled site and compared it to the corresponding Bristol empty site sequence.

MATERIALS AND METHODS

Nematode strains

The N2 strain of C. elegans var. Bristol used in these studies was from the Hirsh laboratory strain collection at the University of Colorado. It was obtained from the MRC strain collection in 1972, which originated from a single N2 nematode isolated by Brenner (6). The Bergerac LY strain used in these studies was obtained in 1977 from Jean Brun of the University of Lyon. This strain was originally isolated in France by Nigon (7).

Isolation of recombinant clones

Construction of a C. elegans Bristol N2 recombinant DNA library in the lambda Charon-10 vector has been described elsewhere (3). A C. elegans Bergerac LY recombinant DNA library in lambda Charon-10 was constructed using the same methods. Phage containing the Bristol empty site adjacent to actin gene III were isolated using a Dictyostelium discoideum actin cDNA clone by standard plaque hybridization procedures (3,8). The corresponding fragment containing Tc1 was selected from the Bergerac LY recombinant DNA library for hybridizing with both the Bristol empty site fragment and the Dictyostelium actin cDNA clone. EcoRI subfragments from the recombinant phage with and without Tc1 were subcloned into pBR325, as described previously (5). The Bergerac recombinant plasmid, containing the Tc1-filled site on a 5.2-kb EcoRI fragment, is designated pCe(Be)T1. The Bristol recombinant plasmid, containing the empty site on a 3.5-kb EcoRI fragment, is designated pCe(Br)T1.

DNA sequencing

DNA restriction fragments from pCe(Be)T1 and pCe(Br)T1 were labeled with

^{32}P-nucleotides (New England Nuclear) either at the 5' termini by T4 polynucleotide kinase (New England Nuclear, Bethesda Research Labs) or at the 3' termini by the Klenow fragment of DNA polymerase I (New England Biolabs, Boehringer-Mannheim) (9,10). Isolation of DNA fragments and chemical cleavage reactions were essentially as described by Maxam and Gilbert (9). Samples were subjected to electrophoresis on sequencing gels of 40 or 80 cm (11). Both strands were sequenced independently for the majority of the Tc1 element. Where only one strand was sequenced, it was sequenced at least twice.

Analysis of sequence data

The Delila computer program was used for sequence data compilation and to search sequences for reading frames, transcriptional control signals, RNA splice sequences and inverted or direct repeats (12).

RESULTS

Tc1 sequence

The subcloned Bergerac and Bristol EcoRI fragments have identical restriction maps except for the transposable element Tc1 that is located between the KpnI site and the right-most HincII site (Fig. 1A; ref. 5). This region of the Bergerac recombinant plasmid pCe(Be)T1 was sequenced according to the strategy outlined in Fig. 1B, and the corresponding Bristol segment in the recombinant plasmid pCe(Br)T1 was sequenced as outlined in Fig. 1C. The limits of Tc1 were defined by comparing the sequences of the Bergerac and Bristol fragments.

The Tc1 element is composed of 1610 base pairs (bp) and has a perfect inverted terminal repeat of 54 bp (Fig. 2), confirming previous heteroduplex results (4,5). Except for the inverted terminal repeat, the only patterns of repeated sequences in Tc1 are short (4-8 bp), irregularly spaced T/A oligomers. Although the Tc1 sequence contains duplications of up to 12 bp, they consist mainly of the T/A oligomers.

Open reading frames

A computer search of the Tc1 sequence revealed that the largest open reading frame potentially encodes a polypeptide of 273 amino acids, initiating with an ATG codon at nucleotide position 523 and terminating with an ochre codon (TAA) at position 1342. (Nucleotide position numbers are from the sequence of the plus strand, 5' to 3', shown in Fig. 2.) The hypothetical protein is quite basic; 20% of the residues are lysine, arginine, or histidine. The second largest open reading frame potentially encodes a

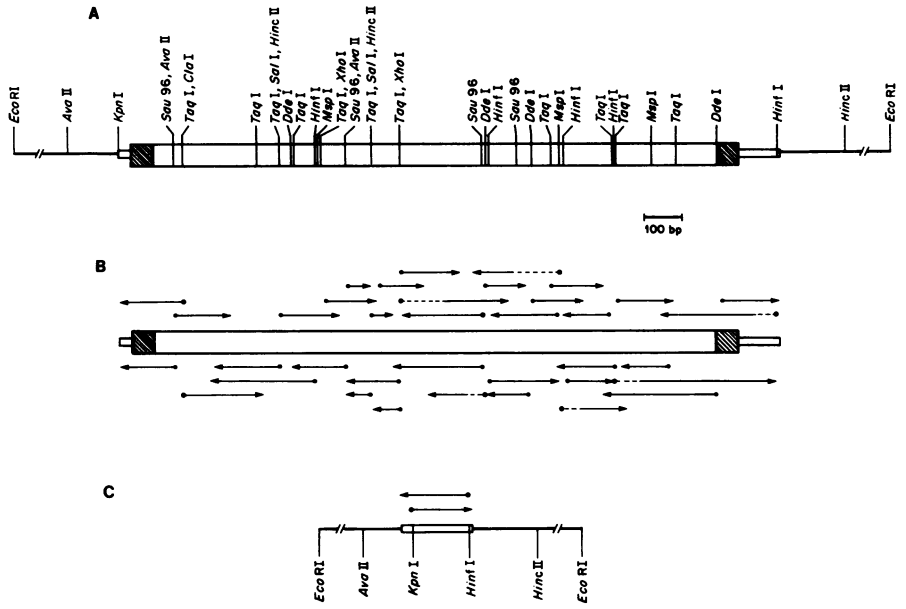

Figure 1. Restriction maps and sequencing strategies of the Bergerac Tc1-filled site and the Bristol empty site. (A) The restriction map delineating the Tc1 insert within the 5.2-kb EcoRI fragment of the Bergerac recombinant plasmid pCe(Be)T1 is shown; however, not every restriction site was used for sequencing. The wide bar represents the Tc1 element, the cross-hatched regions are its inverted terminal repeats, the thin bar is the flanking region for which sequence data were obtained, and the line represents flanking DNA that was not sequenced. (B) Sequence strategy for Tc1. The solid lines with arrows represent the portions of labeled fragments from which sequence data were obtained and the dotted lines represent portions of labeled fragments that were not analyzed. Arrows above the bar represent the 5' strand of Tc1 as drawn; those pointing right depict 5'-end-labeled fragments and arrows pointing left depict 3'-end-labeled fragments. The arrows below the bar are for the 3' strand, with 3'-end-labeled fragments shown as arrows pointing right and 5'-end-labeled fragments shown as arrows pointing left. (C) The restriction map and sequencing strategy of the empty site in the 3.5-kb EcoRI fragment of the Bristol recombinant plasmid pCe(Br)T1. The bar represents the region that was sequenced. Fragments that were sequenced are indicated by arrows, as described above.

112-amino-acid polypeptide on the same strand, but in a different reading frame and nested within the larger polypeptide. Its initiation and termination codons begin at positions 605 and 941, respectively. No other open reading frame in either orientation could encode a polypeptide larger than 75 amino acids.

Possible TATA- and CAAT-box nucleotide sequences associated with transcriptional initiation are found 5' to the putative 273-amino-acid

```
   1  5'CAGTGCTGGCCAAAAAGATATCCACTTTTGGTTTTTTGTGTGTAACTTTTTTCTCAAGCATCCATTTGACTTGAATTTTTCCGTGTGCATAAAGCGAAAT
                    inverted repeat
 101  GTTACGCAAATTTGCGGACCAAACATTACATGATTATCGATTTTTTCTGAATTTTATTTCAATTTTTTGATTTTTTCGTTTTTCCAATTTTCATTATTTT
 201  TTTTGAATTATCAATAAAACGCACTCTGTTTGTTGCACTGGATTTGTTTGGTTGATAAATTATTTTTAAGGTATGGTAAAATCTGTTGGGTGTAAAAATC
 301  TTTCCTTGGACGTCAAGAAAGCCATTGTAGCTGGCTTCGAACAAGGAATACCCACGAAAAGCTCGCGCTGCAAATTCAACGTTCTCCGTCGACTATTTGG
 401  AAAGTAATCAAGAAGTACCAAACTGAGGTGAGTTCGAAAAATATTATTTTTTAATAATAAATGTTTAGAAATCCGTCGCTTTGAGAATCTCGCCCGGCAG
                 CAAT                                          TATA
 501  GCCTCGAGTGACAACCCATAGGATGGATCGCAACATCCTCCGATCAGCAAGAGAAGATCCGCATAGGACCGCCACGGATATTCAAATGATTATAAGTTCT
                 Met
 601  CCAAATGAACCTGTACCAAGTAAACGAACTGTTCGTCGACGTTTACAGCAAGCAGGACTACACGGACGAAAGCCAGTCAAGAAACCGTTCATCAGTAAGA
 701  AAAATCGCATGGCTCGAGTTGCGTGGGCAAAAGCGCATCTTCGTTGGGGACGTCAGGAATGGGCTAAACACATCTGGTCTGACGAAAGCAAGTTCAATTT
 801  GTTCGGGAGTGATGGAAATTCCTGGGTACGTCGTCCTGTTGGCTCTAGGTACTCTCCAAAGTATCAATGCCCAACCGTTAAGCATGGAGGTGGGAGCGTC
 901  ATGGTGTGGGGGTGCTTCACCAGCACTTCCATGGGCCCACTAAGGAGAATCCAAAGCATTATGGATCGTTTTCAATACGAAAACATCTTTGAAACTACAA
1001  TGCGACCCTGGGCACTTCAAAATGTGGGCCGTGGCTTCGTGTTTCAGCAGGATAACGATCCTAAGCATACTTCTCTTCATGTGCGTTCATGGTTTCAACG
1101  TCGTCATGTGCATTTGCTCGATTGGCCAAGTCAGTCTCCGGACTTGAATCCAATAGAGCATTTGTGGGAAGAGTTGGAAAGACGTCTTGGAGGTATTCGG
1201  GCTTCAAATGCAGATGCCAAATTCAACCAGTTGGAAAACGCTTGGAAAGCTATCCCCATGTCAGTTATTCACAAGCTGATCGACTCGATGCCACGTCGTT
1301  GTCAAGCTGTTATTGATGCAAACGGATACGCGACAAAGTATTAAGCATAATTATGTTGTTTTTAAATCCAATTGCTCATATTCCGGTACTTTAATTGTCA
                                                 End
1401  TTTCCTTGCAACCTCGGTTTTTTCAATATTTCTAGTTTTTCGATTTTTTTGAATTTTTCTGAAGTTTTTTCAAAATCTGTTGAACATTTTTGATGAATAT
1501  TGTGTTTTTAGATTTTGTGAACACTGTGGTGAAGTTTCAAAACAAAATAACCACTTAGAAAAAAGTTACACACAAAAAACCAAAAGTGGATATCTTTTTG
                                                 polyA                inverted repeat
1601  GCCAGCACTG3' 1610
```

Figure 2. The 1610-bp sequence of Tc1. Only the sequence of the plus strand (same sense as the putative mRNA) is presented. The 54-bp perfect inverted terminal repeats are underlined. Possible transcriptional control signals consisting of a TATA box, a CAAT box, and a polyA-addition recognition site are indicated. The initiation and termination codons of the hypothetical 273-amino-acid polypeptide are marked. The initiation and termination codons for a potential 112-amino-acid polypeptide begin at positions 605 and 941, respectively (not shown).

polypeptide (Fig. 2). The beginning of the TATA box at position 456 is separated from the initiation codon by 67 bp, which is typical for eukaryotic genes (13,14). The CAAT box begins at position 416, which is 40 bp 5' to the TATA box and is within the spacing (35-57 bp 5' to the TATA box) found for other eukaryotic genes (14,15). No TATA or CAAT sequences with appropriate spacing are found 5' to the putative 112-amino-acid polypeptide (although a TATA sequence begins 14 bp 5' to the initiation codon).

The common eukaryotic polyadenylation signal AATAAA is not found 3' to the coding region (16). However, the sequence AATAA beginning at position 1546 and spaced 201 bp from the termination codon of the 273-amino-acid polypeptide sequence is a possible polyadenylation signal. A computer search failed to find any RNA splice consensus sequence in a position to join open reading frames (17). Several T/A oligomers flank the 273-amino-acid reading frame, making these regions more AT-rich (68% A+T) than the coding region (53% A+T). Similarly, AT-rich regulatory regions flank genes in prokaryotes (18).

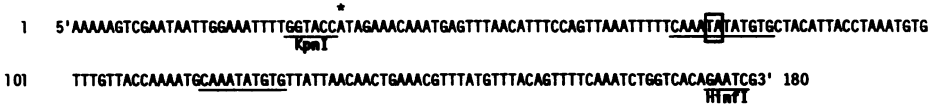

Figure 3. Target sequence. The target sequence from 74 bp 5' to 104 bp 3' to the insertion point (boxed) in the Bristol empty site is shown. The Bergerac target sequence was determined beginning from the KpnI site at the nucleotide marked by the * and continuing to the 3' end. The palindromic 12 nucleotides at the insertion point are underlined, and the 10-nucleotide imperfect repeat starting 40 nucleotides 3' to the insertion point is also underlined. The KpnI and HinfI restriction sites mentioned in the text and shown in Fig. 1 are also marked.

Target sequence

The target sequence flanking Tc1 in the Bergerac strain was analyzed from the KpnI site (44 bp 5' to Tc1) to a HinfI site (105 bp 3' to Tc1) (Fig. 3). The Bristol target sequence was analyzed for this same region and an additional 31 bp 5' to the KpnI site (Fig. 3). The Bristol and Bergerac target sequences are identical for the 149 bp sequenced in both strains, except for a 2-bp duplication in the Bergerac sequence flanking the Tc1 element (Fig. 4). The target sequence is 72% A+T, while C. elegans DNA is 64% A+T overall (19). A snapback structure drawn for the inverted terminal repeat

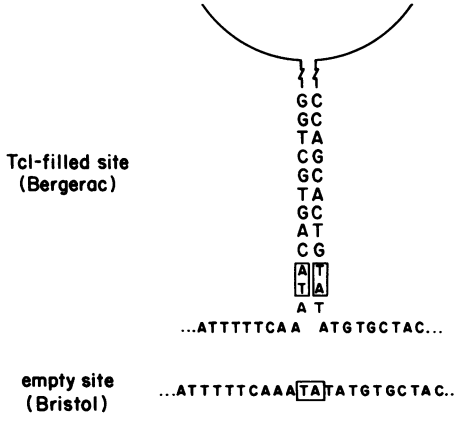

Figure 4. Comparison of Bristol empty site and Bergerac Tc1-filled site. Comparison of the Bristol empty site sequence and the Bergerac Tc1-filled site reveals a two-nucleotide (TA) duplication (boxed) at the ends of Tc1 in the filled site. Its corresponding position in Bristol is also boxed. The filled site is drawn with the inverted repeat of Tc1 forming a snapback structure, which also includes the 3 base pairs of flanking sequence where the TA duplication is located.

of Tc1 can extend 3 bp into the flanking sequence, including the 2-bp duplication (Fig. 4). The extended snapback illustrates that the ends of Tc1 are flanked by an imperfect palindrome (CAAATA:Tc1:TATATG). The 12 bp surrounding the insertion point in the Bristol empty site (without the duplication) are also palindromic for 5 of 6 bp (CAAATA:TATGTG). It is intriguing that 10 of these 12 bp appear again (CAAATATGTG) 40 bp 3' to the insertion point (Fig. 3).

DISCUSSION

Structure of the Tc1 element

The structural organization of Tc1 shares features with other eukaryotic transposable elements. Tc1 elements have inverted terminal repeats, as do P and FB transposable elements in Drosophila (20,21). Tc1 and P elements contain short perfect inverted repeats (54 and 31 bp respectively), while inverted repeats of FB elements are imperfect, are up to 1600 bp long and have extensive reiterations within themselves (22,23). In contrast, Ty elements in yeast and *copia* elements in Drosophila have direct terminal repeats approximately 300 bp long (24,25). The *copia* direct terminal repeat contains a 17-bp imperfect inverted repeat at its ends (25). While P, FB and Ty families exhibit considerable size or sequence heterogeneity, all or nearly all members of the Tc1 family appear identical in size and structure (4,5). In this characteristic, Tc1 resembles *copia*.

Open reading frames

The presence of long open reading frames suggests that Tc1 codes for proteins that function in the element's transposition. The single set of transcriptional control signals in Tc1 could produce a transcript that is processed to synthesize both the 273- and 112-amino-acid polypeptides, one functioning as a repressor and the other as a transposase. Interestingly, the prokaryotic transposon Tn5 produces both a transposase and an inhibitor of transposition from a single transcript (26,27). Abundant transcripts have been found for Ty and *copia* elements; Ty transcripts comprise 5-10% of the total yeast poly(A)$^+$RNA, and *copia* transcripts are 4% of the total RNA in Drosophila cultured cells (28,29). Whereas Tc1, P and FB elements potentially encode polypeptides, no evidence for their transcription has been reported.

Target sequence

Duplication of target sequences upon insertion is common to all transposable elements and is thought to arise from a staggered cleavage at the insertion site during transposition (30). The 2-bp duplication found here is

the smallest reported; the sizes of duplications caused by other transposable elements range from 3 to 12 bp (22,30-33).

Although it is unknown if Tc1 has preferred integration sites, the target sequence studied in this report has features in common with the preferred sites of some prokaryotic transposons. Tn3 inserts into AT-rich regions, similar to this target sequence (34). Tn10 inserts preferentially into imperfect palindromic sequences, and this Tc1 insertion site is an imperfect palindrome (35). The extended snapback structure that includes flanking sequence as well as the inverted terminal repeat might play a role in excision of Tc1. Analysis of other Tc1 insertion sites should determine whether Tc1 integrates at random sequences or at preferred target sites and if preferred sites are randomly distributed in the genome.

ACKNOWLEDGEMENTS

This research was supported by the National Institutes of Health Grant GM26515 to D.H. and a postdoctoral fellowship from the Amerian Cancer Society (PF-1720) to L.W.L.

REFERENCES

1. Emmons, S.W., Klass, M.R. and Hirsh, D. (1979) Proc. Natl. Acad. Sci. U.S.A. 76, 1333-1337.
2. Hirsh, D., Emmons, S.W., Files, J.G. and Klass, M.R. (1979) Eucaryotic gene regulation: ICN-UCLA Symposia on Molecular and Cellular Biology, Axel, R., Maniatis, T. and Fox, C.F. Eds., Vol. 14, pp. 205-218, Academic Press, New York.
3. Files, J.G., Carr, S. and Hirsh, D. (1983) J. Mol. Biol. 164 (in press).
4. Emmons, S.W., Yesner, L., Ruan, K. and Katzenberg, D. (1983) Cell 32, 55-65.
5. Liao, L.W., Rosenzweig, B. and Hirsh, D. (1983) Proc. Natl. Acad. Sci. U.S.A. (in press).
6. Brenner, S. (1974) Genetics 77, 71-94.
7. Nigon, V. (1949) Ann. Sci. Nat. Zool. 11, 1-132.
8. Kindle, K.L. and Firtel, R.A. (1978) Cell 15, 763-778.
9. Maxam, A.M. and Gilbert, W. (1980) Methods Enzymol. 65, 499-560.
10. Klenow, H. and Henningsen, I. (1970) Proc. Natl. Acad. Sci. U.S.A. 65, 168-175.
11. Smith, D.R. and Calvo, J.M. (1980) Nucleic Acids Res. 8, 2255-2274.
12. Schneider, T.D., Stormo, G.D., Haemer, J.S. and Gold, L. (1982) Nucleic Acids Res. 10, 3013-3024.
13. Gannon, F., O'Hare, K., Perrin, F., LePennec, J.P., Benoist, C., Cochet, M., Breathnach, R., Royal, A., Garapin, A., Cami, B. and Chambon, P. (1979) Nature 278, 428-434.
14. Efstratiadis, A., Posakony, J.W., Maniatis, T., Lawn, R.M., O'Connell, C., Spritz, R.A., DeRiel, J.K., Forget, B.G., Weissman, S.M., Slightom, J.L., Blechl, A.E., Smithies, O., Baralle, F.E., Shoulders, C.C. and Proudfoot, N.J. (1980) Cell 21, 653-668.

15. Benoist, C., O'Hare, K., Breathnach, R. and Chambon, P. (1980) Nucleic Acids Res. 8, 127-142.
16. Proudfoot, N.J. and Brownlee, G.G. (1976) Nature 263, 211-214.
17. Breathnach, R., Benoist, C., O'Hare, K., Gannon, F., and Chambon, P. (1978) Proc. Natl. Acad. Sci. U.S.A. 75, 4853-4857.
18. Rosenberg, M. and Court, D. (1979) Annu. Rev. Genet. 13, 319-353.
19. Sulston, J.E. and Brenner, S. (1974) Genetics 77, 95-104.
20. Spradling, A.C. and Rubin, G.M. (1982) Science 218, 341-347.
21. Potter, S., Truett, M., Phillips, M. and Maher, A. (1980) Cell 20, 639-647.
22. Truett, M.A., Jones, R.S. and Potter, S.S. (1981) Cell 24, 753-763.
23. Potter, S.S. (1982) Nature 297, 201-204.
24. Cameron, J.R., Loh, E.Y. and Davis, R.W. (1979) Cell 16, 739-751.
25. Levis, R., Dunsmuir, P. and Rubin, G.M. (1980) Cell 21, 581-588.
26. Johnson, R.C., Yin, J.C.P. and Reznikoff, W.S. (1982) Cell 30, 873-882.
27. Isberg, R.R., Lazaar, A.L. and Syvanen, M. (1982) Cell 30, 883-892.
28. Elder, R.T., St. John, T.P., Stinchcomb, D.T. and Davis, R.W. (1980) Cold Spring Harbor Symp. Quant. Biol. 45, 581-584.
29. Young, M.W. and Schwartz, H.E. (1980) Cold Spring Harbor Symp. Quant. Biol. 45, 629-640.
30. Kleckner, N. (1981) Annu. Rev. Genet. 15, 341-404.
31. Farabaugh, P.J. and Fink, G.R. (1980) Nature 286, 352-356.
32. Gafner, J. and Philippsen, P. (1980) Nature 286, 414-418.
33. Dunsmuir, P., Brorein, W.J. Jr., Simon, M.A. and Rubin, G.M. (1980) Cell 21, 575-579.
34. Tu, C.-P.D. and Cohen, S.N. (1980) Cell 19, 151-160.
35. Halling, S.M. and Kleckner, N. (1982) Cell 28, 155-163.