Volume 11 Number 7 1983 Nucleic Acids Research

A cluster of repetitive elements within a 700 base pair region in the mouse genome

Vernon F.Kalb, Stephan Glasser+, Donna King and Jerry B.Lingrel

Department of Microbiology and Molecular Genetics, University of Cincinnati College of Medicine, Cincinnati, OH 45267, and +Abbott Laboratories, North Chicago, IL 60064, USA

Received 1 February 1983; Accepted 7 March 1983

ABSTRACT

Approximately 39% of the clones from a BALB/c mouse genomic library hybridized with polyadenylated cytoplasmic RNA extracted from anemic mouse spleen. The DNA sequence of a portion of one such clone revealed the presence of three repetitive sequence elements within a 700 bp span. All three elements contain putative RNA polymerase III control regions oriented in the same direction and oligo(dA) tracts at their 3' ends. The first element is a member of the murine B1 family. A comparison of this element with other B1 family members indicates that the B1 family can be divided into two subclasses based on commonly held base changes and deletions. The second element within this 700 bp region may be a member of a new murine Alu family. Its structure is analogous to other murine Alu-equivalent sequences with respect to overall length, the presence of a 3' oligo(dA) tract and putative RNA polymerase III control regions. The third element is a murine type 2 Alu-equivalent sequence.

INTRODUCTION

The mammalian genome is comprised of both single copy and repeated DNA sequences. For the most part the repetitive sequences are 100 to 300 bp long and are interspersed with single copy sequences 1000 to 2000 bp long (1). Within the same organism other types of sequence organization may also occur. For example, in satellite DNA a sequence may be repeated tandemly without the interspersion of single copy sequences (2). Repeated sequences may also be much longer (> 5 kbp) than the more typical 100-300 bp (3). In addition, very long regions of unique DNA can occur without the presence of interspersed repetitive sequences (4).

Primate (5-7) and rodent (8-10) genomes contain a major repetitive DNA family with extensive interspecies homology. The major class of human repetitive DNA has been designated as the "Alu family" because of the presence of an Alu I restriction endonuclease site in many of its members. The human Alu sequence is a dimer formed from two similar direct repeats each approximately 130 bp long. The second monomer contains a 31 bp insertion (7,10-13). The mouse (8) and the Chinese hamster (10) contain repetitive DNA families, each approximately 130 bases long, that show considerable homology to the human Alu sequence (7,10). These are the B1 and type I CHO Alu-equivalent families, respectively.

In this paper we describe the sequence organization of a 700 bp long DNA segment in the BALB/c mouse genome which contains three different elements: (i) a BI repeat, (ii) the murine equivalent of the Chinese hamster type 2 Alu-equivalent sequence and (iii) a sequence which may represent a new repetitive family.

Following the example of Haynes and Jelinek (9), we will refer to the mouse BI sequence as a mouse type I Alu-equivalent sequence and the murine analog of the Chinese hamster type 2 Alu-equivalent sequence as a mouse type 2 Alu-equivalent sequence.

MATERIALS AND METHODS

Enzymes and Labeled Nucleotides. Restriction endonucleases and reverse transcriptase were purchased from New England Biolabs, Bethesda Research Laboratories or Boehringer Mannheim. E. coli DNA polymerase I and the Klenow large fragment were from Boehringer Mannheim. α-^{32}p-dNTPs were purchased from Amersham or New England Nuclear.

Screening the Mouse Genomic Library. A BALB/c mouse genomic library was obtained from Mark Davis (California Institute of Technology). The library was constructed by partially restricting the DNA with Hae III or Alu I and joining the resultant blunt end fragments to EcoRI linkers prior to ligation into the vector λ · Charon 4A. The inserts are 15 to 19 kbp in length. The library was screened as described (14) using a ^{32}P-cDNA probe made to cytoplasmic polyA RNA isolated from anemic mouse spleen (15). After the initial screening, about 30 clones were selected for further plaque purification. The DNA from these clones was isolated, digested with Eco RI, separated on a 0.8% agarose gel, blotted onto nitrocellulose and hybridized to the cDNA probe (16). All 30 clones contained at least one Eco RI fragment that hybridized to the cDNA probe. One of these clones, λ·M58, contained a hybridization-positive 6.7 kb Eco RI fragment. This fragment was nick-translated (17) and used to probe the same set of clones as before. This 6.7 kb Eco RI fragment gave the same pattern of hybridization as was obtained with the cDNA probe. One clone, λ·M3, contained a small (640 bp) Eco RI fragment which hybridized to both probes. It was chosen for sequence analysis and subcloned (18) into pBR325. Growth of E. coli HB101 and the amplification and isolation of plasmid DNA were as described (19).

DNA Sequencing. DNA sequencing protocols, including fragment isolation and labeling, were those of Maxam and Gilbert (20) with modifications as described (21).

The 640 bp Eco RI fragment contains single restriction sites for both Rsa I and Pst I. These sites are located at opposite ends of the fragment. After labeling with ^{32}p-dATP the fragment was cut with either Rsa I or Pst I and the largest fragment from each digest isolated and sequenced. This approach yielded about 130 bp of sequence overlap in the center of the fragment. The 640 bp Eco RI fragment is adjacent to a 1900 bp Eco RI

fragment in λ·M3. This 1900 bp fragment was labeled and digested with Pvu I. The smaller of the two resultant Pvu I fragments is adjacent to the 5' end of the 640 bp Eco RI fragment and was also sequenced.

All work involving recombinant DNA molecules was performed in accordance with National Institute of Health Guidelines for recombinant DNA research.

RESULTS

While screening a BALB/c genomic library with cDNA made using cytoplasmic polyA RNA extracted from anemic mouse spleen, we found that approximately 39% of the genomic clones hybridized to the probe. Plaque purification and restriction analysis of a few of these clones indicated that each positive clone contained at least one Eco RI restriction fragment that hybridized to the probe. A positive restriction fragment from one of the clones cross-hybridized with the positive restriction fragments from other clones indicating that repetitive sequences in our polyA probe were most likely responsible for the observed hybridization. We chose to sequence a small 640 bp fragment from one of these clones (λ·M3) in order to determine the nature of these repetitive elements.

The sequence of the 640 bp Eco RI fragment and part of the 5' flanking 1900 bp Eco RI fragment are presented in Figure 1. Three structural elements can be discerned as indicated by the three shaded segments. The first element, which is flanked by an 18 bp imperfect direct repeat, starts at nucleotide position 38 and ends at nucleotide 203. The second element begins at nucleotide position 234 (where this element begins to share homology with the mouse type 2 Alu-equivalent sequence) and ends at nucleotide 454 after the oligo(dA) rich region. The second element is not flanked by direct repeats. The third element, which is flanked by a 13 bp imperfect direct repeat, starts at nucleotide 506 and ends at nucleotide 700. All three elements have similar overall structures and all three contain oligo(dA) rich regions which start at nucleotide positions 172, 403 and 683, respectively. They also contain sequences, underlined in Fig. 1, that are homologous to the two intragenic RNA polymerase III control regions described by Fowlkes and Shenk (22) and others (23-25) for a variety of genes transcribed either *in vivo* or *in vitro* by RNA pol III. These putative pol III control regions are oriented in the same transcriptional sense in each of our three structural elements. In the 694 bp long region that includes all three elements, about 93% (644 bp) are accounted for by the three elements and their flanking direct repeats. The other 7% is accounted for by 12 bp at the 5' end of the middle element and by 38 bp at the 3' end of the middle element.

The first structural element in Fig. 1 is highly homologous to members of the murine type I Alu-equivalent family. A comparison of this sequence with other members of the same family is shown in Fig. 2. The sequences B1a, B1b and B1c were obtained by Krayev *et al.* (8) from random mouse genomic clones that hybridized to snapback double-stranded

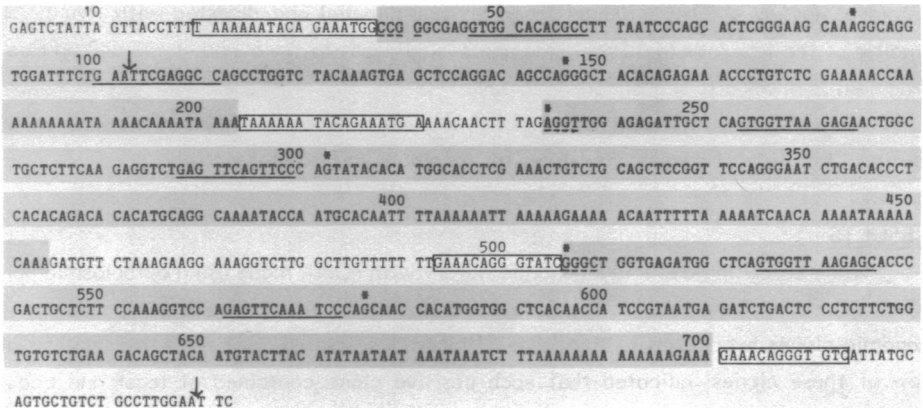

Figure 1
Nucleotide sequence of a portion of the BALB/c mouse genomic clone λ-M3 which contains three closely-linked repetitive elements. The sequence is written in blocks of 10 nucleotides. The three shaded regions contain the three elements. The two pairs of direct repeats are boxed. The first four bases of each element are underlined (dashes). The putative pol III control regions are marked with a solid line. The centers of the two Eco RI restriction sites are marked with vertical arrows. The boundaries of a segment of DNA that is homologous in all three elements (10) are marked by superscript asterisks. Restriction sites for Pst I and Rsa I start at nucleotides 328 and 653, respectively.

hnRNA. The sequence by Coggins et al. (26) is located 2.8 kb to the 3' side of the mouse βmaj globin gene and was identified by its ability to hybridize to a member of the human Alu repetitive family under conditions of low stringency. The sequences described by Monson et al. (27) and Young et al. (28) are located within the intervening sequences of a mouse procollagen gene and the mouse α-fetoprotein gene. They were first detected by hybridization to isotopically-labeled mouse chromosomal DNA.

Within the first 52 bases of the murine type I Alu-equivalent consensus shown in Fig. 2, there occur structures homologous to an RNA pol III control region, a Hogness-Goldberg box and the major T antigen binding site (6). These structures are indicated in Fig. 2 by underlining, superscript dashes and brackets, respectively. The imperfect Hogness-Goldberg box, noted in B1 sequences by Monson et al. (27), occurs immediately after the putative 3' RNA pol III control region in a 14 base string of residues that is invariant in all seven mouse type I Alu-equivalent sequences.

The B1 sequences can be divided into two subclasses by the presence of base changes or deletions at nucleotide positions 7, 85, 105, 117, 121, 122 and 133 of the consensus nucleotide sequence. These positions are indicated in each of the seven sequences by capital letters. The first three repetitive elements contain the same nucleotide or deletion at these positions while the rest share a different nucleotide. It would appear that the two subclasses are descendants of two different progenitor sequences or are

```
                      A  10       ------                          50
Consensus            CCGGGCG-TGG TGGCGCACGC CTTTAATCCC AGCACTC[GG AGGCAGAGGC A]GCGGATTT CTGAGTTCGA GGCCAGCCTG

Bla                  ......A-..a .......t.. .......... .......... .......... ...g...... ........at ..........
Blb                  t.....A-... ...t...... .......... ......t... .......... .......... .......... ...t......
Blc                  ......Ag.a. ...t...t.. .......... ...c...... .......... .......... .......... ..........
1st element/Fig. 1   ......G-a.. ....a..... .......... .......... .a...a.... ...t...... ....a..... ..........
Coggins et al.       ..a...Gt... ....a...t. .......... .......... .......... ....a..... .......... ..........
Monson et al.        ..a..tGg... ...t.a..a. .......... ......aa.. .a........ ...taa...c ........a. .a........
Young et al.         .....tG-... ........a. .......... .......... .......... .......... .......... .......-..

                         T              100    -          G     --
Consensus            GTCTACAGAG TGAGTTCCAG GACAGCCAGG GCTACACAGA GAAACCCTGT CTNN

B1a                  ....T..a.. .......... ....-..... ......G... --........ ..--
B1b                  ....T..... .......... ....-..... ......G... --........ ..--
B1c                  ....T..... .......... ....-..... ...ca.G... --....a... ..--
1st element/Fig. 1   ....A..a.. ....c..... ....G..... ......C... GA........ ..Cg
Coggins et al.       ....A..... ..t....... ....G.t... ......C... GA..a..... ..C-
Monson et al.        ....A..... c....g.... ....G....a ......C... GA........ ..Tg
Young et al.         ....A..a.. ....c.-... ....G-.... ......C... GA........ ..Cc
```

Figure 2
Comparison of mouse type I Alu-equivalent sequences. The consensus sequence is written in blocks of 10 nucleotides. The sequences B1a, B1b and B1c are from Krayev et al. (8) The sequence from Coggins et al. (26) is located in a 7.2 kb Eco RI fragment which contains the β^{maj} globin gene from the BALB/c mouse. The sequence from Monson et al. (27) occurs in the intervening sequence of a mouse proα I(I) procollagen gene. The sequence from Young et al. (28) is found in the first intervening sequence of the mouse α-fetoprotein gene. The consensus sequence here was derived from the sequences listed above. A dash corresponds to a deletion at that position. A period indicates that the nucleotide is the same as the consensus. The letter N designates the absence of a consensus at that position. The putative intragenic control regions for RNA polymerase III are underlined. A 14 base sequence with homology to a T antigen binding site near the origin of papovavirus DNA replication is flanked by brackets. A 6 base sequence which is complementary to an imperfect Hogness-Goldberg box is marked by superscript dashes.

associated with some cellular function responsible for the selection and maintenance of these subclass differences. The existence of different subclasses allows the possibility that these sequences could be associated with different biological roles, though as yet no such role has been identified and the possibility exists that these families represent selfish

```
                      *        10                                         50
Rat type 2 Alu        GGGCTGGAGA GATGGCTCAG TGGTTAAGAG CACCCGACTG CTCTTCCAAA GGTCCTGAGT TCAATTCCCA
3rd element in Fig. 1 .......T.. .......... .......... .......... .......... .....A.... ....A.....
2nd element in Fig. 1 A..T...A.. ...T...... .......... A..-T.G... ......A.G. ....-T.... ...GT.....

                               80                     100
Rat type 2 Alu        GCAACCACAT GGTGGCTCAC AAC--------CATCTGT AAAGAG---ATCC GATGCCCTCT TCTGGTGTGT
3rd element in Fig. 1 .......... .......... ..---------....C.. ..T..---...T ..CT...... ..........
2nd element in Fig. 1 .T.TA..... ..CAC...GA ...TGTCTGCAGC..C.G TTCC..GGA...T ..CA.....A C---------

                               140
Rat type 2 Alu        CTGAAGACAG CTACA-GTGTA CTTA--TATAAT AAACAAATAA ATCTTTAAAAAAAAAAAC...
3rd element in Fig. 1 .......... .....-A.... ....CA...... -..T...... ......AAAAAAAAAAAAAAG...
2nd element in Fig. 1 ------.... AC...CA..C. GGC.AAAT-.CC ..TGC.CA.- --TTTTAAAAAAT...
```

Figure 3
Comparison of the second and third structural elements in Fig. I with a rat type 2 Alu-equivalent sequence. The rat sequence is written in blocks of 10 nucleotides, not counting insertions and deletions. The second and third structural elements pictured above begin at nucleotide position 234 and 506, respectively, in Fig. I. The rat type 2 sequence starts at residue 805 in the rat growth hormone gene sequenced by Page et al. (31). The putative pol III promoter regions are underlined.

DNA (29,30).

The second and third structural elements shown in Fig. 1 are compared in Fig. 3 to a rat type 2 Alu-equivalent sequence. The third structural element is 94% homologous for the first 154 bases to the rat type 2 Alu-equivalent sequence. These two sequences then share a TA_n (n = 1-3) motif for the rest of their lengths before ending with a TCTTT immediately before the oligo(dA) tract. Other murine type 2 Alu-equivalent sequences reported recently (32,33) also show this high homology. An equivalent structure for two Chinese hamster type 2 Alu-equivalent sequences, including a TA_n motif and a TCTT or TCTTTTT immediately preceding the oligo(dA) tract, has also been observed (9).

The second structural element in Fig. 1 is homologous to both the mouse (77%) and rat (81%) type 2 Alu-equivalent sequences for their first 93 residues, but then an 8 bp insertion into the second element marks the beginning of a region of reduced homology as can be seen in Fig. 3. This second element possesses neither the TA_n motif nor the TCTTT sequence immediately before the oligo(dA) region at the 3' terminus. It may be noteworthy that the type 2 Alu-equivalent sequence in Fig. 3 contains an insert with respect to the second structural element (starting at nucleotide 121) with the structure TCTGGTGTGTCTG which shows structural similarities to the sequence GCCTGTGTGTG-GCCTG that Proudfoot and Maniatis have suggested may function as a boundary for recombination events (34).

DISCUSSION

The mammalian Alu and Alu-equivalent families are of interspersed repetitive DNAs typically flanked by direct repeats (11), as are known procaryotic and eukaryotic transposable elements (35,36). However, the structural analogy with transposons is not exact. Alu sequences do not contain the short inverted repeats characteristic of known transposons, and two groups (37,38) have proposed an RNA-mediated mechanism of transposition which requires an RNA transcript of the DNA element to be transposed. This RNA transcript is copied by reverse transcriptase and subsequently integrated into the chromosome at a new location. Many Alu sequences are templates for RNA pol III both *in vitro* (39) and *in vivo* (9,28) and contain a poly dA tract which has been postulated (37) to be important in the reverse transcription of the Alu family pol III transcript. If this insertion process occurs at a staggered break in the chromosome, the integrated sequence would be flanked by direct repeats as is typically observed. Direct repeats would not be observed if the insertion occurs at a blunt break in the chromosome (40) or if the inserted element were the 3' or 5' boundary of a larger element being inserted into a staggered chromosomal break (10).

The second element in Fig. 1 has a strong overall structural resemblance to other Alu-equivalent families as was discussed earlier. Although it shows high homology to the

rat and mouse type 2 Alu-equivalent sequences for the first 93 residues, rat and mouse type 2 Alu-equivalent sequences are much more homologous to each other than to the murine second element. Since this second element contains both RNA pol III control regions and a poly dA tract at the 3' end, the mechanisms described above for transposition via an RNA intermediate might be expected to have generated an interspersed repetitive family with this sequence as a member.

The mechanism by which interspersed repetitive sequence homogeneity is maintained is unclear. Unequal crossing over during meiosis or mitosis (41) which may explain the evolution of tandemly repeated DNA does not seem adequate to explain either the amplification, dispersion or sequence homogeneity of the interspersed repetitive families (13,42). The evolution of interspersed repeats may be more adequately explained by transposition of the repetitive elements to new locations and by gene conversion to maintain sequence homogeneity (42,43).

It has been proposed that Alu family members might function as origins of DNA replication since they contain a 14 base sequence homologous to viral origins of DNA replication and since transcripts from these sequences could serve as primers for DNA synthesis (6). It is interesting to note that the type 2 Alu-equivalent family could provide the RNA primer for DNA replication, but does not possess the 14 base sequence homologous to the viral origin of DNA replication that is found in type I Alu-equivalent sequences.

ACKNOWLEDGMENTS

This work was supported by American Cancer Society Grant NP59L (JBL) and NIH postdoctoral fellowship GM 06705 (VFK).

REFERENCES

1. Davidson, E.H. and Britten, R.J. (1973) Quant. Rev. Biol. 48, 565-613.
2. Corneo, G., Ginelli, E. and Polli, E. (1968) J. Mol. Biol. 33, 331-335.
3. Adams, J.W., Kaufman, R.E., Kretschmer, P.J., Harrison, M. and Nienhuis, A.W. (1980) Nucleic Acids Res. 8, 6113-6128.
4. Wyman, A.R. and White, R. (1980) Proc. Natl. Acad. Sci. U.S.A. 77, 6754-6758.
5. Houck, C.M., Rinehart, F.P. and Schmid, C.W. (1979) J. Mol. Biol. 132, 289-306.
6. Jelinek, W.R., Toomey, T.P., Leinwand, L., Duncan, C.H., Biro, P.A., Choudary, P.V., Weissman, S.M., Rubin, C.M., Houck, C.M., Deininger, P.L., and Schmid, C.W. (1980) Proc. Natl. Acad. Sci. U.S.A. 77, 1398-1402.
7. Schmid, C.W. and Jelinek, W.R. (1982) Science 216, 1065-1070.
8. Krayev, A.S., Kramerov, D.A., Skryabin, K.G., Bayev, A.A. and Georgiev, G.P. (1980) Nucl. Acids Res. 8, 1201-1215.
9. Haynes, S.R. and Jelinek, W.R. (1981) Proc. Natl. Acad. Sci. U.S.A. 78, 6130-6134.
10. Haynes, S.R., Toomey, T.P., Leinwand, L. and Jelinek, W.R. (1981) Mol. Cell. Biol. 1, 573-583.
11. Bell, G.I., Pictet, R. and Rutter, W.J. (1980) Nucleic Acids Res. 8, 4091-4109.
12. Deininger, P.L., Jolly, D.J., Rubin, C.M., Friedmann, T., and Schmid, C.W. (1981) J. Mol. Biol. 151, 17-33.

13. Pan, J., Elder, J.T., Duncan, C.H., and Weissman, S.M. (1981) Nucleic Acids Res. 9, 1151-1170.
14. Robbins, J., Rosteck, P., Jr., Haynes, J.R., Freyer, G., Cleary, M.L., Kalter, H.D., Smith, K., and Lingrel, J.B. (1979) J. Biol. Chem. 254, 6187-6195.
15. Merkel, C.G., Wood, T.G. and Lingrel, J.B. (1976) J. Biol. Chem. 251, 5512-5515.
16. Southern, E.M. (1975) J. Mol. Biol. 98, 503-517.
17. Rigby, P.W.J., Dieckman, M., Rhodes, C. and Berg, P. (1977) J. Mol. Biol. 113, 237-251.
18. Mulligan, R.C., Howard, B.H., and Berg, P. (1979) Nature (London) 277, 108-114.
19. Haynes, J.R., Rosteck, P., Jr., Schon, E.A., Gallagher, P.M., Burks, D.J., Smith, K., and Lingrel, J.B. (1980) J. Biol. Chem. 255, 6355-6367.
20. Maxam, A.M. and Gilbert, W. (1980) In Methods in Enzymology, Grossman, L. and Moldave, K. Eds., Vol. 65, pp. 499-560. Academic Press, New York.
21. Schon, E.A., Wernke, S.M. and Lingrel, J.B. (1982) J. Biol. Chem. 257, 6825-6835.
22. Fowlkes, D.M. and Shenk, T. (1980) Cell 22, 405-413.
23. Ciliberto, G., Traboni, C. and Cortese, R. (1982) Proc. Natl. Acad. Sci. U.S.A. 79, 1921-1925.
24. Fuhrman, S.A., Deininger, P.L., LaPorte, P., Friedmann, T. and Geiduschek, E.P. (1981) Nucleic Acids Res. 9, 6439-6456.
25. Galli, G., Hofstetter, H., and Birnstiel, M.L. (1981) Nature (London) 294, 626-631.
26. Coggins, L.W., Vass, J.K., Stinson, M.A., Lanyon, W.G. and Paul, J. (1982) Gene 17, 113-116.
27. Monson, J.M., Friedman, J. and McCarthy, B.J. (1982) Mol. Cell. Biol. 2, 1362-1371.
28. Young, P.R., Scott, R.W., Hamer, D.H. and Tilghman, S.M. (1982) Nucleic Acids Res. 10, 3099-3116.
29. Dolittle, W.F. and Sapienza, C. (1980) Nature (London) 284, 601-603.
30. Orgel, L.E. and Crick, F.H.C. (1980) Nature (London) 284, 604-607.
31. Page, G.S., Smith, S. and Goodman, H.M. (1981) Nucleic Acids Res. 9, 2087-2104.
32. Krayev, A.S., Markusheva, T.V., Kramerov, D.A., Ryskov, A.P. Skryabin, K.G., Bayev, A.A. and Georgiev, G.P. (1982) Nucleic Acids Res. 10, 7461-7475.
33. Kominami, R. and Muramatsu, M. (1983) Nature (London) 301, 87-89.
34. Proudfoot, N.J. and Maniatis, T. (1980) Cell 21, 537-544.
35. Calos, M.P. and Miller, J.H. (1980) Cell 20, 579-595.
36. Temin, H.M. (1980) Cell 21, 599-600.
37. Jagadeeswaran, P., Forget, B.G. and Weissman, S.M. (1981) Cell 26, 141-142.
38. Van Arsdell, S.W., Denison, R.A., Bernstein, L.B., Weiner, A.M., Manser, T. and Gesteland, R.F. (1981) Cell 26, 11-17.
39. Duncan, C., Choudary, P.A., Elder, J.T., Wang, R.R.C., Forget, B.C., DeRiel, J.K. and Weissman, S.M. (1979) Proc. Natl. Acad. Sci. U.S.A. 76, 5095-5099.
40. Denison, R.A. and Weiner, A.M. (1982) Mol. Cell. Biol. 2, 815-828.
41. Smith, G.P. (1976) Science 191, 528-535.
42. Baltimore, D. (1981) Cell 24, 592-594.
43. Jackson, J.A. and Fink, G.R. (1981) Nature (London) 292, 306-311.