REVIEW ARTICLE

Genome Organization and Expression of Reverse Transcribing Elements: Variations and a Theme

By ROGER HULL* AND SIMON N. COVEY
Department of Virus Research, John Innes Institute, Colney Lane, Norwich NR4 7UH, U.K.

INTRODUCTION

Until recently, reverse transcription was considered to be the sole prerogative of retroviruses. Over the last 4 or 5 years it has been recognized that members of two other virus groups, the hepadnaviruses and the caulimoviruses, undergo reverse transcription during replication. Furthermore, some vertebrate genetic elements, e.g. intracisternal A particle (IAP) genes (Ono et al., 1985), VL30 genes (Hodgson et al., 1983) and L1Md (Loeb et al., 1986), and transposable elements from other taxonomic groups [yeast Ty elements (Clare & Farabaugh, 1985; Hauber et al., 1985), Drosophila copia (Mount & Rubin, 1985) and copia-like elements (Saigo et al., 1984), Dictyostelium DIRS-1 element (Cappello et al., 1985), maize Bs1 element (Johns et al., 1985), and possibly maize Cin1 element (Shepherd et al., 1984)] have been shown to possess structural similarities to integrated retroviruses. The yeast Ty element transcript has recently been found to be contained within virus-like particles which have reverse transcriptase activity (Garfinkel et al., 1985; Mellor et al., 1985a). Details of the molecular biology of most of these viruses and some of the elements have been reviewed recently (retroviruses: Weiss et al., 1982, 1985; hepadnaviruses: Tiollais et al., 1985; caulimoviruses: Hohn et al., 1985; Covey & Hull, 1985; Hull & Covey, 1985; all three groups of viruses: Mason et al., 1986; Ty and copia elements: Boeke et al., 1985; Baltimore, 1985; Varmus, 1985). There is relatively little information on the molecular biology of the other elements and we will not consider them further.

At first sight it might be thought that there were few similarities between these viruses and transposable elements from such disparate sources. However, when their individual genetic organization, expression and replication are compared it can be seen that they share certain features which might be taken to indicate an evolutionary relationship. There are also some differences which suggest that each element has adapted in particular ways to the genetic and cellular environment in which it is propagated. It is the intention of this short review to explain these similarities and differences.

Genetic organization

The genetic organization of reverse transcribing elements is best compared on the genome-length RNA transcript (Fig. 1) which is the template for reverse transcription. In each case, this RNA (—) contains all the genetic information of the element; it also has a terminal redundancy (—) which can be from as little as 16 nucleotides in Rous sarcoma virus (RSV) to about 300 nucleotides in human hepatitis B virus (Cattaneo et al., 1984). The terminal repeat is considered to play an important rôle in the formation of negative-strand DNA from the positive-strand RNA during reverse transcription. Retroviruses package this RNA in virions and the equivalent transcripts of copia and Ty elements (Garfinkel et al., 1985; Mellor et al., 1985a) have also been detected in virus-like particles. In contrast, caulimoviruses and hepadnaviruses encapsidate the DNA phase in virions.

In most retroviruses, the 5' proximal open reading frame (ORF) is the gag (group-specific antigen) gene (— in Fig. 1), the protein products of which interact structurally with genomic RNA to package it within the virions. Next comes the pol (polymerase) gene (— in Fig. 1)
Fig. 1. Genome organization of cauliflower mosaic virus (CaMV), ground squirrel hepatitis virus (GSHV), Rous sarcoma virus (RSV), Moloney murine leukaemia virus (MoMLV), human T cell lymphotropic viruses II and III (HTLV-II and HTLV-III), yeast Ty element, Drosophila copia-like element 17.6 and Drosophila copia element. Each is shown on the genome-length RNA (-----), the terminal repeats being indicated by (11---11). The subgenomic mRNA for the gag gene or its equivalent is shown below the genomic RNA; * indicates a splice. The gag (--), pol (-----) and env (---) genes are highlighted. Genes I to VII of CaMV, the b gene of GSHV, the src gene of RSV, the X gene of HTLV-II and the A and B genes of HTLV-III are noted. The extents of the protease (pro), the reverse transcription (RT) and the integrase (int) domains of MoMLV and various features and regions common to all or most genomes are shown: O, Hydrophilic; ●, hydrophobic; p, phosphoprotein; ◦, nucleic acid binding domain; ▲, protease; ◀, reverse transcriptase; ●, integrase. References to the information in this figure are given in the text.
encoding polypeptides with enzymic activities required in reverse transcription and integration. As will be seen later, although the gag and pol genes of many retroviruses are in different reading frames, they are expressed as a polyprotein. Thus, when discussing the function of domains within them they are best considered as a unit. The retrovirus gag-pol polyprotein has the following domains: NH₂-structural polypeptides-protease-reverse transcriptase-endonuclease. In the reverse transcriptase domain of murine leukaemia virus and RSV, the RNase H activity has been shown to be N-terminal relative to the DNA polymerase activity (Levin et al., 1984; Grandgenett et al., 1985).

The hepadnavirus core antigen and caulimovirus coat protein (gene IV product) (Fig. 1) can be considered to be analogous to gag as they are the gene products which interact with the nucleic acid in the particle. The arrangement of gag and pol regions (the ‘gag-pol’ core) is thus also apparent in hepadnaviruses, cauliflower mosaic virus (CaMV) and in retrotransposons (Fig. 1). Organizational and amino acid sequence similarities between the retrovirus pol genes and analogous genes of retrotransposons, CaMV and hepadnaviruses have already been noted (see Toh et al., 1983; Varmus, 1985). These are highlighted in Fig. 1 where the protease (pro, ▽), reverse transcriptase (RT) and endonuclease domains (int) of Moloney murine leukaemia virus (MoMLV) are identified and the regions where amino acid sequence similarities exist between the various elements are indicated. However, there are some differences in the arrangements of the functional domains when retroviruses and the copia-like element 17.6 are compared with copia and Ty 912 (Mount & Rubin, 1985). Thus, in copia and Ty, the reverse transcriptase (▼) and integrase (⊙) domains of the pol region are apparently inverted relative to the other elements (Fig. 1).

As noted above, a general feature of reverse transcribing elements is that polypeptides derived from the gag ORF interact with genomic nucleic acid to give virus or virus-like particles. This similarity extends even to regions within the gag ORF. The protein derived from the C-terminal region of the gag product is rich in basic amino acids (⊙ in Fig. 1) and in retroviruses is known to bind to the genomic RNA. The amino acid sequence CX₇CX₉C, highly conserved in retrovirus nucleic acid binding proteins (Henderson et al., 1981; Copeland et al., 1984), is found in ORF I of copia (Mount & Rubin, 1985) and the coat protein of CaMV (Covey, 1986; M. Pietrzak, J. Fuetterer & T. Hohn, personal communication), although it is not found in the Ty element, the copia-like 17.6 element or in hepadnavirus core antigen. For many retroviruses, one of the proteins derived from the N-terminal portion of the gag polyprotein is phosphorylated (p in Fig. 1) although the phosphoprotein is apparently absent from the gag product of human T cell lymphotropic virus (HTLV-III) now thought to be a lentivirus (see Rabson & Martin, 1985). The gag proteins of hepadnaviruses, CaMV and the Ty element are also phosphorylated (p in Fig. 1). From the data of Franck et al. (1980) and Hahn & Shepherd (1980) it can be deduced that the phosphorylated region in the CaMV coat protein precursor, which undergoes processing, is in the N-terminal half and so is in an analogous position to the phosphoprotein P10–P17 of various retroviruses. The phosphorylated domains of the gag proteins of hepadnaviruses and Ty elements have not yet been identified. The N-terminal part of CaMV coat protein is hydrophilic (⊙ in Fig. 1) in contrast with the hydrophobic N-termini of some retrovirus gag polyproteins (MoMLV, RSV) and of the equivalent gene product (core antigen) of hepadnaviruses (● in Fig. 1). This probably reflects the fact that retroviruses and hepadnaviruses assemble in association with membranes whereas CaMV assembles in viral inclusion bodies.

For copia and Ty elements, all of the coding capacity is taken up by the gag-pol core whilst the other viruses and retrotransposons have additional coding regions. Those viruses that infect vertebrate cells (the retroviruses and hepadnaviruses) have an env gene or surface antigen (● in Fig. 1). This coding region forms one or more glycoprotein(s) which are exposed on the virus capsid surface and interact with the host cell receptors. The third ORF in the copia-like element 17.6 has been suggested to be analogous to the retrovirus env gene (Saigo et al., 1984). In CaMV, the ORF (gene VI) immediately downstream from the pol gene (gene V) codes for a major protein component of inclusion bodies. These are cytoplasmic proteinaceous structures which contain most of the virus particles and are considered to be the sites of reverse transcription and virion assembly (see Maule, 1985; Hull & Covey, 1985). Thus, this gene product envelopes the
CaMV replication function as do the \textit{env} gene products of retroviruses and hepadnaviruses; however, in caulimoviruses it forms a type of syncytium in that the inclusion body protein forms an aggregate containing many virus genomes.

CaMV has four additional ORFs (genes VII, I, II and III) 5' to the \textit{gag-pol} core (Fig. 1). We have recently suggested (Hull & Covey, 1985) that ORF I is involved in cell-to-cell spread of CaMV within susceptible plants. The ORF II product facilitates aphid transmission of the virus from plant to plant (Armour \textit{et al.}, 1983; Woolston \textit{et al.}, 1983; Givord \textit{et al.}, 1984). No function has yet been ascribed to ORFs III and VII although ORF III is expressed as a protein in plants and is indispensable for virus infection (Xiong \textit{et al.}, 1984). ORF VII is dispensable (Dixon & Hohn, 1984) and it is questionable whether it produces a functional protein.

\textit{Gene expression}

Common features of the genome organization and replicative mechanism are reflected in the strategies of gene expression adopted by reverse transcribing elements. The DNA phase is the starting point for gene expression, although this transcription template generates RNA molecules that serve two functions: replication and protein synthesis. There are differences in the location and structure of transcription templates for each group of elements. For all elements that have been studied, transcription occurs in the nucleus and is directed by the RNA polymerase that normally transcribes host protein-encoding genes. It is not surprising, therefore, that nucleotide sequences involved in the control of transcription initiation and termination of host genes are located in the appropriate positions within the DNA phase of reverse transcribing elements.

In most retroviruses and retrotransposons, the transcribed DNA assumes a linear configuration as it is integrated in the host genome. An exception to this amongst the retroviruses is visna virus, a member of the lentivirus group, the majority of the complementary DNA forms of which are linear molecules that remain extrachromosomal in tissue culture cells (Haase \textit{et al.}, 1982; Harris \textit{et al.}, 1984). In contrast, the transcription template of hepadnaviruses is considered to be an extrachromosomal supercoiled DNA (see Mason \textit{et al.}, 1986 for discussion). Integrated forms of the DNA of some hepadnaviruses have been found (see Mason \textit{et al.}, 1986) although these are not thought to play an essential role in the virus multiplication cycle. CaMV DNA also becomes a supercoiled component of an episomal minichromosome (Olszewski \textit{et al.}, 1982) which is the template for transcription. In addition, a complex population of subgenomic forms of CaMV supercoiled DNA have been detected in infected plants (Olszewski & Guilfoyle, 1983; Rollo & Covey, 1985) but their function remains to be determined.

Following transcription, mRNAs move to the cytoplasm for translation. It has been known for some time that, in retroviruses, translation of the \textit{gag} and \textit{pol} regions is linked. The \textit{gag} polypeptides that constitute the virion core are generated via two routes (see Varmus, 1985). First, an RNA, thought to be similar in structure to genomic RNA, is translated to produce a \textit{gag} precursor polyprotein. Additionally, a longer \textit{gag-pol} polyprotein is generated at a level 1 to 10\% that of the \textit{gag} protein alone. Polyprotein processing is thought to be directed by the protease domain (\textit{\sigma} in Fig. 1) of the \textit{gag-pol} polyprotein discussed above. The mechanism by which the \textit{gag-pol} polyprotein is generated has been the subject of speculation since the two ORFs are separated by a stop codon or by being in different reading frames (Fig. 1). For some retroviruses like MoMLV, the two in-frame genes are separated by an amber termination codon which is thought to be suppressed, resulting in \textit{gag-pol} synthesis (Yoshinaka \textit{et al.}, 1985). For many other retroviruses, e.g. RSV and HTLV-III, the \textit{gag} and \textit{pol} ORFs overlap (in a different frame) and in human T cell leukaemia virus type II (HTLV-II), the \textit{gag}, protease and \textit{pol} ORFs overlap in three different reading frames (see Fig. 1); spliced RNAs containing two or three frames fused have never been found. However, recent studies of the yeast Ty retrotransposon have provided evidence that a frameshift mechanism operates at the overlap of the \textit{gag} and \textit{pol} ORFs such that a \textit{gag-pol} polyprotein is synthesized from the two out-of-frame genes (Mellor \textit{et al.}, 1985\textit{b}; Clare & Farabaugh, 1985). Moreover, Jacks & Varmus (1985) have generated an RNA species \textit{in vitro}
by transcription of an SP6 construct that contained the out-of-frame gag and pol ORFs of RSV and showed that, following in vitro translation in a reticulocyte lysate, a fusion protein was produced. This strongly supports the view that frameshifting occurs more widely amongst reverse transcribing elements in the synthesis of the gag-pol polyprotein.

In hepadnaviruses, the first two ORFs in the genome-length transcript, the core antigen ORF (in Fig. 1) and the polymerase ORF (in Fig. 1), analogous to gag and pol, overlap in a manner reminiscent of retroviruses. This perhaps implies that frame shifting might also generate the appropriate polyprotein since spliced RNAs of hepadnaviruses have not been observed. However, a fusion protein has not been detected in infected cells either and the conserved protease domain (V in Fig. 1), found in the gag-pol region of other reverse transcribing elements, which might function in processing a polyprotein, is absent from hepadnaviruses (Varmus, 1985).

For CaMV, the expression of the gag and pol genes (ORFs IV and V) is also not well understood. Unlike most other reverse transcribing elements, these two ORFs are not located close to the 5' end of the major genome length transcript but instead have four ORFs preceding them. However, there is more than just a suggestion that translation of the first four CaMV ORFs (VII, I, II and III; see Fig. 1) is linked. The evidence comes from studies of the effects, upon viral DNA infectivity, of insertions and deletions in two non-essential ORFs (VII and II; see Fig. 1) which have shown that insertion of an AUG codon leads to loss of infectivity if it is not followed by an in-frame termination codon. Furthermore, initiation and termination codons of adjacent ORFs must be located close to one another for infectivity to be retained (Sieg & Gronenborn, 1982; Dixon & Hohn, 1984). These studies have suggested that a 'relay race' mechanism of translation occurs on the CaMV genome-length transcript in which ribosome small subunits do not completely disengage at the end of a given ORF and protein synthesis is re-initiated internally in the RNA (Sieg & Gronenborn, 1982). Because ORFs I to III tandemly overlap one another slightly (see Fig. 1), this makes possible a frameshifting mechanism similar to that demonstrated for Ty and RSV. However, the appropriate polyprotein that should be generated has not been observed in CaMV-infected cells. Moreover, a frameshifting expression mechanism would not explain why each ORF has a 5'-proximal AUG initiation codon adjacent to the termination codon of the preceding ORF.

Deletion and insertion analysis has not been possible for the CaMV gag and pol ORFs (IV and V) because both are essential genes and so it remains an open question as to whether frameshifting, along the lines proposed for retrovirus and yeast Ty gag-pol ORFs, generates the appropriate CaMV polyprotein(s). The existence of a CaMV polyprotein is suggested, however, because its pol gene (ORF V) contains the protease domain (V) conserved in retroviruses and retrotransposons (Fig. 1). The story is still far from clear though for CaMV, since little is yet known concerning the fine structure of mRNAs that express ORFs VII to V, a situation that is further complicated by the suggested existence of a separate subgenomic mRNA for the pol ORF possibly required for specific gene amplification (Plant et al., 1985), and the inference from the observations of Hirochika et al. (1985) that other ORFs might be fused by RNA splicing.

The env ORF of reverse transcribing elements ( in Fig. 1) in general lies downstream of the gag-pol core. In some elements, e.g. CaMV and HTLV-III, the pol and env ORFs are separated. In others, e.g. RSV and MoMLV, env slightly overlaps the end of pol whilst in hepadnaviruses, env (the surface antigen gene: ) completely overlaps pol. These variations aside, the env ORF in each type of element for which data are available is expressed via a subgenomic mRNA that is 3' co-terminal with the genome length RNA (see Fig. 1). The retrovirus env mRNA has the 5' leader of the genomic RNA spliced onto it ( — — in Fig. 1) whilst the equivalent CaMV mRNA is transcribed under the direction of its own promoter (see Covey, 1985). The means by which the subgenomic mRNA of the hepadnavirus ground squirrel hepatitis virus (GSHV) (Enders et al., 1985) is generated is not yet known although, like the CaMV transcript, it is not spliced. Similarly, a separate transcript for the copia-like retrotransposon 17.6 env gene has not been observed.

Thus, allowing for the limitations of information available noted above, it would appear that for the viruses and elements which use reverse transcription in their replication, the gag-pol core
is one functional unit and the env gene is another; the transcripts of these two units appear to be 3’ co-terminal.

**DISCUSSION**

We have drawn attention above to similarities and differences in the genetic organization of reverse transcribing viruses and elements. We would like to propose that these genomes have a gag–pol core to which is added cassettes of genes necessary for their propagation within their own particular host system. For viruses that infect vertebrates, the env gene product is required for cell-to-cell spread and it also encloses the reverse transcription machinery, a function possibly performed by the gene VI product of CaMV and that of the third ORF in the copia-like element 17.6. In CaMV, there are cassettes comprising the gene required for transmission from plant to plant (gene II) and the putative gene for cell-to-cell spread (gene I; see Fig. 1). The Ty and copia elements do not appear to need host-adapting cassettes of genes. Since the reproduction of yeast is either by binary fusion or cell fusion, and because yeast cells have a thick cell wall, there would appear to be no requirement for cell-to-cell spread factors to transmit Ty sequences. There is no such limitation to the cell-to-cell spread of viruses in Drosophila although it is not known if copia particles spread in this manner. Even if they do not, both Ty and copia can be considered as being transmitted from cell to cell during division as integrated elements in the chromosomal DNA or as episomal particles. Fig. 1 shows that some of the viruses also carry other genes, e.g. A and B in HTLV-III (see Rabson & Martin, 1985) and b in GSHV (Seeger et al., 1984); the full significance of these genes to the host adaptation of the virus is not yet understood although it has been suggested that the gene product of the human T cell leukaemia viruses types I and II ORF X has a trans-acting effect upon transcription and is involved in the cell specificity of transformation (Sodroski et al., 1984). Very recently, a trans-activating gene, called tat-III, has been found for HTLV-III. tat-III is not shown in Fig. 1 as it comprises three exons, one non-coding region from the 5’ long terminal repeat, one coding region from between the A and env genes (Fig. 1) and one within the env gene in a different reading frame (Arya et al., 1985; Sodroski et al., 1985). The product of the tat-III gene is an absolute requirement for virus expression and replication in human T lymphocytes (Fisher et al., 1986).

How these cassettes of genes have been collected together and where they come from are, as yet, unanswered questions. In his protovirus hypothesis, Temin (1974) suggested that retroviruses evolved from normal cellular components. An extension of this hypothesis is that the formation of the gag–pol core represents a single evolutionary event. This core has spread across a wide range of taxonomic groups acquiring other cassettes of genes to enable adaptation to specific hosts. It will be interesting to see if cellular relatives of the pol gene of reverse transcribing elements turn out to be on their own, or associated with progenitor gag sequences.

**REFERENCES**


