The EMBO Journal vol.8 no.12 pp.3867 - 3874, 1989

Identification of four conserved motifs among the RNA-dependent polymerase encoding elements

Olivier Poch[1], Isabelle Sauvaget[2], Marc Delarue[3] and Noël Tordo[4]

[1]Laboratoire de Biochimie II, Institut de Biologie Moléculaire et Cellulaire du CNRS, 15 rue Descartes, 67084 Strabourg Cédex, [2]Unité d'Informatique Scientifique, Institute Pasteur, 25 rue du Docteur Roux, 75724 Paris Cédex 15, [3]Laboratoire de Cristallographie Biologique, Institut de Biologie Moléculaire et Cellulaire du CNRS, 15 rue Descartes, 67084 Strasbourg Cédex, [4] Unité de la Rage, Institut Pasteur, 25 rue du Docteur Roux, 75724 Paris Cédex 15, France

Communicated by D.Kolakofsky

Four consensus sequences are conserved with the same linear arrangement in RNA-dependent DNA polymerases encoded by retroid elements and in RNA-dependent RNA polymerases encoded by plus-, minus- and double-strand RNA viruses. One of these motifs corresponds to the YGDD span previously described by Kamer and Argos (1984). These consensus sequences altogether lead to 4 strictly and 18 conservatively maintained amino acids embedded in a large domain of 120 to 210 amino acids. As judged from secondary structure predictions, each of the 4 motifs, which may cooperate to form a well-ordered domain, places one invariant amino acid in or proximal to turn structures that may be crucial for their correct positioning in a catalytic process. We suggest that this domain may constitue a prerequisit 'polymerase module' implicated in template seating and polymerase activity. At the evolutionary level, the sequence similarities, gap distribution and distances between each motif strongly suggest that the ancestral polymerase module was encoded by an individual genetic element which was most closely related to the plus-strand RNA viruses and the non-viral retroposons. This polymerase module gene may have subsequently propagated in the viral kingdom by distinct gene set recombination events leading to the wide viral variety observed today.

Key words: evolution/homology/polymerase module/reverse transcriptase/RNA-dependent RNA polymerase

Introduction

As a consequence of the rareness of RNA-dependent polymerization processess encoded by their host cells, the RNA viruses were forced to develop very specific polymerase activities for the multiplication of their own RNA genome. Two main types of polymerase exist to perform this task: the RNA-dependent RNA polymerases leading to a strictly RNA life cycle and the RNA-dependent DNA polymerases (reverse transcriptases) in which the RNA genome represents a transient state leading to DNA and possible integration in the host genome. The RNA-dependent RNA polymerases are involved in the multiplication of the plus-, minus- and double-strand RNA viruses (reviewed in Ishihama and Nagata, 1988) while the reverse transcriptases are involved in the replication of retroid-elements including the retroviruses, the transposable integrated elements (non-viral and viral retroposons) and some DNA viruses, such as the hepadnaviruses (Weiner *et al.*, 1986; Doolittle *et al.*, 1989).

Despite wide variations among viruses in morphology, genome organization and sequences of their structural proteins, the polymerase sequences have revealed the conservation of large peptide regions. The RNA-dependent RNA polymerases display extended conserved regions as shown for the plus-strand RNA viruses (Kamer and Argos, 1984; Koonin *et al.*, 1987), the segmented minus-strand RNA viruses (Kemdirim *et al.*, 1986), as well as the unsegmented minus-strand RNA viruses (Tordo *et al.*, 1988). Nevertheless, no clear sequence similarity has been reported between these three distinct groups. On the other hand, in the reverse transcriptases, the published alignments circumscribed five highly conserved regions roughly centred around the ten invariant amino acids, reported by Toh *et al.* (1985). However, the recent characterization of numerous putative reverse transcriptases encoded by non-viral retroposons revealed that only five amino acids are strictly conserved (Hattori *et al.*, 1986; Schwartz-Sommer *et al.*, 1987; Boer and Gray, 1988).

Two recent findings have clearly indicated that interviral relationships across wide evolutionary distances may also exist. Kamer and Argos (1984) first reported that one of the most conserved regions of the RNA-dependent RNA polymerases (YGDD span) was also present in the RNA-dependent DNA polymerases. Argos (1988), recently extended this punctual relationship to the DNA-dependent DNA polymerases. Gorbalenya and Koonin (1988) demonstrated that the polymerase of the infectious bursal disease virus, a double-strand RNA virus, was related to the polymerases of the plus-strand RNA viruses.

Here we report that, at least four common motifs are conserved in the sequences of all the polymerases showing RNA template specificity. The secondary structure predictions of these RNA-dependent polymerases suggest that these four motifs seem to be well-ordered and could build a large domain of 120−210 amino acids that we propose to be a prerequisite 'polymerase module'. The analysis of the sequence similarities suggests that the ancestral genetic element encoding the 'polymerase module' must be searched for in an intermediate position between the plus-strand RNA viruses and some non-viral retroposons. A possible evolutionary scheme which is consistent with sequence similarities over such wide evolutionary distances is then discussed.

Results

First scan of the protein data bank

As a starting point, we selected a set of RNA-dependent DNA polymerase sequences (names underlined in Figure

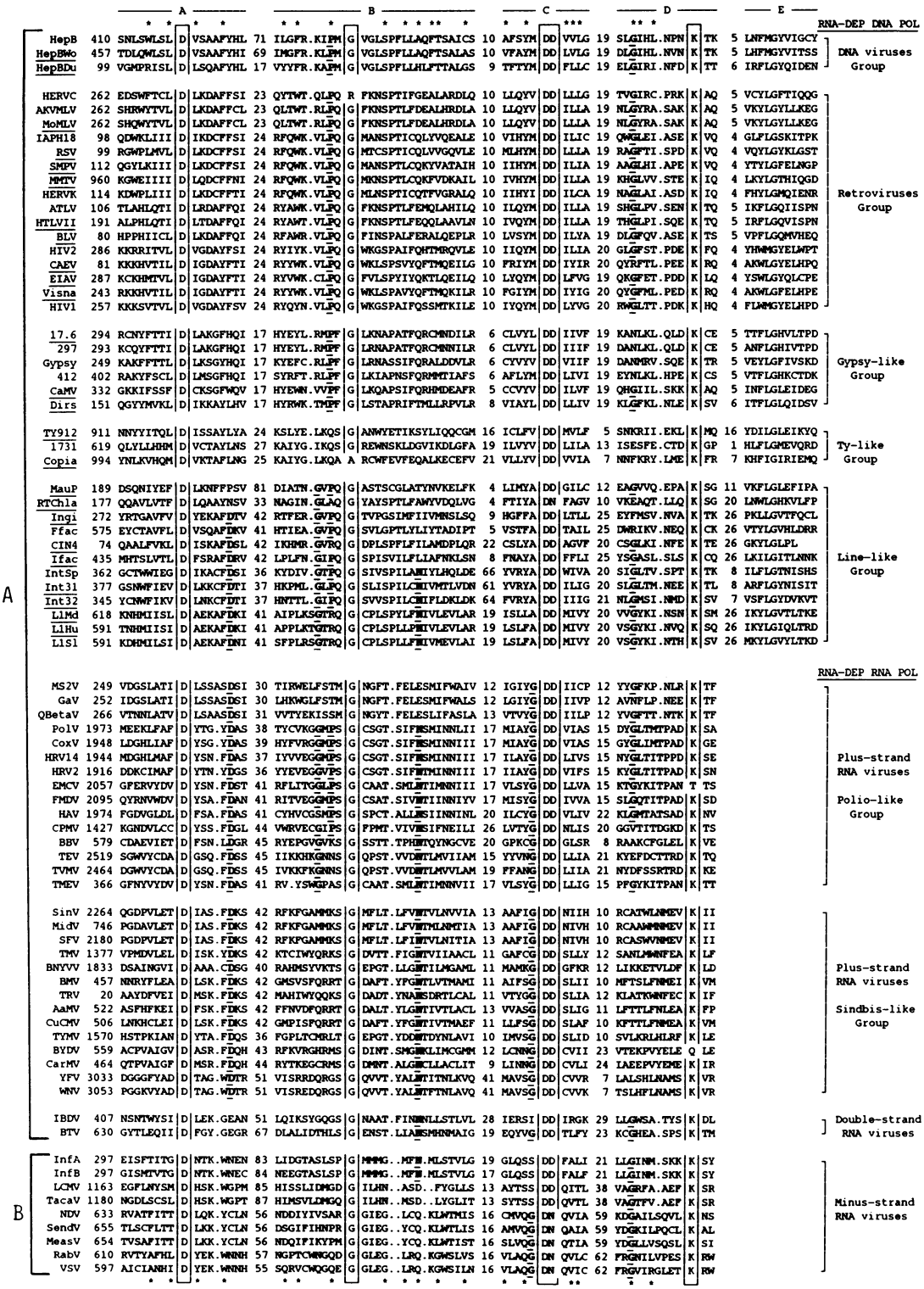

Fig. 1. Alignments of the five conserved motifs within known or probable RNA-dependent polymerases. Names underlined are those of the sequences that were used to construct the initial RT-profiles. **Part A**: polymerase sequences detected by the first profile scan of the protein data bank with the RT-profiles. **Part B**: polymerase sequences detected by the second profile scan with Plus-profiles. Numbers following the abrreviations indicate the sequence position of the first compared amino acid according to the NBRF database. The length separating the motifs are indicated. Gaps (points) were introduced to increase similarity. The (nearly) invariant amino acids are boxed. Stars indicate the identical or chemically similar amino acids conserved in more than 70% of the sequences, at the top of the alignment for the sequences detected by the first scan, bottom for all the sequences shown. Bold-faced and underlined characters outline inter-group conservation of one particular amino acid.

1A). Their five most conserved regions, roughly centered around the ten invariant amino acids reported by Toh *et al.* (1985), were used to construct several reverse transcriptase profiles (RT-profiles; see Materials and methods). These RT-profiles allowed us to perform a first scan of the entire NBRF protein data bank. A subset of quite disparate sequences were selected in order to reflect the variability occurring in each motif and avoid bias in the RT-profiles.

As expected, the reverse transcriptases present in the protein data bank had generally higher scores than the non-reverse transcriptase sequences, using RT-profiles composed of individual as well as concatenated motifs. Strikingly, reverse transcriptases were immediately followed by the RNA-dependent RNA polymerases of plus-strand RNA viruses, particularly those describing the polio-like group, which were detected at the meaningful score level of approximately two standard deviations (SD) above the mean. The number and alignment scores of these polymerases increased when the motifs were concatenated, noticeably with various combinations of the central motifs (B, C and D). It is noteworthy that the matched region within the plus-strand RNA polymerases includes the highly conserved 'polymerase site' as defined by a GDD consensus sequence embedded in hydrophobic residues (motif C) and preceded 21 to 52 amino acids upstream by the consensus sequence (S/T)GxxxTxxxN(S/T) (motif B) (Franssen *et al.*, 1984; Kamer and Argos, 1984; Domier *et al.*, 1987; Zimmern, 1987; Morch *et al.*, 1988). Upstream this 'polymerase site', a strongly conserved region was noted in some plus-strand RNA viruses by Kamer and Argos (1984); motif A is embedded in this conserved region. The similarities reported here link the RNA-dependent RNA and DNA polymerases by involving an overall domain which encompasses four (A to D) out of the five initial RT-motifs conserved in the same linear arrangement and separated by comparable distances (see Figure 1A). The RT-motif E sometimes matched the most C-terminal highly conserved region of the plus- and double-strand RNA virus polymerases (Gorbalenya and Koonin, 1988). However, these similarities were distant and did not result in any additional strictly invariant residue.

The ubiquity of the four motifs is further enhanced through their characterization in the recently reported polymerase sequence of a double-stranded RNA virus (BTV; Roy *et al.*, 1988) confirming the striking relatedness existing between the plus- and double-stranded RNA viruses (Gorbalenya and Koonin, 1988). Analyzing the sequences detected through this first scan, we introduced four gaps to increase the similarities: one in position 13 of motif A and position 16 of motif B of the RNA polymerases, and one in position 6 of motif B and position 7 of motif D of the DNA polymerases (Figure 1A). The resulting alignment includes five residues almost strictly conserved in all sequences and 21 residues that are conservatively maintained in more than 70% of the sequences. Within the five invariant residues, there is a strong predominance of charged amino acids (3 Asp and 1 Lys). The highest degree of similarity is observed for motif C with two invariant Asp residues.

Generally, the gaps, the sequence similarities and the distances between the motifs were mostly characteristic of a polymerase type. The importance of the inter-motif distances, as seen from global inspection of Figure 1, should be considered with caution. Indeed, we observe that closely related sequences belonging to a same polymerase group can widely vary in their inter-motif distances (e.g. 17–71 residues between motifs A–B in DNA viruses, and 9–41 residues between motifs C–D in the sindbis-like group). The bacteriophages (GaV, MS2V and QbetaV) and the line-1 elements (L1S1, L1Hu and L1Md) are noteworthy since they do not share the characteristic gap distribution of their respective polymerase type. Indeed, the gap in position 13 of motif A, present in all other RNA polymerases, is absent in the bacteriophage sequences while their motif D exhibit a gap typical of the reverse transcriptase sequences. On the other hand, line-1 sequences lack the gap in position 6 of motif B present in all other reverse transcriptases. A closer relationship between the line-like group and the polio-like group can be observed by the conservation of additional single residues (position 15 in motif A, postion 7, 9 and 20 in motif B and position 3 in motif D; bold and underlined in Figure 1A). The L1Hu and HRV14 polymerases were of special interest since they exhibit significant sequence similarities in the whole sequence from the first to the fourth motif (Figure 2). The proposed alignment scores 6.26 standard deviations (SD) and this alignment score slightly decreases to 4 SD, when the regions compared were eventually made larger up to 400 amino acids. Following the studies of Barker and Dayhoff (1972), such high alignment scores (more than 4 SD) might reflect an ancient common evolutionary origin of these proteins or of a portion of them.

Additional scans

A second scan of the protein data bank was carried out with Plus-profiles constructed by a disparate subset of the aligned sequences of double- and plus-strand RNA viruses. In agreement with the first scan analysis, the Plus-profiles detected at an interesting level (over 2 SD) some reverse transcriptase sequences, noticeably the line-1 sequences. Furthermore, they detected other sequences, in particular the RNA-dependent polymerases of minus-strand RNA viruses available in the NBRF protein data bank, i.e. PB1 proteins of the influenza A and B viruses (segmented genome) and L proteins of the vesicular stomatitis virus and of the sendai virus (unsegmented genome). The matched regions correspond to those of highest homology when polymerases of the unsegmented or segmented group are compared separately (Tordo *et al.*, 1988, Kemdirim *et al.*, 1986). Figure 1B presents the alignments of the detected polymerase sequences of minus-strand RNA viruses, including those absent from the protein data bank.

Within the four motifs, four of the five strictly invariant amino acids detected by the RT-profiles are maintained and 18 amino acids of similar chemical nature are conserved in more than 70% of the sequences. The conservative change of the second invariant Asp (motif C) to Asn previously observed in the putative reverse transcriptase of *Chlamydomonas reinhardtii* (RTChla) appears as a classical feature in all the polymerase sequences of the unsegmented minus-strand RNA viruses. The RNA polymerases seem more related to one another than to the DNA polymerases. Consistent with this notion are: (i) the detection of the minus-strand RNA viruses by the Plus-profiles, but not by the RT-profiles; (ii) the identical or chemically similar residues shared by members of minus-strand RNA viruses and plus-strand RNA viruses, especially the sindbis-like group (bold and underlined in Figure 1B). This means that, even though RNA and DNA polymerases are clearly related by the conservation of the four motifs, each class of enzyme seems to have developed typical structural features which may be relevant for their distinct catalytic activities.

A third scan of the protein data bank with profiles deduced from all the above sequences did not lead to any notable additional detection. During the different profile scans, some other polymerase sequences (DNA primases, DNA-

Table I. List of the viruses or retrotransposons discussed in Figure 1 and original references for their polymerase sequences

Virus or element name	Abbreviations	Original References	
Hepatitis B human	HepB	Galibert *et al.* (1979)	[1,2]
Woodchuck hepatitis B	HepWo	Galibert *et al.* (1982)	[1,2]
Duck hepatitis B	HepBDu	Mandart *et al.* (1984)	[1,2]
Human endogenous retrovirus C	HERVC	Repaske *et al.* (1985)	[1]
AKV murine leukaemia	AKVMLV	Herr (1984)	
Murine Moloney leukaemia	MoMLV	Shinnick *et al.* (1981)	[1,2]
Hamster intracisternal A particle	IAPH18	Ono *et al.* (1985)	[1]
Rous sarcoma	RSV	Schwartz *et al.* (1983)	[1,2]
Simian Mazon-Pfizer	SMPV	Sonigo *et al.* (1986)	
Murine mammary tumor	MMTV	Moore *et al.* 1987)	[1,5]
Human endogeneous retrovirus K	HERVK	Ono *et al.* (1986)	[1]
Human adult T-cell leukaemia	ATLV	Seiki *et al.* (1983)	
Human T-cell leukaemia type II	HTLVII	Shimotohno *et al.* (1985)	[1,3]
Bovine leukaemia	BLV	Sagata *et al.* (1985)	[1,5]
Human immunodeficiency type 2	HIV2	Guyader *et al.* (1987)	[1]
Caprine arthritis-encephalitis	CAEV	Chiu *et al.* (1985)	[5]
Equine infectious anemia	EIAV	Stephens *et al.* (1986)	[1]
Visna	Visna	Sonigo *et al.* (1985)	[1]
Human immunodeficiency type 1	HIV1	Wain-Hobson *et al.* (1985)	[1]
Drosophila 17.6 element	17.6	Saigo *et al.* (1984)	[1,2]
Drosophila 297 element	297	Inouye *et al.* (1986)	[1,4]
Drosophila gypsy element	Gypsy	Marlor *et al.* (1986)	[1,4]
Drosophila 412 element	412	Yuki *et al.* (1986)	[1,4]
Cauliflower mosaic	CaMV	Franck *et al.* (1980)	[1,2]
Dictyostelium DIRS-1 element	DIRS	Cappello *et al.* (1985)	[1,4]
Ty912 element	TY912	Clare and Farabaugh (1985)	[1,4]
Drosophila 1731 element	1731	Fourcade-Peronnet *et al.* (1988)	
Drosophila copia element	Copia	Mount and Rubin (1985)	[1,4]
Mauriceville plasmid (mtDNA)	MauP	Nargang *et al.* (1984)	[3]
Chlamydomonas intron (mtDNA)	RTChla	Boer and Gray (1988)	
Trypanosoma ingi element	Ingi	Kimmel *et al.* (1987)	[1,6]
Drosophila f-factor	Ffac	Di Nocera and Casari (1987)	[1]
Maize Cin4 element	Cin4	Schwarz-Sommer *et al.* (1987)	
Drosophila I-factor	Ifac	Fawcett *et al.* (1986)	[1,6]
Yeast class I introns (mtDNA)	Intsp	Lang *et al.* (1985)	[1,3]
Yeast class II introns (mtDNA)	Int31,Int32	Bonitz *et al.* (1980)	[1,3]
Mouse line-1 element	L1Md	Loeb *et al.* (1986)	[1,6]
Prosimian, hum. line-1 elements	L1Sl, L1Hu	Hattori *et al.* (1986)	[1,5]

Virus or element name	Abbreviations	Original References	
Bacteriophage MS2	MS2V	Fiers *et al.* (1976)	[7]
Bacteriophage Ga	GaV	Inokuchi *et al.* (1986)	[7]
Bacteriophage Q-Beta	QBetaV	Inokuchi *et al.* (1988)	
Poliovirus	PolV	Racaniello and Baltimore (1981)	[7,8]
Coxsackievirus	CoxV	Stalhandske *et al.* (1984)	[7]
Human rhinovirus type 14	HRV14	Callahan *et al.* (1985)	[7]
Human rhinovirus type 2	HRV2	Skern *et al.* (1984)	[7]
Encephalomyocarditis	BMCV	Palmenberg *et al.* (1984)	[7,8]
Foot-and-mouth disease	FMDV	Carroll *et al.* (1984)	[7,8]
Hepatitis A	HAV	Najarian *et al.* (1985)	[7]
Cowpea mosaic	CPMV	Lomonosoff and Shanks (1983)	[7,8]
Black beatle	BBV	Dasmahapatra *et al.* (1985)	[7]
Tobacco etch	TEV	Allison *et al.* (1986)	[7]
Tobacco vein mottle	TVMV	Domier *et al.* (1986)	[7]
Theiler's murine encephalomyel.	TMEV	Ozden *et al.* (1986)	
Sindbis, Middleburg	SinV,MidV	Strauss *et al.* (1984)	[7,8]
Semliki forest	SFV	Takkinen (1986)	
Tobacco mosaic	TMV	Goelet *et al.* (1982)	[7,8]
Beet necrotic yellow vein	BNYVV	Bouzoubaa *et al.* (1986)	
Brome mosaic	BMV	Ahlquist *et al.* (1984)	[7,8]
Tobacco rattle	TRV	Boccara *et al.* (1986)	[7]
Alfalfa mosaic	AaMV	Cornelissen *et al.* (1983)	[7,8]
Cucumber mosaic	CucMV	Resaian *et al.* (1984)	[7]
Turnip yellow mosaic	TYMV	Morch *et al.* (1988)	
Barley yellow dwarf	BYDV	Miller *et al.* (1988)	

Table I (*continued*)

Virus or element name	Abbreviations	Original References	
Carnation mottle	CarMV	Guilley *et al.* (1985)	[7]
Yellow fever	YFV	Rice *et al.* (1985)	[7]
West Nile	WNV	Castle *et al.* (1986)	[7]
Infectious bursal disease	IBDV	Morgan *et al.* (1988)	[9]
Bluetongue	BTV	Roy *et al.* (1988)	
Influenza A, B	InfA, InfB	Kemdirim *et al.* (1986)	
Tacaribe	TacaV	Iapalucci *et al.* (1989)	
Lymphocytic choriomeningitis	LCMV	Salvato *et al.* (1989)	
Newcastle disease	NDV	Yusoff *et al.* (1987)	[10]
Sendai	SendV	Shioda *et al.* (1986)	[10]
Measles	MeasV	Blumberg *et al.* (1988)	[10]
Rabies	RabV	Tordo *et al.* (1988)	[10]
Vesicular stomatitis	VSV	Schubert *et al.* (1984)	[10]

The number in brackets indicates the article in which alignments of larger conserved regions are available for, (i) reverse transcriptases: Doolittle *et al.*, 1989 [1]; Toh *et al.*, 1985 [2]; Michel and Lang, 1985 [3]; Stucka *et al.*, 1986 [4]; Hattorie *et al.*, 1986 [5]; Schwarz-Sommer *et al.*, 1987 [6]; (ii) polymerases of plus- and double-strand RNA viruses: Koonin *et al.*, 1987 [7], Kamer and Argos, 1984 [8]; Gorbalenya and Koonin, 1988 [9]; (iii) polymerases of minus-strand RNA viruses: Tordo *et al.*, 1988 [10]. mtDNA: DNA from mitochondrial origin.

```
HRV14   MDGHLMAFDYSN-FDASLSPVWFVCLEKVLTKLGFAGSSLIQSICNTHHIFRDEIY--------------
         :  ::::|    ||    |:    : | | |||: |: |        :|:|  ||
LIHU    TNHMIISIDAEKAFDKIQQPF----MLKPLNKLGIDGTYL--------KIIRA-IYDKPTANIILNGQKL

HRV14   ---VVEGGMPSGCSGTSIFNSMINNIIIRTLILD-AYKGIDLDK--LKILAYGDDLIVSYPYELDPQVLA
            : :|   ||: ::::  :: ::: |::  :  :|||:| |  :|:  ::||:||   |  :| | |
LIHU    EAPPLKTGTRQGCPLSPLLPNIVLEVLARAIRQEKEIKGIQLGKEEVKLSLFADDMIV---YLENPIVSA

HRV14   --------TLGKNYGLTITPPDKSETF
                 ::|  |: |   :||::|
LIHU    QNLLKLISNFSKVSGYKI-NVQKSQAF
```

Fig. 2. Comparison of the entire sequences encompassing the five motifs between HRV14 (positions 1944–2099) and L1Hu (positions 591–775) polymerases. The sequences corresponding to the motifs are underlined.

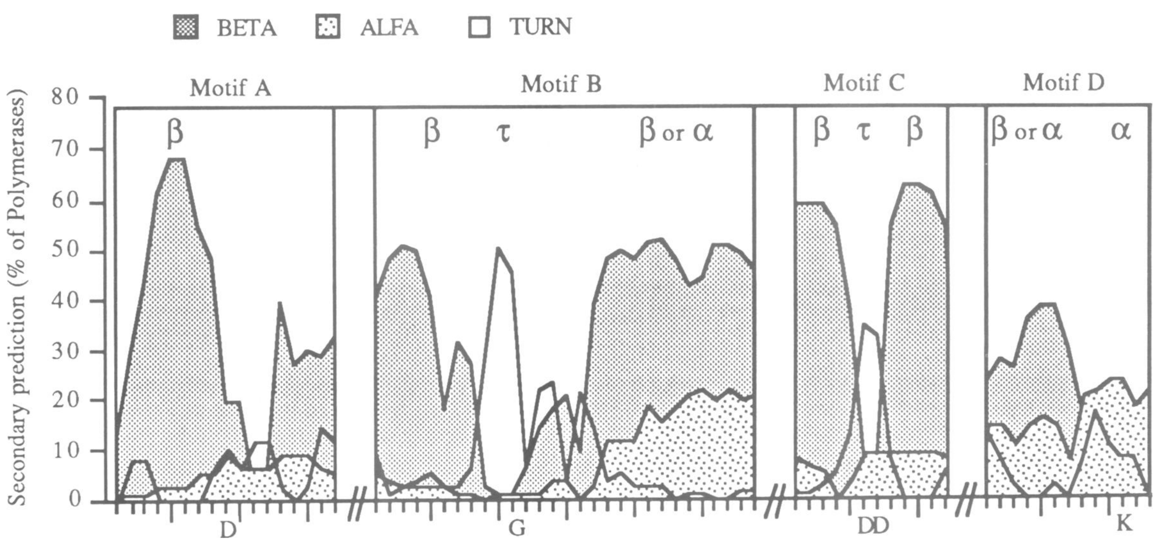

Fig. 3. Percentage of polymerases with an α, β or turn predicted secondary structure. The three curves of the different percentages are superimposed. Amino acid position is indicated by a graduation in abscisse and the five invariant amino acids are mentioned. The secondary structure more frequently predicted within a region of a motif is indicated in the top of the square by a β (beta strand), α (alpha helix) and τ (turn structure).

dependent DNA and RNA polymerases, terminal transferases) were detected, but they generally lacked some of the motifs, or displayed inconsistent inter-motif distances. The significance of this finding is currently being investigated by other methods (i.e. not only sequence comparisons).

Secondary structure predictions

Secondary structure predictions of the aligned polymerase sequences were performed as described in Materials and methods. Essentially, only strong predictions were taken into account. Although such predictions are of limited reliability,

their concordance over numerous sequences may lend more credence (Figure 3). Motif C contains a clear $\beta-$turn$-\beta$ structure, while the beginning of motif A is clearly a β strand. In Figure 3, the end of motif A seems to be predicted as a β structure. However, taken into account both strong and weak predictions (data not shown), it appears that it is not possible to decide if it is an α helix or a β strand. The carboxy-terminal region (and possibly all) of motif D is an α helix. Analyzing the position of the invariant amino acids within the predicted structures, we noticed that they are frequently located within or near tight turns: the invariant Asp residue of motif A is located near the end of a β strand; the Gly residue of motif B is predicted to be in a turn structure as well as the two Asp of motif C located precisely in the turn of a $\beta-$turn$-\beta$ supersecondary structure.

Discussion

This paper presents evidence for the existence of four highly conserved motifs, involving a large domain of 120–210 amino acids, in all the investigated RNA-dependent polymerases encoded by retroviruses, viral and non-viral retroposons, plus- and minus-strand RNA viruses and by the two known double-strand RNA viruses. The significance of these similarities is mainly highlighted by (i) the presence of four invariant and 18 strongly conservatively maintained amino acids within the 69 residues describing the four motifs; (ii) the conservation of additional single residues between members belonging to different groups (e.g. polio-like and non-viral retroposons; minus-strand RNA viruses and sindbis-like); (iii) the identical linear arrangement of the motifs; (iv) the roughly comparable distances separating each motif. In addition, the four motifs are consistently located in regions of greatest homology in each polymerase group.

Thus, the four motifs are attractive targets for site-directed mutagenesis experiments and their concatenation constitutes a useful tool to specifically identify, in sequences of unknown function, a putative polymerase with RNA template specificity.

Functional and structural implications

It is likely that the high degree of conservation of the aforesaid regions in all the RNA-dependent polymerases reflect their crucial importance for the RNA template recognition and/or polymerase activity. Site-directed mutagenesis experiments have recently been performed on a region of the reverse transcriptase of HIV1 encompassing the four conserved motifs (Larder *et al.*, 1987). Within all the mutated amino acids, just two mutations totally destroyed the polymerase activity. They involve the invariant Asp residue of motif A and the first invariant Asp residue of motif C. Within the other mutated amino acids falling in the sequence shown in Figure 1, drastic loss of activity was observed when Tyr residue (position 4 of motif C) was mutated while the other mutations (Asp and Ala in position 12 and 13 of motif A, respectively and Gly in position 11 of motif C) have limited effects. These site-directed mutagenesis experiments are not exhaustive and in particular do not involve the invariant Gly residue in motif B and Lys residue in motif D. However, the integrity of motif B seems also required since insertion of amino acids at position 3 or 10 induces the loss of reverse transcriptase activity of the HIV1 virus (Hizi *et al.*, 1988, 1989). Considering motif C, its key functional role is further highlighted by mutation experiments within the replicase of a plus-strand RNA virus, the Q Beta bacteriophage, in which substitution of the G of the YGDD sequence by A, S, P, M or V residues totally destroyed the activity (Inokuchi and Hirashima, 1987).

In addition, emphasizing the possible universal nature of this motif, Argos (1988) has proposed that the YGDD sequence may be related to the almost invariant YGDTD sequence present in the DNA-dependent DNA polymerases. In these latter proteins, we noticed that the YGDTD motif is preceded by two additional consensus sequences, Vx*D*xSLYP and NSxY*G* (Bernad *et al.*, 1987), where the invariant D and G residues recall those observed in motifs A and B of the RNA-dependent polymerses. As noted above, these regions were sometimes detected by the profile scans. Work is now in progress to assess whether or not these coincidences are of real significance.

The preservation, in such widely divergent proteins, of concatenated motifs which can encompass up to 210 amino acids, strongly suggests their cooperative implication in a well-defined functional unit after proper folding of the protein. It is possible that some of the β strands predicted in Figure 3 may cooperate in a β sheet. The predicted turns frequently contain the charged invariant residues, noticeably the two invariant Asp in motifs A and C previously shown as critical for polymerase activity. The location of these residues in tight turns may be required for proper orientation of these residues for cation binding, template specificity or the catalytic process. On the other hand, the strictly conserved Gly residue of motif B is likely to bear a structural role.

Evolutionary implications

As a convergent mechanism cannot account for the colinearity along the four consensus regions, it seems reasonable to assume that the sequence similarities reported here may be linked to the existence of a common ancestral genetic element bearing a polymerase function.

In addition, it is likely that these sequence relationships arise from a modular evolution which supposes that viral genomes have been built from different combinations of 'modules' such as genes or parts of genes (Zimmern, 1987). Such combinations of genes, reflected by the different genomic organizations, have previously been observed for the plus-strand RNA viruses (Goldbach and Wellink, 1988) and for the reverse transcriptase encoding elements (Doolittle *et al.*, 1989). In this way, the consensus regions detected here may represent a prerequisite 'polymerase module' which has propagated, by RNA or DNA recombination, in the genetic elements encoding either RNA-dependent RNA polymerase or RNA-dependent DNA polymerase. Therefore, one is led to the question, regardless of some recent horizontal transmission, what could be the most parsimonious evolutionary pathway which can account for the sequence similarities observed?

The sequence similarities, gap distribution and inter-motif distances distinguish RNA from DNA polymerases. Nevertheless, some closer relationships are observed between the non-viral retroposons and the plus-strand RNA viruses and especially, between the line-like group and the polio-like group. Indeed, this last group was detected at a meaningful level of 2 SD by the initial RT-profiles and exhibits additional conserved residues with the sequences of

the line-like group. Such a relatedness is further illustrated by the strong homology observed between the L1Hu and the HRV14 (6,26 SD) which suggests a possible common evolutionary origin (over 4 SD). On the other hand, the three plus-strand RNA bacteriophages share with all the retroid elements a very similar gap distribution within motifs A and D. These data clearly suggest that the putative ancestral element encoding the 'polymerase module' should be located in an intermediate position between the retrotransposons and the plus-strand RNA viruses, namely between RNA and DNA polymerase encoding elements. This leads to the following evolutionary scheme, in which the polymerase function of minus- and double-strand RNA viruses would have emerged from a plus-strand RNA virus (RNA polymerase life cycle) while retroid elements and retroviruses (DNA polymerase life cycle) originated from the retroposons, especially from the line-like elements.

Recently, several authors (Doolittle *et al.*, 1989; Temin, 1989) have suggested that the retroviruses emerged very late in evolution, perhaps after the mammal emergence. They proposed that the retrotransposons, and especially the line-like group members, might constitute the ancestors of all the reverse transcriptase encoding elements. Indeed, reverse transcriptases are thought to have existed before retroviruses, namely before the divergence between prokaryotes and eukaryotes (Temin, 1989). Therefore, the putative common ancestor encoding the original polymerase module is likely to be extremely ancient. In accordance with the hypothesis of a primordial RNA world (reviewed in Wintersberger and Wintersberger, 1988), it is reasonable to postulate, as suggested by Lazcano *et al.* (1988), that this ancestor originally encoded an RNA-dependent RNA polymerase module.

Materials and methods

Sequence data

The amino acid sequences compared were mainly collected from the National Biomedical Research Foundation (NBRF) protein database (release 16.8;) and the PSeqIP data bank (release 5.0; Claverie and Bricault, 1986).

Profile analysis

All programs used come from the UWGCG (University of Wisconsin Genetic Computer Group) software package, release 5.3 (Devereux *et al.*, 1984). Amino acid sequence similarities were detected using the Profile analysis method that allows the scanning of target sequences with 'fuzzy' probes (profiles) deduced from a group of aligned sequences (Gribskov *et al.*, 1987, 1988). A set of the 5 most conserved regions of reverse transcriptase sequences as defined by previous sequence comparisons (Michel and Lang, 1985; Toh *et al.*, 1985; Hattori *et al.*, 1986; Stucka *et al.*, 1986; Yuki *et al.*, 1986; Doolittle *et al.*, 1989) were aligned. A subset of quite disparate sequences (names underlined in Figure 1) were selected in order to reflect the variability occurring in the conserved regions. This results in the defintion of five individual consensus sequence motifs. The profiles were constructed by a single motif or by a concatenation of 2, 3, 4 or 5 motifs in the appropriate linear arrangement. Within a motif, the gap and gap-length penalties were defined as 4.5 and 0.5, respectively, for a position where a gap never appears and as 1 and 0.05, respectively, for the position where a gap linked to a particular polymerase group appears (see Figure 1). Two undetermined residues were introduced between each motif to allow the non-conserved interregions separating each motif to vary without constraint during the profile alignment process. Their values for gap and gap-length penalties were both defined as equal to 0. The profiles were used to scan the entire NBRF protein data bank and the sequences manually entered. The profiles verified the validation properties as defined by Gribskov *et al.* (1988). The alignment obtained were analysed by considering two main criteria: (i) the level of significance of the score; (ii) the location of the matching regions with respect to the extended alignments existing between polymerase sequences of a same group, with special emphasis to the invariant amino acids previously appointed.

The human rhinovirus 14 polymerase (HRV14) and the human line-1 reverse transcriptase (L1Hu) were compared with the program Align based on the Needleman and Wunsch algorithm (1970) with a gap penalty equal to 12.

Secondary structure predictions

The program Peptidestructure of the UWGCG software package based on the Chou and Fasman algorithm (Chou and Fasman, 1978) and the Garnier, Osguthorpe and Robson algortihm (Garnier *et al.*, 1978) were used to predict the secondary structure of the entire polymerase sequences. For each amino acid position of the conserved motifs, the percentage of polymerase with an α, β or turn predicted structure were calculated.

Acknowledgements

We wish to thank Dr P.Argos for advice and discussions concerning the sequence alignments, Dr M.A.McClure, Dr M.T.Franze-Fernandez, Dr M.Salvato and Dr M.Billeter for providing data before publication and for helpful discussions. We are pleased to thank Dr B.M.Blumberg, Dr M.Vincentz, Dr L.Bougueleret and Dr J.L.Prato for pertinent comments on the mansucript. We thank the staff of the IBMC and especially Dr M.Zerbib and A.Mouchaboeuf for constant computer assistance. We are deeply indebted to Professors Dirheimer and Sureau in whose laboratories this work was carried out. This work was supported by grants from the Comité Consultatif des Applications de la Recherche de l'Institut Pasteur (7830) and from the Centre National de la Recherche Scientifique (ATP no. 3682).

References

Ahlquist,P., Dasgupta,R. and Kaesberg,P. (1984) *J. Mol. Biol.*, **172**, 369–383.

Allison,R., Johnston,R.E. and Dougherty,W.G. (1986) *Virology*, **154**, 9–20.

Argos,P. (1988) *Nucleic Acids Res.*, **16**, 9909–9916.

Baker,W.C. and Dayhoff,M.O. (1972) In Dayhoff,M.O. (ed.), *Atlas of protein sequence and structure*. National Biomedical Research Foundation, Georgetown University, Washington, vol. 5, pp. 101–110.

Bernad,A., Zaballos,A., Salas,M. and Blanco,L. (1987) *EMBO J.*, **6**, 4219–4225.

Blumberg,B.M., Crowley,J.C., Silverman,J.I., Menonna,J., Cook,S.D. and Dowling,P.C. (1988) *Virology*, **164**, 487–497.

Boccara,M., Hamilton,W.D.O. and Baulcombe,D.C. (1986) *EMBO J.*, **5**, 223–229.

Boer,P.H. and Gray,M.W. (1988) *EMBO J.*, **7**, 3501–3508.

Bonitz,S.G., Coruzzi,G., Thalenfeld,B.E., Tzagoloff,A. and Macino,G. (1980) *J. Biol. Chem.*, **255**, 11927–11941.

Bouzoubaa,S., Ziegler,V., Beck,D., Guilley,H., Richards,K. and Jonard,G. (1986) *J. Gen. Virol.*, **67**, 1689–1700.

Callahan,P.L., Mizutani,S. and Colonno,R.J. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 732–736.

Cappello,J., Handelsman,K. and Lodish,H.F. (1985) *Cell*, **43**, 105–115.

Carroll,A.R., Rowlands,D.J. and Clarke,B.E. (1984) *Nucleic Acids Res.*, **12**, 2461–2472.

Castle,E., Leidner,U., Nowak,T., Wengler,G. and Wengler,G. (1986) *Virology*, **149**, 10–26.

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.

Chou,P.Y. and Fasman,G.D. (1978) *Adv. in Enzymol.*, **47**, 45–148.

Clare,J. and Farabaugh,P. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 2829–2833.

Claverie,J.M. and Bricault,L. (1986) *Proteins*, **1**, 60–65.

Cornelissen,B., Brederode,E., Veeneman,G., van Boom,J. and Bol.J. (1983) *Nuelcic Acids Res.*, **11**, 3019–3025.

Damahapatra,B., Dasgupta,R., Ghosh,A. and Kaesberg,P. (1985) *J. Mol. Biol.*, **182**, 183–189.

Devereux,J., Haeberli,P. and Smithies,O. (1984) *Nucleic Acids Res.*, **12**, 387–395.

Di Nocera,P.P. and Casari,G. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 5843–5847.

Domier,L.L., Franklin,K.M., Shahabuddin,M., Hellmann,G.M., Overmeyer,J.H., Hiremath,S.T., Siaw,M.F.E., Lomonosoff,G.P., Shaw,J.G. and Rhoads,R.E. (1986) *Nucleic Acids Res.*, **14**, 5417–5430.

Domier,L.L., Shaw,J.G. and Rhoads,R.E. (1987) *Virology*, **158**, 20–27.
Doolittle,R.F., Feng,D.F., Johnson,M.S. and McClure,M.A. (1989) *Q Rev. Biol.*, **64**, 1–30.
Fawcett,D.H., Lister,C.K., Kellet,E. and Finnegan,D.J. (1986) *Cell*, **47**, 1007–1015.
Fiers,W., Contreras,R., Duerinck,F., Haegeman,G., Iserentant,D., Merregaert,J., Min Jou,W., Molemans,F., Raeymaekers,A., Van den Berghe,A., Volckaert,G. and Ysebaert,M. (1976) *Nature*, **260**, 500–507.
Fourcade-Peronnet,F., D'Auriol,L., Becker,J., Galibert,F. and Best-Belpomme,M. (1988) *Nucleic Acids Res.*, **16**, 6113–6125.
Franck,A., Guilley,H., Jonard,G., Richards,K. and Hirth,L. (1980) *Cell*, **17**, 285–294.
Franssen,H., Leunissen,J., Goldbach,R., Lomonosoff,G. and Zimmern,D. (1984) *EMBO J.*, **3**, 855–861.
Galibert,F., Mandart,E., Fitoussi,F., Tiollais,P. and Charnay,P. (1979) *Nature*, **281**, 646–650.
Galibert,F., Chen,T.N. and Mandart,E. (1982) *J. Virol.*, **41**, 51–65.
Garnier,J., Osguthorpe,D.J. and Robson,B. (1978) *J. Mol. Biol.*, **120**, 97–120.
Goelet,P., Lomonosoff,G.P., Butler,P.J.G., Akam,M.E., Gait,M.J. and Karn,J. (1982) *Proc. Natl. Acad. Sci., USA*, **79**, 5818–5822.
Golbach,R. and Wellink,J. (1988) *Intervirology*, **29**, 260–267.
Gorbalenya,A.E. and Koonin,E.V. (1988) *Nucleic Acids Res.*, **15**, 7735.
Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 4355–4358.
Gribskov,M., Homyak,M., Edenfield,J. and Eisenberg,D. (1988) *CABIOS*, **4**, 61–66.
Guilley,H., Carrington,J.C., Balàzs,E., Jonard,G., Richards,K. and Morris,T.J. (1985) *Nucleic Acids Res.*, **13**, 6663–6677.
Guyader,M., Emerman,M., Sonigo,P., Clavel,F., Montagnier,L. and Alizon,M. (1987) *Nature*, **326**, 662–669.
Hattori,M., Kuhara,S., Takenaka,O. and Sakaki,Y. (1986) *Nature*, **321**, 625–628.
Herr,W. (1984) *J. Virol.*, **49**, 471–478.
Hizi,A., McGill,C. and Hughes,S.H. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 1218–1222.
Hizi,A., Barber,A. and Hughes,S.H. (1989) *Virology*, **170**, 378–384.
Iapalucci,S., Lopez,R., Rey,O., Lopez,N., Franze-Fernandez,M.T., Cohen,G.N., Lucero,M., Ochoa,A. and Zakin,M.M. (1989) *Virology*, **170**, 40–47.
Inokuchi,Y., Takahashi,R., Hirose,T., Inayama,S., Jacobson,A.B. and Hirashima,A. (1986) *J. Biochem.*, **99**, 1169–1180.
Inokuchi,Y. and Hirashima,A. (1987) *J. Virol.*, **61**, 3946–3949.
Inokuchi,Y., Jacobson,A.B., Hirose,T., Inayama,S. and Hirashima,A. (1988) *Nucelic Acids Res.*, **16**, 6205–6221.
Inouye,S., Yuki,S. and Saigo,K. (1986) *Eur. J. Biochem.*, **154**, 447–454.
Ishihama,A. and Nagata,K. (1988) CRC Critical Rev. Biochem., **23**, 27–76.
Kamer,G. and Argos,P. (1984) *Nucleic Acids Res.*, **12**, 7269–7282.
Kemdirim,S., Palefsky,J. and Briedis,D.J. (1986) *Virology*, **152**, 126–135.
Kimmel,B.E., Ole-Moiyoi,O.K. and Young,J.R. (1987) *Mol. Cell. Biol.*, **7**, 1465–1475.
Koonin,E.V., Gorbalenya,A.E., Chamakov,K.M., Donchenko,A.P. and Blinov,V.M. (1987) *Molek. Genetika*, **7**, 27–39.
Lang,B.F., Ahne,F. and Bonen,L. (1985) *J. Mol. Biol.*, **184**, 353–366.
Larder,B.A., Purifoy,D.J.M., Powell,K.L. and Darby,G. (1987) *Nature*, **327**, 716–717.
Lazcano,A., Fastag,J., Gariglio,P., Ramirez,C. and Oro,J. (1988) *J. Mol. Evol.*, **27**, 365–376.
Loeb,D.D., Padgett,R.W., Hardies,S.C., Shehee,W.R., Comer,M.B., Edgell,M.H. and Hutchinson III,C.A. (1986) *Mol. Cell. Biol.*, **6**, 168–182.
Lomonosoff,G. and Shanks,M. (1983) *EMBO J.*, **2**, 2253–2258.
Mandart,E., Kay,A. and Galibert,F. (1984) *J. Virol.*, **49**, 782–792.
Marlor,R.L., Parkhurst,S.M. and Coeces,V.G. (1986) *Mol. Cell. Biol.*, **6**, 1129–1134.
Michel,F. and Lang,B.F. (1985) *Nature*, **316**, 641–643.
Miller,W.A., Waterhouse,P.M. and Gerlach,W.L. (1988) *Nucleic Acids Res.*, **16**, 6097–6111.
Moore,R., Dixon,M., Smith,R., Peters,G. and Dickson,C. (1987) *J. Virol.*, **61**, 480–490.
Morch,M.D., Boyer,J.C. and Haenni,A.L. (1988) *Nucleic Acids Res.*, **16**, 6157–6173.
Morgan,M.M., Macreadie,I.G., Harpley,V.R., Hudson,P.J. and Azad,A.A. (1988) *Virology*, **163**, 240–242.
Mount,S.M. and Rubin,G.M. (1985) *Mol. Cell. Biol.*, **5**, 1630–1638.
Najarian,R., Caput,D., Gee,W., Potter,S.J., Renard,A., Merryweather,J., Van Nest,G. and Dina,D. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 2627–2631.
Nargang,F.E., Bell,J.B., Stohl,L.L. and Lambowitz,A.M. (1984) *Cell*, **38**, 441–453.
Needleman,S.B. and Wunsch,C.D. (1970) *J. Mol. Biol.*, **48**, 443–453.
Ono,M.H., Toh,T., Miyata,T. and Awaya,T. (1985) *J. Virol.*, **55**, 387–394.
Ono,M.H., Yasunaga,T., Miyata,T. and Ushikubo,H. (1986) *J. Virol.*, **60**, 589–598.
Ozden,S., Tangy,F., Chamorro,M. and Brahic,M. (1986) *J. Virol.*, **60**, 1163–1165.
Palmenberg,A.C., Kirby,E.M., Janda,M.R., Drake,N.L., Duke,G.M., Potratz,K.F. and Collett,M.S. (1984) *Nucleic Acids Res.*, **12**, 2969–2996.
Racaniello,V.R. and Baltimore,D. (1981) *Proc. Natl. Acad. Sci. USA*, **78**, 4887–4891.
Repaske,R., Steele,P.E., O'Neill,R.R., Rabson,A.B. and Martin,M.A. (1985) *J. Virol.*, **54**, 764–772.
Rezaian,A., Williams,R.H.V., Gordon,K.H.J., Gould,A.R. and Symonds,R.H. (1984) *J. Biochem.*, **143**, 277–284.
Rice,C.M., Lenches,E.M., Eddy,S.R., Jung Shin,S., Sheets,R.L. and Strauss,J.H. (1985) *Science*, **229**, 726–733.
Roy,P., Fukusho,A., Ritter,G.D. and Lyon,D. (1988) *Nucleic Acids Res.*, **24**, 11759–11767.
Sagata,N., Yasunaga,T., Tsuzuku-Kawamura,J., Oshihi,K., Ogawa,Y. and Ikawa,Y. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 677–681.
Saigo,K., Kugimiya,W., Matsuo,Y., Inouye,S., Yoshioka,K. and Yuki,S. (1984) *Nature*, **312**, 659–661.
Salvato,M., Shimomaye,E. and Oldstone,M.B.A. (1989) *Virology*, **169**, 377–384.
Schubert,M., Harmison,G.G. and Meier,E. (1984) *J. Virol.*, **51**, 505–514.
Schwartz,D.E., Tizard,R. and Gilbert, (1983) *Cell*, **52**, 631–633.
Schwarz-Sommer,Z., Leclercq,L., Göbel,E. and Saedler,H. (1987) *EMBO J.*, **6**, 3873–3880.
Seiki,M., Hattori,S., Hirayama,Y. and Yoshida,M. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 3618–3622.
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.
Shinnick,T.M., Lerner,R.A. and Sutcliffe,J.G. (1981) *Nature*, **293**, 543–548.
Shioda,T., Iwaski,K. and Shibuta,H. (1986) *Nucleic Acids Res.*, **14**, 1545–1563.
Skern,T., Sommergruber,W., Blass,D., Pieler,C. and Kuechler,E. (1984) *Virology*, **136**, 125–132.
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.
Sonigo,P., Barker,C., Hunter,E. and Wain-Hobson,S. (1986) *Cell*, **45**, 375–385.
Stalhandske,P.O.K., Lindberg,M. and Pettersson,U. (1984) *J. Virol.*, **51**, 742–746.
Stephens,R.M., Casey,J.W. and Rice,N.R. (1986) *Sceince*, **231**, 589–594.
Strauss,E.G., Rice,C.M. and Strauss,J.H. (1984) *Proc. Natl. Acad. Sci. USA*, **80**, 5271–5275.
Stucka,R., Hauber,J. and Feldmann,H. (1986) *Curr. Genet.*, **11**, 193–200.
Takkinen,K. (1986) *Nucleic Acids Res.*, **14**, 5667–5682.
Temin,H.M. (1989) *Nature*, **339**, 254–255.
Toh,H., Kikuno,R., Hayashida,H., Miyata,T., Kugimiya,W., Inouye,S., Yuki,S. and Saigo,K. (1985) *EMBO J.*, **4**, 1267–1272.
Tordo,N., Poch,O., Ermine,A., Keith,G. and Rougeon,F. (1988) *Virology*, **165**, 565–576.
Wain-Hobson,S., Sonigo,P., Danos,O., Cole,S. and Alizon,M. (1985) *Cell*,**40**, 9–17.
Weiner,A.M., Deininger,P.L. and Efstratiadis,A. (1986) *Annu. Rev. Biochem.*, **55**, 631–661.
Wintersberger,U. and Wintersberger,E. (1988) *Trends Genet.*, **3**, 198–202.
Yuki,S., Inouye,S., Ishimaru,S. and Saigo,K. (1986) *Eur. J. Biochem.*, **158**, 403–410.
Yusoff,K., Millar,N.S., Chambers,P. and Emmerson,P.T. (1987) *Nucleic Acids Res.*, **15**, 3961–3976.
Zimmern,D. (1987) In Holland,J., Domingo,E. and Ahlquist,P. (eds) *RNA Genetics*. CRC Press, Boca Raton, Florida.

Received on July 7, 1989; revised on August 5, 1989