Volume 14 Number 6 1986 Nucleic Acids Research

An Epstein-Barr virus transcription unit is at least 84 kilobases long

Myriam Bodescot, Olivier Brison[1] and Michel Perricaudet

ER272 CNRS, Institut de Recherches Scientifiques sur le Cancer, 94800 Villejuif and
[1]UA1158 CNRS, Institut Gustave Roussy, 94800 Villejuif, France

Received 5 February 1986; Accepted 17 February 1986

ABSTRACT

We have studied the structure of the Epstein-Barr virus mRNAs expressed in B95-8, a productively-infected Marmoset cell line established from *in vitro*-infected B-lymphocytes. We constructed a cDNA library from the cytoplasmic polyadenylated RNAs of B95-8 in the λgt10 bacteriophage. We present here the analysis of a 3.5 kbp cDNA containing exons transcribed from the US, IR and UL regions of the viral genome. The corresponding transcription unit is at least 84 kbp long. Two exons are transcribed from the US region, five from the IR region and two from the UL region. The exons from the IR region consist of two tandem repeats of a unit containing two exons, 66 and 132 nucleotides, and of a third copy of the 66 nucleotide exon. The exons from the UL region contain an open reading frame coding for a 944 amino acid polypeptide. The C-terminal end of this polypeptide harbors three types of repeated sequences. The corresponding mRNA is the second described of a family of mRNAs produced by alternative splicing of exons transcribed from the US, IR and UL regions.

INTRODUCTION

Epstein-Barr virus (EBV) is a herpesvirus ubiquitous in humans (for reviews, see 1 and 2). The virus is associated with nasopharyngeal carcinoma and *in vitro*, some fragments of the viral genome immortalize monkey epithelial cells (3). The virus is also associated with two lymphoproliferative diseases : Burkitt's lymphoma and infectious mononucleosis. *In vitro*, the virus converts human or simian B-lymphocytes into continuously-dividing lymphoblasts. These cells harbour the viral genome in an episomal or in an integrated state and express a set of viral antigens.

The viral genome is nearly 175 kbp long. Two clusters of tandemly-repeated sequences, designated TR and IR, delimit the US and UL regions (Figure 1A). The genome of the B95-8 virus, which

is taken as a prototype, was sequenced (4). Several regions of the viral genome encode mRNAs transcribed in latently or productively-infected cell lines. The structure of some of them was found from S1 mapping and primer extension experiments (5, 6, 7, 8). From the analysis of a first cDNA, we showed that an mRNA expressed in Raji, a latently-infected Burkitt's lymphoma cell line, contains exons transcribed from the IR and UL regions (9). This mRNA belongs to a family of mRNAs produced by alternative splicing of exons transcribed from the US, IR and UL regions. Such a family was detected as well in B95-8, a productively-infected cell line established from *in vitro*-infected Marmoset B-lymphocytes (Bodescot *et al.*, unpublished results). We report here the analysis of a second cDNA, which corresponds to another mRNA of the family expressed in B95-8.

MATERIALS and METHODS

Culture of cells and preparation of cytoplasmic polyadenylated RNAs. B95-8 cells (10) were maintained in exponential growth in RPMI 1640 medium with 10% fetal calf serum (Gibco Laboratories). Cytoplasmic RNAs were prepared as described (11, 12). Polyadenylated RNAs were selected by chromatography on oligo(dT)-cellulose (Collaborative Research) as described (12, 13).

Synthesis and cloning of cDNA. Double-stranded cDNA molecules were synthesized from cytoplasmic polyadenylated RNAs (14) and made blunt-ended by treatment with S1 nuclease (Sigma) and Klenow fragment of *E.coli* DNA polymerase I (Biolabs). The cDNA molecules were cloned in the λgt10 bacteriophage as described (15), with modifications. They were methylated by EcoRI methylase (Biolabs) before ligation with EcoRI linkers (Biolabs) and digestion by EcoRI restriction endonuclease (Boehringer Mannheim). They were separated from the remaining linkers by chromatography on Sephadex G-75. The longest cDNA molecules were then selected by centrifugation through a sucrose density gradient and ligated with the EcoRI arms of the λgt10 bacteriophage. The resultant molecules were packaged *in vitro* (16). The library was titered on *E.coli* C600r^-m^+ and amplified on the *Hfl*A150 strain BNN102.

In situ hybridization. *In situ* hybridization was performed as

described (12). The probe was prepared by nick-translation (17). DNA sequencing. DNA sequencing was performed as described (18, 19), with modifications (20). The pUC13 plasmid (21) and the mp8 and mp9 derivatives of the M13 bacteriophage (22) were used as vectors.

RESULTS

We have constructed a cDNA library from the cytoplasmic polyadenylated RNAs of B95-8 cells into the λgt10 bacteriophage. About 1 x 10^6 recombinants were screened by in situ hybridization. The pDK10 plasmid (23) containing the B95-8 BamHI-C fragment, which is located in the US region, was used as a probe (Figure 1A and 1B). One of the isolated clones, designated T2, is presented here. The T2 cDNA is 3.5 kbp. A restriction map and the sequencing strategy are shown in Figure 2 ; the nucleotide sequence is given in Figure 3.

The nucleotide sequence of the T2 cDNA was compared with the nucleotide sequence of the B95-8 viral genome (4). This comparison showed that the corresponding mRNA contains exons transcribed from the US, IR and UL regions (Figure 1). Two exons are transcribed from the BamHI-C fragment, which is located in the US region, five exons from the BamHI-W fragment, which corresponds to the IR region, and two exons from the BamHI-L and E fragments, which are located in the UL region. The 5' end of the T2 cDNA is located at nucleotide 11,382 in the BamHI-C fragment and the first exon ends at nucleotide 11,479. The second

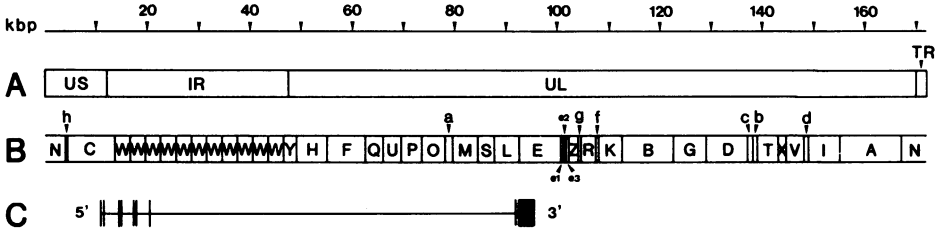

Figure 1 : A/ Linear representation of the EBV genome. Two clusters of tandemly repeated sequences, designated TR and IR, delimit the US and UL regions. B/ BamHI restriction map of the B95-8 viral genome. C/ The structure of the T2 cDNA and the position of the exons onto the viral genome. The exons are represented by vertical bars.

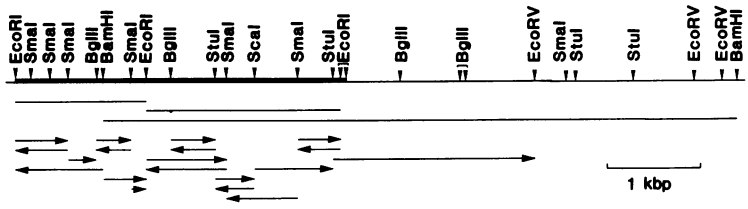

Figure 2 : The T2 cDNA as inserted in the λgt10 vector : a restriction map and the sequencing strategy. The T2 cDNA and the adjacent regions of the λgt10 vector are represented by thick and thin lines, respectively. Three restriction fragments overlaying the T2 cDNA were cloned into the pUC13 plasmid and are represented by horizontal lines. Several restriction fragments were prepared from the three recombinant plasmids and cloned into the M13 mp8 and mp9 vectors. These fragments and the direction of DNA synthesis in the sequencing reactions are shown by horizontal arrows.

exon, 32 nucleotides, starts at nucleotide 11,626 in the BamHI-C fragment and at nucleotide 99 of the T2 cDNA. Two exons, 66 and 132 nucleotides, are transcribed from the BamHI-W fragment and start at nucleotides 14,554 and 14,701 relative to the first copy of the BamHI-W fragment. These two exons constitute a unit which is tandemly repeated twice ; the repeats start at nucleotides 131 and 329 of the T2 cDNA. A third copy of the 66 nucleotide exon starts at nucleotide 527 of the T2 cDNA. The eighth exon, 344 nucleotides, starts at nucleotide 92,238 in the BamHI-L fragment and at nucleotide 593 of the T2 cDNA. The ninth exon, 2,579 nucleotides, starts at nucleotide 92,670 in the BamHI-L fragment and at nucleotide 937 of the T2 cDNA. This exon ends at nucleotide 95,248 in the BamHI-E fragment. The sequences of the viral genome corresponding to the junctions between the exons and the introns (Table 1) follow the rule previously established (24, 25).

A long open reading frame extends from nucleotide 598 of the T2 cDNA in the eighth exon to nucleotide 3,429 in the ninth exon (Figure 3). The corresponding polypeptide is 944 amino acids long and has a predicted molecular weight of 103,377. Its amino acid composition is given in Table 2. It contains 46 % nonpolar, 31 % polar, 12 % acidic and 11 % basic side chains. Its proline residue content is 13 %, which is unusually high. Two potential

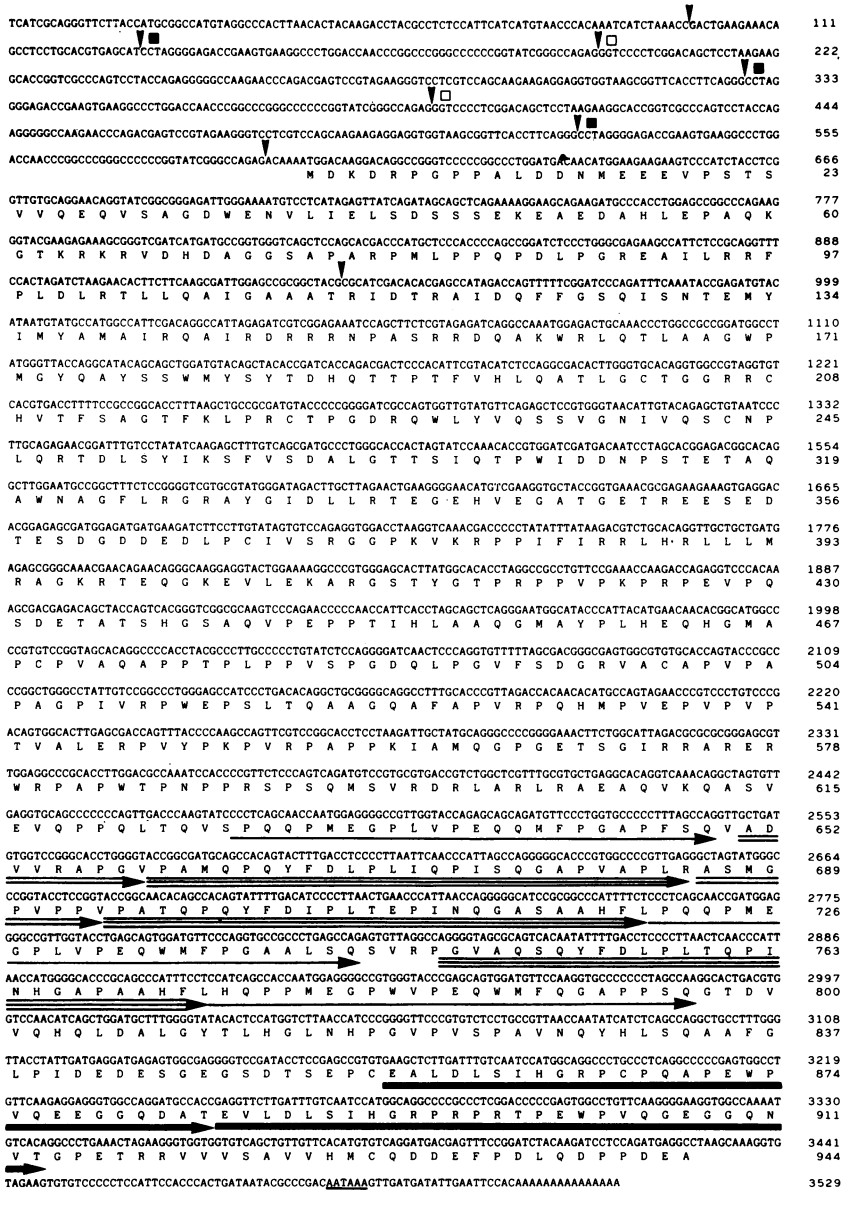

Figure 3 : The nucleotide sequence of the T2 cDNA and the amino acid sequence corresponding to the long open reading frame. The junctions between exons are indicated by vertical arrows. The beginning of the 66 and 132 nucleotide exons are shown by black and white squares, respectively. The type A, B, C and D sequences are underlined by single, double, triple and thick arrows, respectively. The AATAAA sequence is underlined.

Table 1 : Nucleotide sequences around the donor and acceptor sites corresponding to the T2 cDNA. Numbers refer to the position of the adjacent nucleotide in the T2 cDNA. The consensus sequences are underlined.

	Exon		Intron		Exon	
95	ACC	GTAAGT	...TTCCCTCTAG	GA		101
127	CAT	GTATCT	...GCCATCCAAG	CC		133
193	GAG	GTAAGT	...CCCGTCTCAG	GG		199
325	GGG	GTAAGT	...GCCATCCAAG	CC		331
391	GAG	GTAAGT	...CCCGTCTCAG	GG		397
523	GGG	GTAAGT	...GCCATCCAAG	CC		529
589	GAG	GTAAGT	...TTTGTTGCAG	AC		595
933	ACG	GTGAGC	...TTGGTTTCAG	CG		939
	A	A	CCCCCC C	G		
	AG	GT AGT	...	X AG	G	
	C	G	TTTTTT T		T	

N-linked glycosylation sites (26) are found at amino acids 312 and 911.

The 3' end of the long open reading frame contains repetitive sequences. They are made up of direct repeats of four types of sequences designated as A, B, C and D, which are 70, 26, 78 and 87 nucleotides long, respectively. There are three copies of the type A and C sequences and two copies of the type B and D sequences (Figure 3). The type A, C and D sequences correspond to repeated amino acid sequences in the polypeptide. At the amino acid level, the type A sequences exhibit 74 and 87 % homology, the type C sequences exhibit 58, 61 and 69 % homology and the type D sequences exhibit 76 % homology (Figure 4). The type B sequences are poorly conserved at the amino acid level.

The 3' end of the T2 cDNA is made up of a poly(dA) tail. The 3' non-coding region contains an AATAAA sequence located at nucleotide 3,488, 27 nucleotides upstream from the beginning of the tail. This sequence should correspond to the polyadenylation signal (27).

Table 2 : Amino acid composition of the polypeptide corresponding to the long open reading frame of the T2 cDNA.

Residue	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Number	85	10	50	60	27	69	24	32	18	63	25	14	124	71	64	58	50	65	16	19

```
Type A :
626  P Q Q P M E G P L V P E Q Q M F P G A P F S Q  650
     * * * * * * * * * * * *    * * * * *      * *          87 %
720  P Q Q P M E G P L V P E Q W M F P G A A L S Q  744
        *    * * * * *    * * * * * * * *    * *      * *    74 %
773  H Q P P M E G P W V P E Q W M F Q G A P P S Q  797
        *    * * * * *    * * * *    * *    * * *    * *    74 %
626  P Q Q P M E G P L V P E Q Q M F P G A P F S Q  650

Type C :
659  P A M Q P Q Y F D L P L I Q P I S Q G A P V A P L R  686
     * *    * * * * * *    * *       * *    * * *      *          61 %
694  P A T Q P Q Y F D I P L T E P I N Q G A S A A H F L  721
        *    * * * *    * * *    * * *    * *    * * * * *       69 %
747  G V A Q S Q Y F D L P L T Q P I N H G A P A A H F L  774
           *    * * * * * * * *    * * *       * * *    *       58 %
659  P A M Q P Q Y F D L P L I Q P I S Q G A P V A P L R  686

Type D :
855  E A L D L S I H G R P C P Q A P E W P V Q E E G G Q D A T  885
     *    * * * * * * * * * *    *       * * * * * * *    * * *      76%
884  E V L D L S I H G R P R P R T P E W P V Q G E G G Q N V T  914
```

Figure 4 : Comparison of the type A, C and D amino acid sequences. Numbers refer to the position of the adjacent amino acid in the long open reading frame of the T2 cDNA. The residues which are identical are shown by stars. The percentages of identical residues are indicated.

DISCUSSION

We have studied the structure of the EBV mRNAs expressed in the B95-8 cell line. We report here the characterization of a cDNA, designated T2, which contains exons transcribed from the US, IR and UL regions. The corresponding transcription unit is at least 84 kbp. Two exons are transcribed from the US region, five from the IR region and two from the UL region. The exons from the IR region consist of two tandem repeats of a unit containing two exons, 66 and 132 nucleotides, and of a third copy of the 66 nucleotide exon. The exons from the UL region contain a long open reading frame.

The T2 cDNA contains an unusually long 5' untranslated region, 597 nucleotides. The mRNA coding for the mouse ornithine decarboxylase (28) and those coding for the mouse and the human myc proteins (29, 30) contain long 5' untranslated regions as well, which seem to be involved in regulation of the translation process (31, 32). The 5' untranslated region of the T2 cDNA contains the exons transcribed from the IR region. For that reason, its structure is reminiscent of that of the late Polyoma virus mRNAs. Indeed, the 5' untranslated region of these mRNAs contains a tandemly-repeated exon, 57 nucleotides (33).

The IR region of the viral genome consists of eleven copies of a unit corresponding to the BamHI-W fragment. However, the T2 cDNA contains only two tandem repeats of the unit consisting of the 66 and 132 nucleotide exons transcribed from the IR region, and a third copy of the 66 nucleotide exon. Therefore, it is impossible to know from which repeats of the viral genome these exons were transcribed and how the repeats of the precursor RNA were spliced. For example, we do not know whether a given repeat of the T2 cDNA was transcribed from a single repeat of the viral genome. Furthermore, DNA rearrangements might have occurred in bacteria so that the T2 cDNA might be a derivative of a cDNA molecule containing more repeats. At any rate, since the T2 cDNA contains an incomplete copy of the unit, consisting of the third copy of the 66 nucleotide exon, all the repeats of the precursor RNA were not spliced in the same way.

The C-terminal end of the polypeptide corresponding to the long open reading frame of the T2 cDNA contains three types of repeated sequences. Whether they appeared recently or some constraint prevented them from diverging remains to be elucidated. The viral genome harbors several other types of repeated sequences and some of them are located in translated regions (4). For example, the EBNA-1 nuclear antigen contains a repetitive amino acid sequence whose size varies depending on the strain (34, 35, 36). Furthermore, an mRNA was described which contains tandem repeats of exons and should enable translation of a repetitive polypeptide (9).

It seems reasonable to assume that the T2 cDNA represents an almost complete copy of the corresponding mRNA. Indeed, the 3' untranslated region of the T2 cDNA ends with a poly(dA) tail and contains the AATAAA sequence characteristic of polyadenylation signals (27). Furthermore, a CCAAT and a TACAAAA sequence, 37 nucleotides apart, are located on the viral genome 115 and 77 nucleotides upstream from the 5' end of the T2 cDNA, respectively (4). These sequences are close to the consensus sequences for RNA polymerase II promoters (24) and should be part of the promoter which enables transcription of the corresponding RNA. In this case, only about 50 nucleotides should be missing in the T2 cDNA.

The comparison of the structure of the T2 cDNA with that of

the previously-described T1 cDNA (9) showed that the corresponding mRNAs were produced by alternative splicing. Indeed, the T1 and T2 cDNAs contain the same exons from the IR region and different exons from the UL region. Whereas the 132 nucleotide exon joins the exons from the IR region to those from the UL region in the T1 cDNA, the 66 nucleotide exon does this in the T2 cDNA. Furthermore, other mRNA species have been shown to contain the exons from the US and IR regions which are present in the T2 cDNA. The exons which differ are transcribed from the UL region and, as observed for those of the T2 cDNA, they contain open reading frames (Bodescot et al., in preparation).

ACKNOWLEDGEMENTS

We thank Dr. R. Young and Dr. R. Davis for the gift of the λ gt10 bacteriophage and Dr. E. Kieff for the gift of the pDK10 plasmid. We acknowledge Dr. P. Farrell and Dr. P. Sheldrick for helpful discussions. This investigation was supported by grants from the "Centre National de la Recherche Scientifique", the "Fondation pour la Recherche Médicale Française" and the "Association pour la Recherche sur le Cancer".

REFERENCES

1. Kieff, E., Dambaugh, T., Hummel, M. and Heller, M. (1983), in Klein, G. (ed.), Advances in Viral Oncology, Raven Press, New York, 3, pp. 133-182.
2. Henle, W. and Henle, G. (1985), in Klein, G. (ed.), Advances in Viral Oncology, Raven Press, New York, 5, pp. 201-238.
3. Griffin, B. and Karran, L. (1984), Nature, 309, 78-82.
4. Baer, R., Bankier, A., Biggin, M., Deininger, P., Farrell, P., Gibson, T., Hatfull, G., Hudson, G., Satchwell, S., Seguin, C., Tuffnell, P. and Barrell, B. (1984), Nature, 310, 207-211.
5. Fennewald, S., Van Santen, V. and Kieff, E. (1984), J. Virol., 51, 411-419.
6. Beisel, C., Tanner, J., Matsuo, T., Thorley-Lawson, D., Kezdy, F. and Kieff, E. (1985), J. Virol., 54, 665-674.
7. Hudson, G., Farrell, P. and Barrell, B. (1985), J. Virol., 53, 528-535.
8. Weigel, R. and Miller, G. (1985), J. Virol., 54, 501-508.
9. Bodescot, M., Chambraud, B., Farrell, P. and Perricaudet, M. (1984), The EMBO J., 3, 1913-1917.
10. Miller, G., Shope, T., Lisco, H., Stitt, D. and Lipman, M. (1972), Proc. Natl. Acad. Sci. USA, 2, 383-387.
11. Favaloro, J., Treisman, R. and Kamen, R. (1980), in Grossman, L. and Moldave, K. (eds), Methods in Enzymology, Academic Press, New York, 65, pp. 718-749.

12. Maniatis, T., Fritsch, E. and Sambrook, J. (1982), Molecular cloning : a laboratory manual, Cold Spring Harbor Laboratory.
13. Aviv, H. and Leder, P. (1972), Proc. Natl. Acad. Sci. USA, 69, 1408-1412.
14. Gubler, U. and Hoffman, B. (1983), Gene, 25, 263-269.
15. Huynh, T., Young, R. and Davis, R. (1985), in Glover, D. (ed.), DNA cloning : a practical approach, IRL Press, Oxford, England, 1, pp. 49-78.
16. Enquist, L. and Sternberg, N. (1979), in Wu, R. (ed.), Methods in Enzymology, Academic Press, New York, 68, pp. 281-298.
17. Rigby, P., Dieckmann, M., Rhodes, C. and Berg, P. (1977), J. Mol. Biol., 113, 237-251.
18. Sanger, F., Nicklen, S. and Coulson, A. (1977), Proc. Natl. Acad. Sci. USA, 74, 5463-5467.
19. Sanger, F., Coulson, A., Barrell, B., Smith, A. and Roe, B. (1980), J. Mol. Biol., 143, 161-178.
20. Biggin, M., Gibson, T. and Hong, G. (1983), Proc. Natl. Acad. Sci. USA, 80, 3963-3965.
21. Messing, J. (1983), in Wu, R., Grossman, L. and Moldave, K. (eds), Methods in Enzymology, Academic Press, New York, 101, pp. 20-78.
22. Messing, J. and Vieira, J. (1982), Gene, 19, 269-276.
23. Dambaugh, T., Beisel, C., Hummel, M., King, W., Fennewald, S., Cheung, A., Heller, M., Raab-Traub, N. and Kieff, E. (1980), Proc. Natl. Acad. Sci. USA, 77, 2999-3003.
24. Breathnach, R. and Chambon, P. (1981), Annu. Rev. Biochem., 50, 349-383.
25. Mount, S. (1982), Nucleic Acids Res., 10, 459-472.
26. Hubbard, C. and Ivatt, R. (1981), Annu. Rev. Biochem., 50, 555-583.
27. Proudfoot, N. and Brownlee, G. (1976), Nature, 263, 211-214.
28. Kahana, C. and Nathans, D. (1985), Proc. Natl. Acad. Sci. USA, 82, 1673-1677.
29. Bernard, O., Cory, S., Gerondakis, S., Webb, E. and Adams, J. (1983), The EMBO J., 2, 2375-2383.
30. Stanton, L., Watt, R. and Marcu, K. (1983), Nature, 303, 401-406.
31. Saito, H., Hayday, A., Wiman, K., Hayward, W. and Tonegawa, S. (1983), Proc. Natl. Acad. Sci. USA, 80, 7476-7580.
32. Darveau, A., Pelletier, J. and Sonenberg, N. (1985), Proc. Natl. Acad. Sci. USA, 82, 2315-2319.
33. Treisman, R. (1980), Nucleic Acids Res., 8, 4867-4888.
34. Hennessy, K., Heller, M., Van Santen, V. and Kieff, E. (1983), Science, 220, 1396-1398.
35. Hennessy, K. and Kieff, E. (1983), Proc. Natl. Acad. Sci. USA, 80, 5665-5669.
36. Dillner, J., Sternas, L., Kallin, B., Alexander, H., Ehlin-Henriksson, B., Jörnvall, H., Klein, G. and Lerner, R. (1984), Proc. Natl. Acad. Sci. USA, 81, 4652-4656.