Amino acid sequence homology in *gag* region of reverse transcribing elements and the coat protein gene of cauliflower mosaic virus

Simon N.Covey

Department of Virus Research, John Innes Institute, Colney Lane, Norwich NR4 7UH, UK

Received 25 October 1985; Revised and Accepted 10 December 1985

ABSTRACT

A nucleic acid binding protein (NBP) derived from the gag gene of retroviruses that is thought to interact with genomic RNA in virion cores, contains a highly conserved arrangement of cysteine residues. A search of available nucleic acid and protein sequences has revealed that the motif $CysX_2CysX_4HisX_4Cys$ (NBPcys) is invarient in all replication competent retroviruses, a Syrian hamster intracisternal A-particle gene, the Drosophila retrotransposon copia and in cauliflower mosaic virus (CaMV). In each case, NBPcys is located in that part of the 'gag-pol' region just preceding a conserved protease amino acid sequence. This is of special significance for CaMV as NBPcys is in the coat protein gene (ORF IV) upstream of the putative reverse transcriptase gene (ORF V) and demonstrates that the gag-pol arrangement of reverse transcribing elements is preserved in CaMV. Moreover, CaMV differs from all other known NBPcys-containing elements in that it packages a DNA genome in virions.

INTRODUCTION

Recently, there has been considerable interest in relationships amongst retroviruses, hepatitis B viruses (HBVs), Drosophila copia and yeast Ty transposable elements (retrotransposons), and cauliflower mosaic virus (CaMV), each considered to undergo reverse transcription [1,2]. In particular, amino acid sequence comparisons of putative pol gene products have revealed positional regions of homology in a protease domain, a reverse transcriptase domain and, for those elements that integrate a DNA 'provirus', an endonuclease domain [3-5].

In most retroviruses, the pol gene is preceded by an open reading frame (ORF), the gag (group specific antigen) gene, that encodes virion core structural proteins. The primary translation product of the gag gene is a polyprotein subsequently cleaved to generate individual virion core polypeptides to which certain functions have been ascribed. One of these is the highly basic nucleic acid binding protein (NBP) that originates from a domain in the C-terminal half of the gag polyprotein and is thought to interact directly with genomic RNA in retrovirus particles [see 6]. It has been noted [7,8], that retrovirus NBPs contain a highly conserved arrangement of cysteine residues: $CysX_2CysX_9Cys$ (here designated NBPcys). In some retroviruses, for example Moloney murine leukemia virus (MoMLV) [9], there is one copy of NBPcys whilst in others, for example Rous sarcoma virus (RSV) [10], the sequence is duplicated in tandem.

Because NBPcys is so highly conserved in both sequence and positional terms, I undertook a search for, and comparison of, this domain amongst reverse transcribing elements in general for which sequence data are available. Amongst other findings, the discovery of NBPcys in the coat protein gene of CaMV is of particular interest as it extends retroviral gag gene homology to this plant virus which is unusual in that it packages DNA rather than RNA in virions.

Survey and Sources

A computer-assisted search was with the Protein Identification Resource (PIR) data base, supported by the Division of the Research Resources of the National Institutes of Health, USA. Additionally, DNA and/or protein sequence data were extracted from published articles for the following elements: feline leukemia virus (FeLV) [7]; Moloney murine leukemia virus (MoMLV) [9]; Rous sarcoma virus (RSV) strain Pr-C [10]; cauliflower mosaic virus (CaMV) isolates Strasbourg [11], CM1841 [12], D/Hungary [13]; simian sarcoma virus (SSV) [14]; baboon endogenous virus (BaEV) [15]; avian sarcoma virus (ASV) [16]; AKR murine leukemia virus (AKV) [17]; Moloney murine sarcoma virus (MoMSV) [18]; human T-cell leukemia virus-1 (HTLV-1) [19]: human T-cell leukemia virus-2 (HTLV-2) [20]; bovine leukemia virus (BLV) [21]; human T-cell lymphotropic virus-3 (HTLV-3) [22]; lymphadenopathy-associated virus (LAV) [23]; AIDS-associated retrovirus (ARV-2) [24]; visnavirus [25]; Syrian hamster intracisternal A-particle gene (IAP-H18) [26]; Drosophila copia element [27] and copia-like element 17.6 [28]; hepatitis B viruses [29-32]. Single letter amino acid abbreviations are: A, alanine; R, arginine; N, asparagine; D, aspartate; C, cysteine; Q, glutamine; E, glutamate; G, glycine; H, histidine; I, isoleucine; L, leucine; K, lysine; M, methionine; F, phenylalanine; P, proline; S, serine; T, threonine; W, tryptophan; Y, tyrosine; V, valine.

Distribution of NBPcys

A computer-assisted search of the PIR data base for the sequence CX_2CX_9C, where X is any amino acid, was performed and this motif was found to be present in a number of protein sequences apparently not related to retroviral genes (data not shown) in addition to those expected for retroviruses. However, closer inspection of all available sequence data revealed that the arrangement of cysteine residues in reverse transcribing elements as a whole contained, in addition, an invarient histidine residue 8 amino acids downstream from the first cysteine (denoted n) at position n+8 (see Fig. 2). Thus, the arrangement $CX_2CX_4HX_4C$ (NBPcys) was found to be conserved amongst retrovirus -related elements and was not found in any unrelated protein sequences. The occurrence of NBPcys has been reported previously for a number of retroviral elements [7,22-25,27] and this search found that it is present in all replication competent retroviruses, in three isolates of CaMV and also in IAP-H18. NBPcys was not detected in hepatitis B viruses and, as previously noted [27], it is absent from the yeast Ty element and the *Drosophila copia*- like element 17.6 [28].

Genomic location of NBPcys

The co-ordinates of NBPcys domains of reverse transcribing elements are summarized in the Table. In those cases where positions have been determined from nucleotide sequence data, the first cysteine residue (n) is numbered from the first amino acid of the ORF in which it occurs. For NBPs that have been directly sequenced (marked with an asterisk in the table) the cys residue (n) is numbered from the N-terminal amino acid in the particular NBP.

From the table it can be seen that several mammalian retroviruses and also CaMV and *copia*, contain only one copy of NBP cys, in others (including IAP-H18) it is duplicated in tandem (labelled a & b in the table). The duplicates are separated by

Table. Locations of NBPcys in reverse transcribing elements

		amino acid residues			
Virus	ORF/protein	NBPcys(a)	spacing	NBPcys(b)	ref.
RSV	*gag* P12	509	12	535	10
ASV*	*gag* P12	21	11	47	16
SSV	*gag* P10	491	-	-	14
FeLV*	*gag* P10	30	-	-	7
MoMULV	*gag* P10	503	-	-	9
AKV	*gag* P10	503	-	-	17
MoMSV	*gag* P10	503	-	-	18
HTLV1	*gag* P15	356	9	379	19
HTLV2	*gag* P15	361	9	384	20
BLV	*gag* P12	347	11	372	21
HTLV3	*gag* P15	391	7	412	22
LAV	*gag* P15	391	7	412	23
ARV2	*gag* P15	393	7	414	24
visna	*gag* P14	387	5	406	25
IAP-N18	ORF 2	116	11	141	26
copia	ORF 1	232	-	-	27
CaMV	ORF IV coat	414	-	-	11-13

between 5 and 12 amino acids depending upon the virus. In all replication competent retroviruses, NBPcys is present in that virion core protein, the nucleic acid binding protein, derived from a region in the C-terminal half of the *gag* gene product. In IAP-H18, the duplicated sequence is in ORF 2 considered to be equivalent to part of the retroviral *gag* gene [26] and in *copia* [27] NBPcys is towards the N-terminal region of a putative gene product from a single long ORF in this element. In each of three sequenced CaMV isolates [11-13], NBPcys is located towards the C-terminal part of ORF IV, the viral coat protein gene [33] (Fig. 1). Within the genome of each element, the NBPcys sequences are located just upstream of a highly conserved arrangement of amino acids thought to constitute part of a protease domain [5]. This protease 'core' sequence is positioned in an N-terminal region of the *pol* gene of most retroviruses and ORF V, the putative reverse transcriptase gene, of CaMV. In RSV, the protease core is downstream of the duplicated NBPcys sequence but in an extended portion of the *gag* ORF. The protease domain of HTLV-2 is in a small ORF that overlaps both *gag* and *pol* genes and in

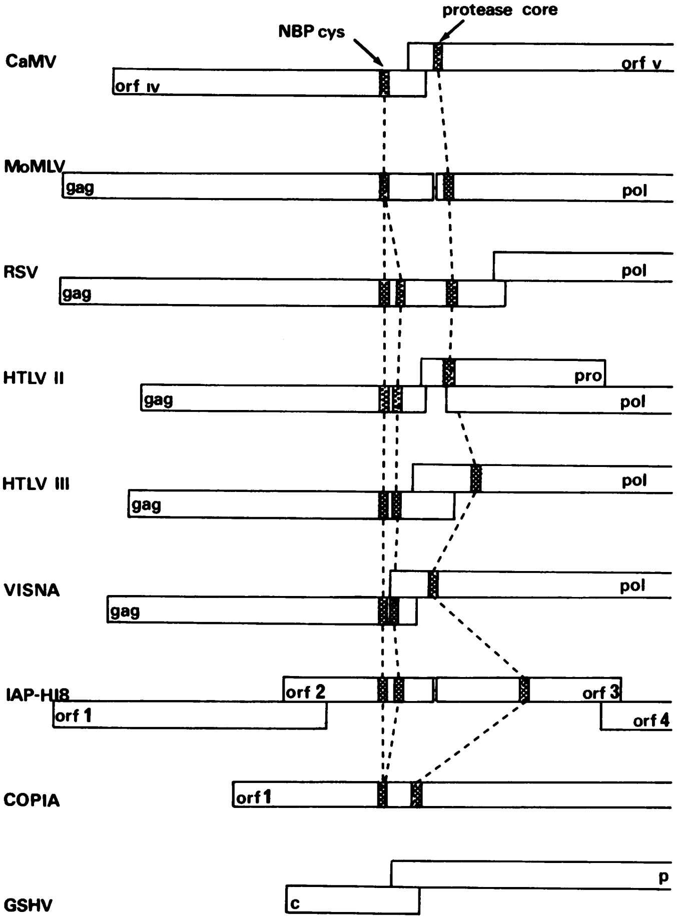

Figure 1. Location of the conserved $CX_2CX_4HX_4C$ motif (NBPcys) in the *gag* gene of retroviruses and other reverse transcribing elements. Part of the adjacent *pol* ORFs are shown with the conserved protease 'core' amino acid sequence; other homologies within *pol* have been described previously [4]. For ground squirrel hepatitis virus (GSHV) [32], the core antigen ORF (c) and part of the putative *pol* ORF (p) are shown. The elements are drawn to scale and for reference the CaMV coat protein gene (ORF IV) is 1500 bp long.

IAP-H18, it is present in ORF 3. The protease core sequence of *copia* is downstream of NBPcys in the single long ORF. Thus, NBPcys is conserved in both sequence and positional terms within the genomes of reverse transcribing elements (Fig. 1).

In replication-defective retroviruses that carry transform-

				n			n+3					n+8					n+13			tRNA
avian	RSV(a)/ASV(a)	G	L	C	Y	T	C	G	S	P	G	H	Y	Q	A	Q	C	P	K	trp
	RSV(b)	E	R	C	Q	L	C	N	G	M	G	H	N	A	K	Q	C	R	K	
	ASV(b)	E	R	C	E	L	C	N	G	M	G	H	N	A	K	Q	C	R	K	
	SSV	D	Q	C	A	Y	C	K	E	K	G	H	W				C	T	G	
	BaEV	D	Q	C	A	Y	C	K	E	R	G	H	W	T	K	D	C	P	K	
	FeLV	D	Q	C	A	Y	C	K	E	K	G	H	W	V	R	D	C	P	K	
	MoMLV/AKV	D	Q	C	A	Y	C	K	E	K	G	H	W	A	K	D	C	P	K	
	MoMSV	D	Q	C	T	Y	C	E	E	Q	G	H	W	A	K	D	C	P	K	
mammalian	HTLV1(a)	Q	P	C	F	R	C	G	K	A	G	H	W	S	R	D	C	T	Q	pro
	HTLV2(a)	Q	P	C	F	R	C	G	K	V	G	H	W	S	R	D	C	T	Q	
	BLV(a)	G	P	C	Y	R	C	L	K	E	G	H	W	A	R	D	C	P	T	
	HTLV1(b)	G	P	C	P	L	C	Q	D	P	T	H	W	K	R	D	C	P	R	
	HTLV2(b)	G	P	C	P	L	C	Q	D	P	S	H	W	K	R	D	C	P	Q	
	BLV(b)	G	P	C	P	I	C	K	D	P	S	H	W	K	R	D	C	P	T	
	HTLV3(a)	V	K	C	F	N	C	G	K	E	G	H	T	A	R	N	C	R	A	
	LAV(a)	V	K	C	F	N	C	G	K	E	G	H	I	A	R	N	C	R	A	
	ARV2(a)	V	K	C	F	N	C	G	K	E	G	H	I	A	K	N	C	R	A	
lentiviruses	visna(a)	Q	K	C	Y	N	C	G	K	P	G	H	L	A	R	Q	C	R	Q	lys
	HTLV3(b)/LAV(b)	K	G	C	W	K	C	G	K	E	G	H	Q	M	K	D	C	T	E	
	ARV2(b)	K	G	C	W	R	C	G	R	E	G	H	Q	M	K	D	C	T	E	
	visna(b)	I	I	C	H	H	C	G	K	R	G	H	M	Q	K	D	C	R	Q	
	IAP-H18(a)	K	A	C	F	N	C	G	R	M	G	H	L	K	K	D	C	Q	A	phe
	IAP-H18(b)	K	L	C	Y	R	C	G	K	G	Y	H	R	A	S	E	C	R		
	copia	V	K	C	H	H	C	G	R	E	G	H	I	K	K	D	C	F	H	?
	CaMV	C	R	C	W	I	C	N	I	E	G	H	Y	A	N	E	C	P	N	met
	Ty			--------------------------																met
	17.6			--------------------------																?
	HBVs			--------------------------																-

Figure 2. Amino acid sequences of NPBcys in reverse transcribing elements. Boxed residues are invarient. The sequences are grouped according to familial similarities. Sequences imperfectly duplicated within any one element are denoted (a) and (b). In the replication defective retrovirus SSV, seven unrelated amino acids (DEEIAPA) are located between n+10 and n+12. The tRNA primers of DNA minus-strand synthesis used by each group of elements is shown to the right.

ing genes, NBPcys is usually absent because the gag domain containing it is either deleted or substituted by onc sequences. Notable exceptions are MoMSV [18] in which deletions in the pol gene and a substitution of env sequences for the onc gene mos are downstream of the gag NBP domain, and SSV [14] which has a sis oncogene within the pol region. However, NBPcys in SSV overlaps the start of the pol ORF in a different reading frame and its composition is unusual in that although the first ten residues (n to n+9) are identical to those of MoMLV, seven unlelated amino acids (DEEIAPA) are located between n+10 and n+12 (see Fig. 2).

Amino acid variations

The NBPcys amino acid sequences of reverse transcribing elements are shown in Fig. 2 together with adjacent amino acids. Some general features of a few retroviral NBPcys sequences have been noted previously [7] and the comparison is extended here to other elements. The arrangement of 3 cysteine residues (n, n+3, n+13) together with a histidine at position n+8 is totally invarient. Position n+1 is an aromatic or heterocyclic amino acid except in those mammalian retroviruses that have only one NBPcys sequence in which n+2 is the aromatic amino acid tyrosine. Visnavirus and copia have the heterocyclic amino acid histidine at both n+1 and n+2. In general, amino acids in positions n+4 to 6, n+9 and n+10 are variable although there are similarities that group certain elements together (discussed below). At n+7, all elements have at least one glycine residue and for HTLV-1, HTLV-2, BLV and IAP-H18 this is in the first of the duplicated NBPcys sequences. Position n+11 is a conserved basic amino acid except for RSV(a), ASV(a), IAP-H18(b) and CaMV. An acidic amino acid is usually in position n+12 except in avian retroviruses and visnavirus(a) where a glutamine residue has probably resulted from a point mutation of a glutamate triplet.

In addition to general similarities of amino acid positions within NBPcys noted above, there are others that group elements in families (Fig. 2). In mammalian retroviruses, n+9 is an invarient tryptophan residue and n+12 an invarient aspartate regardless of whether NBPcys is duplicated. The lentiviruses have invarient glycines at n+4 and n+7, and n+5 is a highly conserved lysine. Moreover, in the first of the duplicated NBPcys sequences of lentiviruses, n+2 is an invarient asparagine, n+10 an invarient alanine and outside of the cys domain, n-1 is always lysine and n+14 an invarient arginine. In the second of the duplicated lentivirus NBPcys sequences, positions n+11 and n+12 are invarient lysine and aspartate residues respectively. Placing the AIDS retroviruses in the lentivirus family is consistent with recent observations [25,34]. The NBPcys amino acids of avian retroviruses exhibit close familial similarities but differ from those of other retroviruses. The remaining elements are distinct although IAP-H18, with two NBP cys sequences, and copia with one, have features of the lentivirus group. The variable residues in the CaMV NBPcys sequence exhibit perhaps the greatest difference compared with the other elements and the amino acids at positions n+5 and n+11 are unique to CaMV presumably reflecting its evolutionary divergence.

In any one reverse transcribing element that has two NBPcys sequences, the second always differs from the first in a number of positions and so it is an imperfect repeat. Moreover, within families of retroviruses, the first NBPcys exhibits greater homology with that of other viruses in its group than it does with its own duplicate (Fig. 2). This suggests that the difference in the duplicated sequences of each virus is itself a conserved feature that might have functional significance for the nucleic acid binding protein. Quite why some elements have only one NBPcys sequence and others have two, remains at present unknown.

DISCUSSION

This survey has examined the distribution and composition of an amino acid sequence (NBPcys) that is highly conserved in the 'gag' gene of reverse transcribing elements isolated from sources as diverse as mammals, birds, flies and plants. A notable observation is that NBPcys is in the coat protein gene of CaMV demonstrating that the gag-pol arrangement of retroviruses extends to this unusual plant virus. Like the gag gene product of retroviruses, the CaMV 57K mol. wt. coat protein undergoes processing and phosphorylation [35] during assembly of virus particles. Similarly, the CaMV NBPcys sequence is located in that part of the coat protein rich in basic amino acids [11].

Conservation of the NBPcys sequence in CaMV suggests that it might be involved in packaging nucleic acid in virions. However, one of the fundamental differences between CaMV and retroviruses is that the former packages double-stranded DNA, whilst the latter RNA. It has been shown [36] that the RSV NBP interacts specifically with domains of genomic RNA although *in vitro* it exhibits a general and non-specific affinity for RNA and single-stranded DNA but a relatively low affinity for double-stranded DNA [37]. The NBPcys sequence appears, however, to be restricted to reverse transcribing elements and is not a general feature of virus coat protein nucleic acid binding domains, although an almost perfect reversal of it ($CX_3HX_5CX_2C$) has been observed [8] in the bacteriophage T_4 DNA binding protein [38]. Hepatitis B viruses fulfil many of the criteria for possession of NBPcys in that they replicate by reverse transcription, encapsidate a DNA molecule with extensive single-stranded regions and apparently retain the '*gag*-*pol*' arrangement of retroviruses [29-32] (in HBVs, the core antigen ORF can be considered equivalent to the '*gag*' gene) (Fig. 1). Somewhat unexpectedly, the search for NBPcys reported here, failed to detect this sequence in any of the HBVs.

One possible explanation for the absence of NBPcys in HBVs, and its specific distribution amongst retroviruses and CaMV, is that the sequence of cysteine residues is involved in sequestering the tRNA primer of DNA-minus strand synthesis within virus particles. HBVs do not use tRNA to prime minus-strand synthesis, a protein is thought to perform this function [39].

It is not known whether CaMV, which has a $tRNA^{met}$ primer [40], packages tRNA *per se* although appreciable quantities of minus-strand strong-stop DNA with the tRNA primer covalently attached are associated with CaMV virions in a DNase resistant form [41]. Furthermore, the familial groupings of retroviral elements, based solely upon similarities in NBPcys sequences, can be correlated with use of a specific tRNA primer (see Fig. 2) that is presumably packaged in virus particles in each case. However, it is possible that this correlation is fortuitous since NBPcys is absent from the *copia*-like element 17.6 and from the yeast Ty element although the definition of these retrotransposons as infectious virus entities, in the strict sense, that package RNA molecules for transmission, remains questionable at present; the *copia* element which contains NBPcys might be an exception to this.

ACKNOWLEDGEMENT

I thank Ian Evans for assistance with computer searches.

REFERENCES

1. Varmus, H.E. (1985). Nature **314,** 583-584.
2. Baltimore, D. (1985). Cell **40,** 481-482.
3. Toh, H., Hayashida, H. and Miyata, T. (1983). Nature **305,** 827-829.
4. Toh, H., Kikuno, R., Hayashida, H., Miyata, T., Kugimiya,W., Inouye, S., Yuki, S. and Saigo, K. (1985). EMBO J. **4,** 1267-1272.
5. Toh, M., Ono, M., Saigo, K. and Miyata, T. (1985). Nature **315,** 691.
6. Dickson, C., Eisenman, R. and Fan, H. (1985). In RNA Tumor Viruses, Eds. Weiss, R., Teich, N., Varmus, H. and Coffin,J. (Cold Spring Harbor Laboratory, U.S.A.) pp.135-145.
7. Copeland, T.D., Morgan, M.A. and Oroszlan, S. (1984). Virology **133,** 137-145.
8. Henderson, L.E., Copeland, T.D., Sowder, R.C., Smythers,G.W. and Oroszlan, S. (1981). J. Biol. Chem. **256,** 8400-8403.
9. Shinnick, T.M., Lerner, R.A., and Sutcliffe, J.G. (1981). Nature **293,** 543-548.
10. Schwartz, D.E., Tizard, R. and Gilibert,W. (1983). Cell **32,** 853-869.
11. Franck, A., Guilley, H., Jonard, G. and Richards, K. (1980). Cell **21,** 285-294.
12. Gardner, R.C., Howarth, A.J., Hahn,P., Brown-Luedi,M., Shepherd, R.J. and Messing, J. (1981). Nucleic Acids Res. **9,** 2871-2888.
13. Balàzs, E., Guilley, H., Jonard, G. and Richards, K. (1982). Gene **19,** 239-249.
14. Devare,S.G., Reddy,E.P., Law,J.D., Robbins,K.C. and Aaronson S.A. (1983). Proc. Natl. Acad. Sci. USA **80,** 731-735.
15. Tamura, T-A. (1983). J. of Virol. **47,** 137-145.
16. Misono, K.S., Sharief, F.S. and Leis, J. (1980). Fed. Proc. **39,** 1611.
17. Herr, W. (1984). J. Virol. **49,** 471-478.
18. van Beveran,C., van Straaten, F., Galleshaw, J.A. and Verma, I.M. (1981). Cell **27,** 97-108.
19. Seiki, M., Hattori, S., Hirayama, Y., and Yoshida,M. (1983). Proc. Natl. Acad. Sci. USA **80,** 3618-3622.
20. Shimotohno,K., Takahashi,Y., Shimizu,N., Gojobori,T., Golde, D.W., Chen,I.S.Y., Miwa, M., and Sugimura, T. (1985). Proc. Natl. Acad. Sci. USA **82,** 3101-3105.
21. Sagata, N., Yasunaga, T., Tsuzuku-Kawamura, J., Ohishi, K., Ogawa, Y. and Ikawa, Y. (1985). Proc. Natl. Acad. Sci. USA **82,** 677-681.
22. Ratner,L., Haseltine,W., Patarca, R., Livak, K.J., Starcich, B., Josephs,S.F., Doran,E.R., Rafalski,J.A., Whitehorn,E.A., Baumeister,K., Ivanoff,L., Petteway,S.R. Jr., Pearson, M.L., Lautenberger, J.A., Papas, T.S., Ghrayeb,J., Chang, N.T., Gallo, R.C. and Wong-Staal, F. (1985). Nature **313,** 277-284.
23. Wain-Hobson, S., Sonigo, P., Danos, O., Cole, S. and Alizon, M. Cell **40,** 9-17.
24. Sanchez-Pescador, R., Power, M.D., Barr, P.J., Steimer,K.S., Stempien, M.M., Brown-Shimer, S.L., Gee, W.W., Renard, A., Randolph, A., Levy, J.A., Dina, D. and Luciw, P.A. (1985). Science **227,** 484-492.
25. Sonigo, P., Alizon, M., Staskus, K., Klatzmann, D., Cole, S., Danos, O., Retzel, E., Tiollais, P., Haase, A. and Wain-

Hobson, S. (1985). Cell **42**, 369-382.
26. Ono, M., Toh, H., Miyata, T. and Awaya, T. (1985). J. Virol. **55**, 387-394.
27. Mount, S.M. and Rubin, G.M. (1985). Molec. Cell Biol. **5**, 1630-1638.
28. Saigo, K., Kugimiya, W., Matsuo, Y., Inouye, S., Yoshioka,K. and Yuki, S. (1984). Nature **312**, 659-661.
29. Galibert, F., Mandart, E., Fitoussi, F., Tiollais, P. and Charnay, P. (1979). Nature **281**, 646-650.
30. Mandart, E., Kay, A. and Galibert, F. (1984). J. Virol. **49**, 782-792.
31. Galibert, F., Chen, T.N. and Mandart, E. (1982). J. Virol. **41**, 51-65.
32. Seeger, C., Ganem, D. and Varmus, H.E. (1984). J. Virol **51**, 367-375.
33. Daubert, S., Richins, R., Shepherd, R.J. and Gardner, R.C. (1982). Virology **122**, 444-449.
34. Chiu, I-M., Yaniv, A., Dahlberg, J.E., Gazit, A., Skuntz, S. F., Tronick, S.R. and Aaronson, S.A. (1985). Nature **317**, 366-368.
35. Hahn, P. and Shepherd, R.J. (1980). Virology **107**, 295-297.
36. Darlix, J-L. and Spahr, P-F. (1982). J. Mol. Biol. **160**, 147-161.
37. Méric, C., Darlix, J-L. and Spahr, P-F. (1984). J. Mol. Biol. **173**, 531-538.
38. Williams, K.R., LoPresti, M.B. and Setoguchi, M. (1981). J. Biol. Chem. **256**, 1754-1762.
39. Molnar-Kimber, K.L., Summer, J., Taylor, J.M. and Mason,W.S. (1983) J. Virol. **45**, 165-172.
40. Hull, R. and Covey, S.N. (1983). TIBS **8**, 119-121.
41. Turner, D.S. and Covey, S.N. (1984). FEBS Letters **165**, 285-289.