Sequence and expression of the mouse mammary tumour virus *env* gene

Shelagh M.S. Redmond* and Clive Dickson

Imperial Cancer Research Fund Laboratories, London WC2A 3PX, UK

Communicated by L.V. Crawford
Received on 29 November 1982

We have determined the DNA sequence of the envelope gene region of the GR strain of mouse mammary tumour virus. The sequence extends for 3012 nucleotides from the single *Eco*RI site to beyond the *Pst*I site in the 3′ long terminal repeat (LTR) of the provirus. There is a major open reading frame from nucleotides 752 to 2818 which encompasses the entire *env* gene. This reading frame extends through a polypurine tract and into the LTR. There is another open reading frame from the first nucleotide to position 803, presumably corresponding to the end of the *pol* gene. The splice acceptor site which generates *env* mRNA has been mapped experimentally to nucleotide 750. The *env* gene products, gp52 and gp36, have been positioned on the sequence using the directly determined amino acid sequences of the amino terminus of gp52; and both the amino and carboxyl terminini of gp36. The start of gp52 is preceded by a series of 19 uncharged amino acids which could function as a typical signal sequence, but this sequence is only part of a much longer leader peptide. The tetrad Arg-Ala-Lys-Arg is the presumed cleavage site in the gPr73^{env} precursor, and occurs just before the gp36 amino terminus. There are five potential asparagine-linked glycosylation sites which agrees with previous experimental results. The gp36 has two long hydrophobic regions at its amino and carboxy termini, these are suggested to act as a fusion peptide and the trans-membrane anchor, respectively.

Key words: DNA sequence/envelope gene/mouse mammary tumour virus/virus glycoproteins

Introduction

The enveloped viruses have provided useful systems in which to study the synthesis and maturation of cell surface glycoproteins, and the introduction of gene cloning and DNA sequencing now makes it possible to correlate experimental observations with the predictions from the primary sequence. The envelope glycoprotein complex of a virus must perform several functions, such as the recognition of virus core particles during assembly, binding to cell surface molecules and initiating the fusion of cellular and viral membranes. As a result, they will contribute to a large extent to the tissue tropism, host-range and viral interference properties of viruses. Thus the envelope glycoproteins of mouse mammary tumour virus (MMTV) are of particular interest in view of some of the unusual features which distinguish MMTV from other retroviruses.

MMTV is morphologically and morphogenetically different from the majority of retroviruses (C-type) and is defined as the prototype of the B-type retroviruses (see Weiss *et al.*, 1982). Electron microscopy studies show that B-type particles mature from a distinct intracytoplasmic form, and are characterised by an eccentrically located electron-dense nucleoid surrounded by an envelope with well-defined surface glycoprotein spikes. As in other enveloped viruses, these spikes are composed of two virally-coded glycoproteins. For MMTV these are gp52, which may form the exposed 'knob' and gp36, which is thought to form a supporting trans-membrane 'stalk' (Sarkar *et al.*, 1976). The MMTV envelope glycoproteins may contribute to a unique tissue tropism in that MMTV productively infects, almost exclusively, mammary epithelial cells *in vivo*. This suggests that there could be a highly specific cell surface receptor molecule recognised by MMTV virions. Another interesting feature of the MMTV *env* products is the reported cross-reactivity with antigens associated with human breast carcinomas, especially as there is no cross-reactivity with benign breast tumours or other organ carcinomas (Mesa-Tejada *et al.*, 1979; Ohno *et al.*, 1979; Day *et al.*, 1981).

To gain some understanding of how the MMTV *env* gene may perform its complex functions we have determined its DNA sequence. Using this, and the derived amino sequence, several features concerning the structure and function of this viral glycoprotein have been illuminated.

Results

The DNA sequence

The nucleotide sequence of a 3-kb region from the 3′ half of the MMTV provirus, extending from the single *Eco*RI site to beyond the *Pst*I site in the long terminal repeat (LTR) (see Figure 1) was determined from fragments cloned into bacteriophage M13 and using Sanger's dideoxynucleotide sequencing method. The strategy used to obtain the complete sequence is shown in Figure 1 and the complete sequence is presented in Figure 2.

Examination of the sequence reveals one extensive open reading frame from nucleotide 752 to 2818, which corresponds to the expected position of the envelope gene. The open reading frame (*orf*) contained within the LTR (Donehower *et al.*, 1981; Dickson and Peters, 1981; Fasel *et al.*, 1982) begins at nucleotide 2766 and therefore overlaps with the end of the *env* coding region. Another open reading frame extends from the start of the sequence and terminates at nucleotide 803. This latter open reading frame may be the end of the polymerase gene. Several features deduced from the sequence, especially relevant to the expression of the envelope gene, are described below.

Expression of the envelope gene

Infected cells contain two major MMTV-related mRNAs, one of genomic size (35S), which encodes the *gag* and *gag-pol* precursors and a subgenomic spliced mRNA (24S) which encodes the *env* precursor (Sen *et al.*, 1979; Groner *et al.*, 1979; Dickson and Peters, 1981; Robertson and Varmus, 1981; Dudley and Varmus, 1981). The 24S mRNA is believed to contain 200–300 bases derived from the 5′ end of the provirus which is spliced to a coding region located in the 3′ half of the provirus (Majors, 1982).

*To whom reprint requests should be sent.

From sequencing eukaryotic genes and the mRNAs they produce, a consensus for splicing signals has been recognised (Breathnach and Chambon, 1978; Mount, 1982). The consensus splice acceptor site consists of a pyrimidine-rich sequence followed by the dinucleotide AG such that cleavage at the intron/exon boundary occurs on the 3′ side of the G nucleotide. A computer search of the sequence reveals several possible splice acceptor sites (see Figure 2). One of these, between nucleotides 732 and 750, is very close to the beginning of the presumed *env* coding sequences. To determine whether this potential splice site is authentic, the RNA from viral producing cells and tissues was subjected to S1 mapping (Berk and Sharp, 1979). The DNA probe used was the *Sau*3A fragment from nucleotide 648 to 865, which spans the presumed splice site (see Figure 2). The result of the S1 mapping is shown in Figure 3. Using RNA from uninfected M2 cells, or carrier RNA alone, no protection of the single-stranded DNA probe was seen. Using RNA from MMTV-infected cells (MGR-4 and D55) or a virally induced tumour (W8), two bands were observed. One of 217 nucleotides corresponds to protection of the full length DNA probe, the other of 118 nucleotides corresponds to the distance between the labelled 5′ end of the probe and the splice site. This position maps exactly to the splice point predicted by the nucleotide sequence.

Several other consensus splice regions are present in the sequence (see Figure 2), one of which (nucleotide 2672–2692) may be used in the production of an mRNA encoding the product of the open reading fraqme in the LTR (Wheeler *et al.*, 1983; van Ooyen, personal communication). There are also several minor RNA species seen on Northern blots of RNA fractionated on denaturing agarose gels (unpublished data; Robertson and Varmus, 1981). The relationship between these RNAs and other putative splice points is at present under investigation.

Amino acid sequence of the envelope gene

The deduced amino acid sequence is shown in Figure 4 and, beginning at the first methionine codon after the splice point, reveals a continuous open reading frame between nucleotides 752 and 2818 which would result in a primary translation product of 688 amino acids. Since the amino acid sequences of the amino-terminal end of gp52 (Arthur *et al.*, 1982), and both the amino and carboxyl termini of gp36 (Henderson *et al.*, 1983) have been determined, they can be aligned on the derived sequence as illustrated in Figure 4. For gp52, 41 out of 43 amino acids agree with the predicted sequence, and for gp36, 27 out of 27 agree for the amino terminus and five out of five for the carboxyl terminus. The amino terminus of gp52 appears to be many amino acids away from the potential start of the precursor polyprotein. This had been expected from a comparison of the sizes of the *in vitro* translated precursor and the *in vivo* apoprotein (Dickson and Peters, 1981; Dickson and Atterwill, 1980; Ser *et al.*, 1979; Dudley and Varmus, 1981; Arthur *et al.*, 1982), which suggested the presence of a long leader peptide of ~7000–9000 daltons. From the predicted amino acid sequence there are three potential methionine starts for the precursor, which would give rise to hypothetical leaders of 11 000, 7000 and 5700 daltons, respectively (see Figure 4). Most membrane and secreted proteins have a hydrophobic 'signal sequence' which facilitates trans-membrane synthesis (Blobel and Dobberstein, 1975). The signal sequence is usually appended to the amino-terminal end of protein forming part of a leader peptide. Usually this peptide and the signal sequence are synonymous and at a minimum consists of a string of ~11 hydrophobic amino acids. The potential leader peptides of MMTV are unusually long, (98, 63 or 53 amino acids depending on which methionine start is used), but all would contain the stretch of 19 uncharged amino acids immediately adjacent to the amino terminus of gp52 which could act as a signal peptide (see Figure 4). After transmembrane synthesis has been initiated, the leader peptide is thought to be cleaved from the precursor by a cellular protease, to expose what is to become the amino-terminal end of $gPr73^{env}$. A further proteolytic cleavage would then be necessary to liberate gp52 and gp36 from the intracellular precursor $gPr73^{env}$. The tetrapeptide Arg-Ala-Lys-Arg occurs just before the amino terminus of gp36 suggesting that a trypsin-like protease could function here.

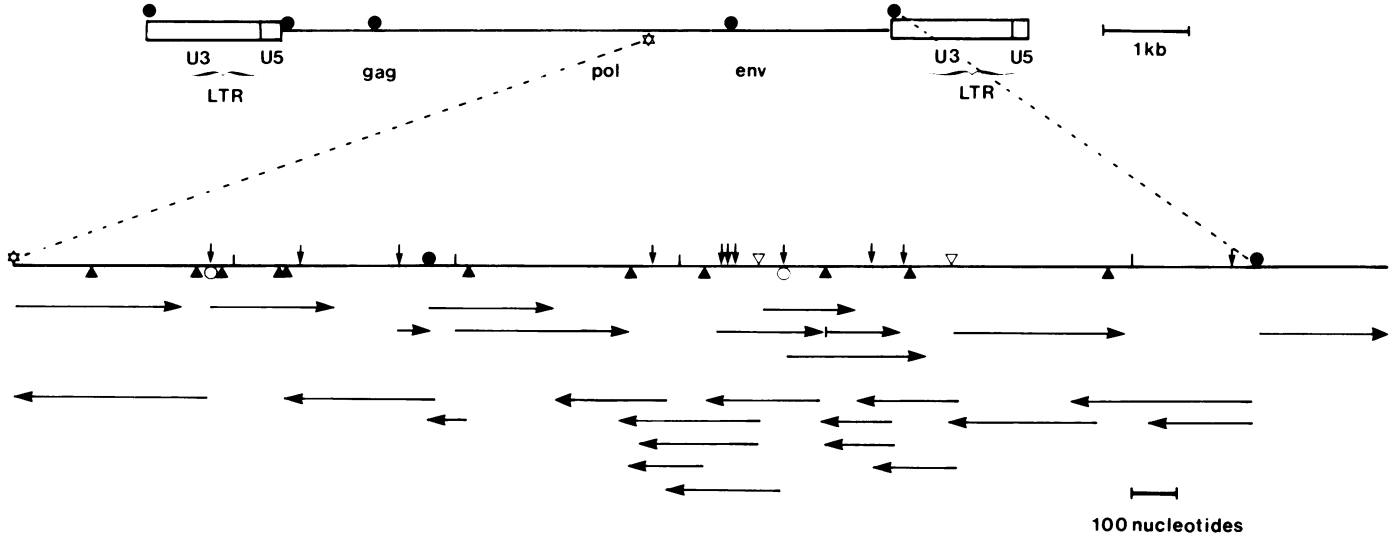

Fig. 1. A restriction map of the *env* region and the strategy used to obtain a complete sequence. A schematic representation of the MMTV provirus showing the approximate position of the structural genes *gag, pol* and *env* and the LTRs composed of sequences derived from the 5′ and 3′ ends of the viral RNA (U5 and U3). The envelope region is expanded below, demonstrating important restriction sites and the strategy of sequencing. The following restriction sites are marked: *Eco*Ri (✡); *Bam*HI (○); *Pst*I (●); *Sau*3A (↓); *Taq*I (▽); and *Hae*III (▲). The horizontal arrows below indicate the length and direction of sequences obtained from different M13 clones.

```
                                                                    50                                                                 100
GAATTCTGAC CGCTTTAGAG TCAGCTCAAG AAAGCCACGC ACTACACCAT CAAAACGCCG CGGCGCTTAG GTTTCAGTTT CACATCACTC GTGAACAAGC

                                                                   150                                                                 200
ACGAGAAATA GTAAAACTGT GTCCAAATTG CCCCGACTGG GGCAGTGCGC CACAACTAGG AGTAAATCCT AGGGGCCTTA AGCCGAGTTC TATGGCAAAT

                                                                   250                                                                 300
GGATGTTACT CATGTCTCAG AATTTGGAAA ATTAAAATAT GTACATGTGA CAGTGGATAC TTATTCTCAT TTTACTTTCC TACCGCCCGA ACGGGCGAAG

                                                                   350                                                                 400
CAACCAAGGA TGTGTTACAA CACTTGGCTC AAAGCTTTGC ATACATGGGC ATTCCTCAAA AAATAAAAAC AGATAATGCC CCTGCATATG TGTCTCGTTC

                                                                   450                                                                 500
AATACAAGAA TTTCTGGCCA GATGGAAAAT ATCTCACGTC ACGGGGATCC CTTACAATCC CCAAGGACAG GCCATTGTTG AACGAACACA CCAAAATATA

                                                                   550                                                                 600
AAGGCACAGC TTAATAAACT TCAAAAGGCT GGAAAATACT ATACACCCCA TCATCTATTG GCACACGCTC TTTTTGTGCT GAATCATGTA AATATGGACA

                                                                   650                                                                 700
ATCAAGGCCA TACAGCGGCC GAAAGACATT GGGGTCCAAT CTCAGCCGAT CCAAAACCTA TGGTCATGTG GAAAGACCTT CTCACAGGGT CCTGGAAAGG

                                                                   750                                                                 800
ACGATGTCCT AATAACAGCC GGACGAGGCT ATGCTTGTGT TTTTCCACAG GATGCCGAAT CACCAATCTG GGTCCCCGAC CGGTTCATCC GACCTTTTAC

                                                                   850                                                                 900
TGAGCGGAAA GAAGCAACGC CCACACCTGG CACTGCGGAG AAAACGCCGC CGCGAGATGA GAAAGATCAA CAGGAAAGTC CGGAGGATGA ATCTAGCCCC

                                                                   950                                                                1000
CATCAAAGAG AAGACGGCTT GGCAACATCT GCAGGCGTTA ATCTCCGAAG CGGAGGAGGT TCTTAAAACC TCACAAACTC CCCAAAACTC TTTGACTTTA

                                                                  1050                                                                1100
TTTCTTGCTT TGTTGTCTGT CCTCGGCCCC CCGCCTGTGA CCGGGGAAAG TTATTGGGCT TACCTACCTA AACCACCTAT TCTCCATCCC GTGGGATGGG

                                                                  1150                                                                1200
GAAGTACAGA CCCCATTAGA GTTCTGACCA ATCAAACCAT GTATTTGGGT GGGTCGCCTG ACTTTCACGG GTTTAGAAAC ATGTCTGGCA ATGTACATTT

                                                                  1250                                                                1300
TGAGGGGAAG TCTGATACGC TCCCCATTTG CTTTTCCTTC TCCTTTTCTA CCCCCACGGG CTGTTTTCAA GTAGATAAGC AAGTATTTCT TTCTGATACA

                                                                  1350                                                                1400
CCCACGGTTG ATAATAATAA ACCTGGGGGA AAGGGTGATA AAAGGCGTAT GTGGGAACTT TGGTTGCATA CTTTGGGGAA CTCAGGGGCC AATACAAAAC

                                                                  1450                                                                1500
TGGTCCCTAT AAAAAAGAAG TTGCCCCCCA AATATCCTCA CTGCCAGATC GCCTTTAAGA AGGACGCCTT CTGGGAGGGA GACGAGTCTG CTCCTCCACG

                                                                  1550                                                                1600
GTGGTTGCCT TGCGCCTTCC CTGACAAGGG GGTGAGTTTT TCTCCAAAAG GGGCCCTTGG GTTACTTTGG GATTTCTCCC TTCCCTCGCC TAGTGTAGAT

                                                                  1650                                                                1700
CAGTCAGATC AGATTAAAAG CAAAAAGGAT CTATTTGGAA ATTATACTCC CCCTGTCAAT AAAGAGGTTC ATCGATGGTA TGAAGCAGGA TGGGTAGAAC

                                                                  1750                                                                1800
CTACATGGTT CTGGGAAAAT TCTCCTAAGG ATCCCAATGA TAGAGATTTT ACTGCTCTAG TTCCCCATAC AGAATTGTTT CGCTTAGTTG CAGCCTCAAG

                                                                  1850                                                                1900
ACATCTTATT CTCAAAAGGC CAGGATTTCA AGAACATGAA ATGATTCCTA CATCTGCCTG TGTTACTTAC CCTTATGCCA TATTATTAGG ATTACCTCAG

                                                                  1950                                                                2000
CTAATAGATA TAGAGAAAAG AGGATCTACT TTTCATATTT CCTGTTCTTC TTGTAGATTG ACTAATTGTT TAGATTCTTC TGCCTACGAC TATGCAGCGA

                                                                  2050                                                                2100
TCATAGTCAA GAGGCCGCCA TACGTGCTGC TACCTGTAGA TATTGGTGAT GAACCATGGT TTGATGATTC TGCCATTCAA ACCTTTAGGT ATGCCACAGA

                                                                  2150                                                                2200
TTTAATTCGA GCTAAGCGAT TCGTCGCTGC CATTATTCTG GGCATATCTG CTTTAATTGC TATTATCACT TCCTTTGCTG TAGCTACTAC TGCCTTAGTT

                                                                  2250                                                                2300
AAGGAGATGC AAACTGCTAC ATTTGTTAAT AATCTTCATA GAAATGTTAC ATTAGCCTTA TCCGAACAAA GAATAATAGA TTTAAAATTA GAAGCTAGAC

                                                                  2350                                                                2400
TTAATGCTTT AGAAGAAGTA GTTTTAGAGT TGGGACAAGA TGTGGCAAAC TTAAAGACCA GAATGTCCAC TAGGTGTCAT GCAAATTATG ACTTTATCTG

                                                                  2450                                                                2500
CGTTACACCT TTACCATATA ATGCTACTGA GGACTGGGAA AGAACCAGGG CCCATTTATT AGGCATTTGG AATGATAATG AGATTTCATA TAACATACAA

                                                                  2550                                                                2600
GAATTAACTA ACCTAATTAG TGATATGAGC AAACAACACA TTGACGCAGT GGACCTCAGT GGCTTGGCTC AGTCTTTTGC CAATGGAGTG AAGGCTTTAA

                                                                  2650                                                                2700
ATCCATTAGA TTGGACTCAA TATTTCATTT TTATAGGTGT TGGAGCCCTG CTTTTAGTCA TAGTACTTAT GATTTTCCCC ATTGTTTTCC AGTGCCTTGC

                                                                  2750                                                                2800
GAAGAGCCTT GACCAAGTGC AGTCAGATCT TAACGTGCTT CTTTTAAAAA AGAAAAAAGG GGGAATGCC GCGCCTGCAG CAGAAATGGT TGAACTCCCG

                                                                  2850                                                                2900
AGAGTGTCCT ACACTTAGGG GAGAAGCAGC CAAGGGGTTG TTTCCCACCA AGGACGACCC GTCTGCGCAC AAACGGATGA GCCCATCAGA CAAAGACATA

                                                                  2950                                                                3000
CTCATTCTCT GCTGCAAACT TGGCATAGCT CTGCTTTGCC TGGGGCTATT GGGGGAAGTT GCGGTTCGTG CTCGCAGGGC TCTCACCCTT GATTCTTTCA

ATAATAACTC TC
```

Fig. 2. The sequence of the MMTV *env* region. The dotted underlines indicate the potential splice acceptor sites. Those mentioned in the text extend from nucleotides 732–750, and 2672–2692. The dashed underlines indicate the *Sau*3A sites used to generate the DNA probe used in S1 mapping the *env* splice site. The boxes show the initiation codons of the three potential *env* starts, at nucleotides 752, 857 and 887; and for the open reading frame extending into the LTR at nucleotide 2766. The solid overlines point out the termination codons at nucleotide 801 for the presumed *pol* gene, and at nucleotide 2818 for the *env* gene.

A study of cellular and viral glycoproteins shows that the primary site of glycosylation occurs on the amino acid residue asparagine, where the sequence is Asn-X-Ser or Asn-X-Thr (Neuberger *et al.*, 1972). The drug tunicamycin inhibits glycosylation at these sites (Leavitt *et al.*, 1977), and experiments using a variety of tunicamycin concentrations has suggested that the MMTV envelope precursor (gPr73^{env}) has five asparagine linked sites of glycosylation (Dickson and Atterwill, 1980; Arthur *et al.*, 1982). The predicted amino acid sequence also has five such sites, three in gp52 and two in gp36 – which agrees with the experimental data.

An examination of the amino acid sequence of gp36 shows that it has a very high percentage of a non-charged amino acids (166/208) with two major regions of hydrophobicity at the amino- and carboxyl-terminal ends (see Figure 4). This agrees with the hydrophobic nature of gp36 determined experimentally (Marcus *et al.*, 1978).

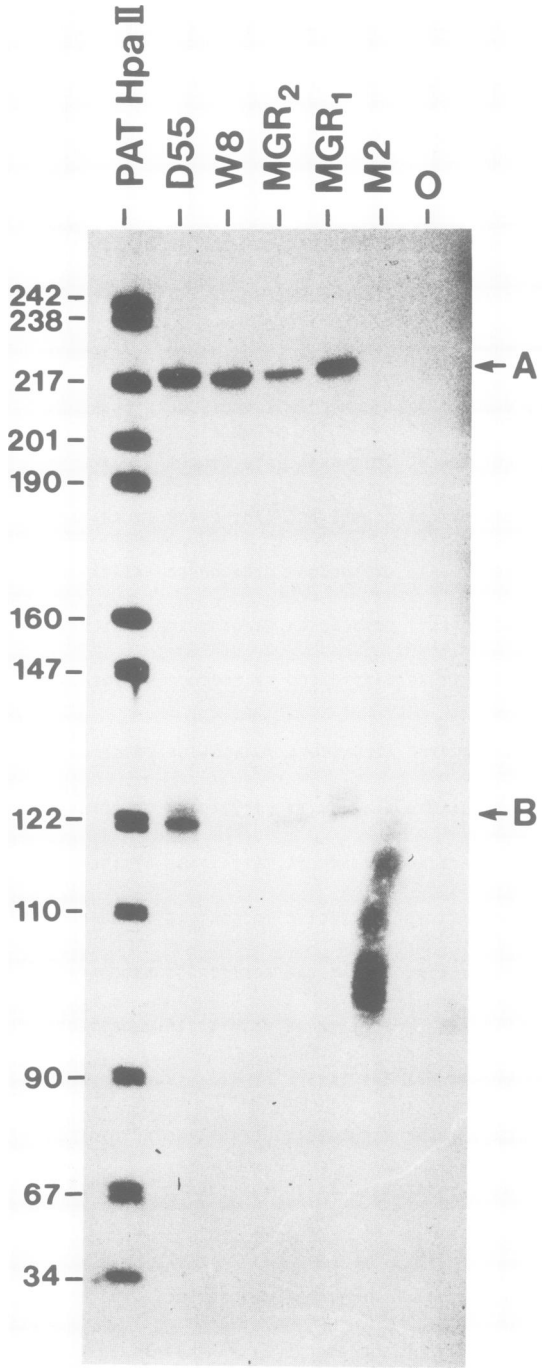

Fig. 3. The position of the *env* splice site. S1 nuclease analysis of RNA from MMTV-infected cells, and an MMTV-induced tumour. The probe used was the 5′ ^{32}P-labelled anti-sense strand *Sau3*A fragment spanning nucleotides 652 to 869 (see Figure 2). Hybridization reactions contained 10 μg RNA from the following cell lines, M2: a clone of the mink lung cell line (CCL64; MGR 1 + 2: two different preparations of RNA from mink lung cells chronically infected with GR strain MMTV; W8: RNA from an MMTV-induced tumour in a BR6 mouse; D55: rat XC cells infected with C3H strain MMTV. The total amount of RNA in all the tracks was made up to 25 μg with yeast carrier RNA. Track 0 contained carrier RNA alone. The marker used was 5′ ^{32}P-labelled PAT153 cut with *Hpa*II. Marker lengths are given in nucleotides.

Discussion

The organisation of the coding sequence of the 3′ half of the MMTV genome is extremely economical. The beginning of the *env* open reading frame overlaps for 51 nucleotides with the presumed end of the *pol* coding region. This is similar to the situation in Moloney murine leukaemia virus, where there is an analogous overlap of 63 nucleotides (Shinnick *et al.*, 1981). At the 3′ end of the genome the *env* coding region runs through the polypurine track, thought to have a role in priming positive strand DNA synthesis and overlaps for 53 nucleotides with the beginning of the open reading frame (*orf*) of the LTR.

There are three possible methionine codons at which the primary product of translation of the *env* gene could be initiated. All three (at positions 752, 857, 887 in the sequence; see Figures 2 and 4) have the correct 'context' to be possible initiation codons (Kozak, 1981). Circumstantial evidence suggests that the second or third methionine codon is the most likely start. This is partly based on the estimated size of the primary protein products from the three possible initiator methiones, which are 77 132, 73 159 and 71 822 daltons, respectively, in comparison with the size of the *in vitro* translated product of *env* mRNA, estimated on SDS-polyacrylamide gels to be 68 000–70 000 daltons (Sen *et al.*, 1979; Dudley and Varmus, 1981). This calculation would favour the third methionine. However, treatment of MMTV-infected cells *in vivo* with tunicamycin results in the production of a 60 000 dalton unglycosylated *env* precursor (Dickson and Atterwill, 1980; Arthur *et al.*, 1982), and comparing this apoprotein with the size of the primary product of translation shows that the leader sequence is 7000–9000 daltons long. Calculation of the sizes of the three potential leader sequences (from the three possible methionine starts to the amino terminus of gp52) gives values of 11 000, 7000 and 5700 daltons, respectively, a result which favours the second methionine. Taken together, this evidence suggests that the second or possibly the third methionine will turn out to be the authentic *env* gene initiator methionine although only direct amino acid sequencing will distinguish between these possibilities. The MMTV *env* gene appears to be more like the avian sarcoma virus *env* gene which has an unusually long leader sequence of 64 amino acids, rather than the Moloney murine leukaemia virus (MoMLV) which has a 33 amino acid leader (see Weiss *et al.*, 1982).

Within the long *env* leader sequence there is a stretch of 19 uncharged amino acids (see Figure 4) which could function as the signal sequence for membrane insertion (Blobel and Dobberstein, 1975; Engelman and Steitz, 1981). It is more common for the signal peptide and leader peptide to be the same size, so there may be another role for the 7000–9000

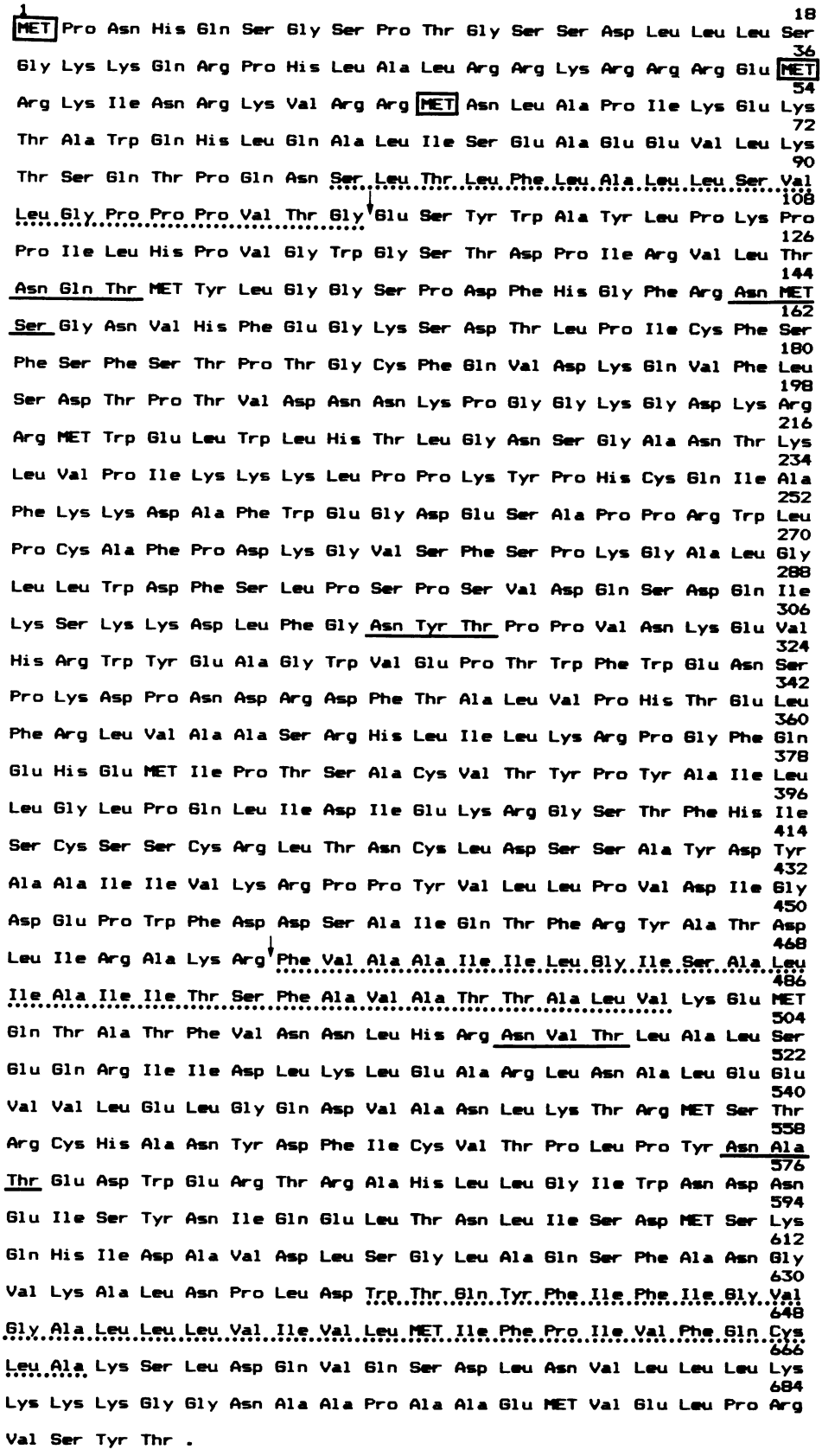

Fig. 4. The amino acid sequence of the *env* open reading frame. The three possible methionine starts are boxed. The arrowheads show the amino termini of gp52 and gp36. The solid underlines indicate the five potential carbohydrate addition sites. The dotted underlines draw attention to the hydrophobic signal sequence in the leader peptide of gp52; and the amino- and carboxy-terminal hydrophobic regions of gp36.

dalton *env* leader peptide once it has been released from the *env* precursor. Cleavage occurs just after a Gly residue, which is in accordance with the usual cleavage of leader peptides just after a small amino acid residue.

There are two differences between the derived amino acid sequence presented here and the published amino acid sequence of the amino terminus of gp52 (Arthur *et al.*, 1982). Both arise from single base differences, a G-A change for the Met ATG at position 1139 resulting in an Ile residue, and G-A change for the Ser AGT at position 1103, resulting in an Asn residue. These minor discrepancies probably reflect a strain difference as the amino acid sequence was determined using the C3H strain of MMTV, and the nucleotide sequence using the GR strain. The predicted amino acid sequences for the amino and carboxyl termini of gp36 are in perfect agreement with that derived by protein sequencing (Henderson *et al.*, 1983). The DNA sequence indicates that the amino terminus of gp36 falls just after the tetrapeptide Arg-Ala-Lys-Arg. A similar group of basic amino acids is found just before the site of cleavage in many other virus glycoproteins; for example, Rous sarcoma virus (Arg Arg Lys Arg, Schwartz *et al.*, 1982); MoMLV (Arg His Lys Arg, Shinnick *et al.*, 1981); AK virus (Lys Tyr Lys Arg, Lenz *et al.*, 1982); fowl plague virus HA (Lys Arg Glu Lys Arg, Porter *et al.*, 1979). An enzyme with the specificity of trypsin could perform this cleavage.

gp52 is accessible to labelling by lactoperoxidase, whereas gp36 is not (Witte *et al.*, 1973; Parks *et al.*, 1974). Although this is often taken to mean that gp52 is more exposed on the cell surface than gp36, another reason for the inability to label gp36 is suggested upon examination of the amino acid sequence. This reveals that gp52 has 11 Tyr residues which could be sites for lactoperoxidase catalysed iodination plus two other Tyr residues which are unlikely to be accessible due to their proximity to carbohydrate attachment sites, gp36 has only three Tyr residues, one of which occurs in a likely transmembrane region. Nevertheless, gp52 is thought to be the 'knob' on the MMTV virion, since it can be depleted by protease treatment of virions (Cardiff *et al.*, 1974), and may also be removed from the particle by washing with 0.05 M HCl (Sarkar *et al.*, 1976); properties consistent with a surface location. gp52 is held in close association with the envelope membrane *via* gp36 (Dion *et al.*, 1979; Racevskis and Sarkar, 1980), but their linkage is not by disulphide bridges as in AK virus, avian leukaemia virus and influenza HA (Pinter and Fleissner, 1977; Leamnson and Halper, 1976; Waterfield *et al.*, 1980). However, the absence of disulphide bonds is not a unique feature, as in some strains of murine leukaemia virus only a minor fraction of gp70 is disulphide-linked to p15(E), and none to p12(E), (Leamnson *et al.*, 1977; Pinter *et al.*, 1978).

The smaller glycoprotein gp36, is thought to be the hydrophobic trans-membrane 'stalk' supporting gp52. Its amino acid sequence reveals that it has two hydrophobic regions of 27 and 30 amino acids at the amino and carboxyl termini, respectively, one or both of which could be trans-membrane regions (Engelman and Steitz, 1981). However, by analogy with other glycoproteins, it seems likely that the hydrophobic region at the carboxyl terminus spans the viral membrane, leaving the NH_2-terminal portion of the molecule free to interact with gp52. The carboxy-terminal hydrophobic region is flanked by charged Asp and Lys residues which could interact with the polar head groups of the phospholipid bilayer. There would remain 38 amino acids of the carboxyl terminus exposed on the cytoplasmic side of the membrane, which would be free to interact with virion core particles during virus assembly. These contain the unusual string Lys-Lys-Lys-Lys as a result of translating through the polypurine tract. The hydrophobic amino terminus could be 'concealed' from the aqueous environment by the tertiary structure of gp36 or gp52. In this model, the amino terminus of gp36 could function as a fusion peptide, as seen for the influenza HA, vesicular stomatitis virus and Semliki Forest virus glycoproteins (White *et al.*, 1981). In these viruses low pH *in vitro*, or presumably in lysosomes *in vivo*, causes a change in conformation and/or the neutralisation of a few charged amino acids on an otherwise hydrophobic region of the glycoprotein complex. This unveiling of the hydrophobic region causes fusion of viral and cellular membranes. It is easy to extrapolate this to the MMTV *env* glycoproteins, where the low pH of the lysosomes could remove gp52 from gp36 (by neutralising ionic bonds between them), and consequently allow fusion and release of the virion core into the cytoplasm.

The DNA sequence presented here demonstrates the unusual coding organisation of the MMTV genome. In addition to the unique open reading frame (*orf*) contained within the LTR; the MMTV envelope gene is unusual in overlapping both with *orf* and with the resumed *pol* coding sequences. The DNA sequence also verifies a number of previous experimental observations on the MMTV envelope glycoproteins. It has provided their complete amino acid sequences which allow us to predict the function of various regions of the *env* gene. These predictions are currently under investigation.

Materials and methods

Source of DNA for sequencing

A recombinant plasmid 7-1a (kindly provided by J. Majors and H.E. Varmus, University of California, San Francisco), containing a deleted form of unintegrated circular DNA from the GR strain of MMTV, was the source of MMTV DNA. The *env* region of MMTV is contained mostly within the 0.95-kb *Eco*RI to *Pst*I and 1.85-kb *Pst*I to *Pst*I fragments located in the 3′ half of the provirus (see Figure 1). These fragments had been subcloned into PAT153 (G. Peters and S. Brookes, unpublished results) and were used to prepare DNA, as described by Birnboim and Doly (1979). The sequence extending into the LTR was obtained using a 1.4-kb *Pst*I fragment from unintegrated linear MMTV DNA also cloned into PAT153 (G. Peters, unpublished results). Plasmid DNA was digested with the appropriate restriction enzymes, and the DNA fragments separated on 0.8% agarose gels. The MMTV specific fragments were then electroeluted and concentrated by ethanol precipitation.

M13 cloning and sequencing

The purified MMTV DNA fragments described above were either used directly or further digested with a variety of restriction enzymes such as *Bam*HI, *Taq*I, *Hae*III, *Sau*3A (all purchased from New England Biolabs Inc.) before being cloned into M13. Digestions were performed in a standard buffer containing 20 mM Tris-HCl pH 7.6; 50 mM NaCl; 10 mM $Mg(OAc)_2$, 5 mM mercaptoethanol and 100 μg/ml bovine serum albumin. The DNA was ligated into the double-stranded replicative form (RF) of bacteriophages M13 mp7, 8 or 9 (Messing *et al.*, 1981; Messing and Viera, 1982), which had been cut with the appropriate restriction enzyme(s). M13 mp7 has symmetrically disposed restriction sites within its polylinker cloning site and so was used for the shotgun cloning of small fragments (100–400 nucleotides) produced by digestion with *Sau*3A, *Taq*I or *Hae*III. M13 mp8 and mp9 have asymmetric and oppositely arranged restriction sites and were used to clone fragments with two different restricted ends in one or both orientations, as determined by the vector. M13 mp8 and mp9 RFs were a gift from M. Jones (ICRF). *Escherichia coli* (JM101 or JM103) were transformed with the mixture of ligated RF forms of M13 using the calcium chloride method of Winter and Fields (1980).

Recombinant plaques were toothpicked into 1.5 ml of 2 x TY, containing early exponential JM101, and grown for 5 h at 37°C. The single-stranded

phage DNA was isolated as described by Winter and Fields (1980). A sample of phage DNA was analysed on agarose gels to determine the size of the insert. The presence of MMTV-specific DNA was confirmed by dot-blotting 0.5 μl of the DNA onto nitrocellulose and hybridising with ^{32}P-labelled MMTV cDNA (Kafatos *et al.*, 1979).

The single-stranded phage DNA was sequenced by the dideoxynucleotide method of Sanger *et al.* (1977). The single-stranded 15 nucleotide 'Universal primer', as described by Messing *et al.* (1981) (purchased from BRL) was used for most of the sequencing, although the earliest results were obtained using the cloned primer isolated from the plasmid pSP14 (Anderson *et al.*, 1980). The products of the sequencing reaction were run on thin 6% acrylamide/urea gels (Sanger and Coulson, 1978) from 1.5 to 6 h at 1500 V. Gels were routinely fixed in 10% acetic acid for 10 min, rinsed and dried with Kleenex tissues before being covered with Saran Wrap and autoradiographed overnight.

Cell culture

A clone (M2) of the mink lung cell line (CCL64) and the same cells chronically infected with MMTV strain GR (MGR-4) were obtained from G. Peters. A clone (D55) of transformed rat XC cells infected with MMTV strain C3H was a gift from D. Robertson and H.E. Varmus. All the cells were cultured in Dulbecco's modified Eagle's medium, supplemented with either 5% calf serum and 2% foetal calf serum (M2 and MGR) or with 5% foetal calf serum (D55).

RNA preparation

Subconfluent monolayers of cells were treated with 3 x 10^{-6} M dexamethasone for 24 h to stimulate MMTV RNA synthesis (Ringold *et al.*, 1975). The RNA was prepared from the treated cell cultures by the guanidinium thiocyanate method of Ullrich *et al.* (1977). RNA from a virally-induced mammary tumour (W8), prepared as above was also used in some experiments.

S1 mapping

Single-stranded 5′ ^{32}P-labelled probes were prepared as described by Maxam and Gilbert (1980). The probe used here was a 217 base *Sau*3A fragment, shown in the DNA sequence from nucleotide 648 to 865 (see Figure 2). All buffers and procedures used were as described by Favaloro *et al.* (1980). Hybridisation was carried out overnight at 30°C after denaturation at 85°C for 15 min. S1 digestion was for 30 min at 37°C with 100 U/ml of S1 nuclease (Sigma). The S1-resistant products were analysed on 6% acrylamide/urea gels (as above), exposed at −70°C to preflashed Fuji film, with intensifying screens.

Acknowledgements

We wish to thank John Majors for the generous gift of cloned MMTV DNA, and for communicating recently his unpublished sequence, and to Steven Oroszlan and Gordon Hager for communicating the work of Henderson *et al.* and Wheeler *et al.* before publication. Many thanks to Gordon Peters and Sharon Brookes for the gifts of plasmids, and to Mick Jones for RFs and a lot of help with M13 cloning and sequencing. We thank Gordon Peters, Mike Owen and John Arrand for critical reading of the manuscript and Audrey Gibson and Audrey Symons for its preparation.

References

Anderson,S., Gait,M.J., Mayol,L. and Young,I.G. (1980) *Nucleic Acids Res.*, **8**, 1741-1743.
Arthur,L.O., Copeman,T.D., Oroszlan,S. and Schochetman,G. (1982) *J. Virol.*, **41**, 414-422.
Berk,A.J. and Sharp,P.A. (1977) *Cell*, **12**, 721-732.
Birnboim,H.C. and Doly,J. (1979) *Nucleic Acids Res.*, **7**, 1513-1523.
Blobel,G. and Dobberstein,B. (1975) *J. Cell Biol.*, **67**, 835-851.
Breathnach,R., Benoist,C., O'Hare,K., Gannon,F. and Chambon,P. (1978) *Proc. Natl. Acad. Sci. USA*, **75**, 4853-4857.
Cardiff,R.D., Puentes,M.J., Teramoto,Y.A. and Lund,J.K. (1974) *J. Virol.*, **14**, 1293-1303.
Day,N.K., Witkin,S.S., Sarkar,N.H., Kinne,D., Jussawalla,D.J., Levin,A., Hsia,C.C., Geller,N. and Good,R.A. (1981) *Proc. Natl. Acad. Sci. USA*, **78**, 2483-2487.
Dickson,C. and Atterwill,M. (1980) *J. Virol.*, **35**, 349-361.
Dickson,C. and Peters,G. (1981) *J. Virol.*, **37**, 36-47.
Dion,A.S., Pomenti,A.A. and Farwell,D.C. (1979) *Virology*, **96**, 249-257.
Donehower,L., Huang,A. and Hager,G.L. (1981) *J. Virol.*, **37**, 226-238.
Dudley,J.P. and Varmus,H.E. (1981) *J. Virol.*, **39**, 207-218.
Engelman,D.M. and Steitz,T.A. (1981) *Cell*, **23**, 411-422.
Fasel,N., Pearson,K., Buetti,E. and Diggelmann,H. (1982) *EMBO J.*, **1**, 3-7.
Favaloro,J.M., Treisman,R.H. and Kamen,R. (1980) in Moldave,K. and Grossmann,L. (eds.), *Methods in Enzymology*, Vol. **65**, Academic Press, NY, pp. 718-749.
Groner,B., Hynes,N.E. and Diggelmann,H. (1979) *J. Virol.*, **30**, 417-420.
Henderson,L.E., Sowder,R., Smythers,G. and Oroszlan,S. (1983) *J. Virol.*, in press.
Kafatos,F.C., Jones,C.W. and Efstratiadis,A. (1979) *Nucleic Acids Res.*, **7**, 1541-1552.
Kozak,M. (1981) *Nucleic Acids Res.*, **9**, 5233-5252.
Leamnson,R.N. and Halpern,M.S. (1976) *J. Virol.*, **18**, 956-968.
Leamnson,R.N., Shander,M.H.M. and Halpern,M.S. (1977) *Virology*, **76**, 137-139.
Leavitt,R., Schlesinger,S. and Kornfield,S. (1977) *J. Virol.*, **21**, 375-385.
Lenz,J., Crowther,R., Straceski,A. and Haseltine,W. (1982) *J. Virol.*, **42**, 519-529.
Majors,J. (1982) in Weiss,R., Teich,N., Varmus,H.E. and Coffin,J. (eds.), *The Molecular Biology of Tumour Viruses: RNA Tumour Viruses*, 2nd Edn., Cold Spring Harbor Laboratory Press, NY, p.460.
Marcus,S.L., Smith,S.W., Racevskis,J. and Sarkar,N.H. (1978) *Virology*, **86**, 398-412.
Maxam,A.M. and Gilbert,W. (1980) in Moldave,K. and Grossmann,L. (eds.), *Methods in Enzymology*, Vol. **65**, Academic Press, NY, pp. 499-560.
Mesa-Tejada,R., Keydar,I., Ramanarayanan,M., Bausch,J. and Spiegelman, S. (1979) *Proc. Natl. Acad. Sci. USA*, **76**, 2460-2464.
Messing,J., Crea,R. and Seeburg,P.H. (1981) *Nucleic Acids Res.*, **9**, 309-321.
Messing,J. and Viera,J. (1982) *Gene*, in press.
Mount,S.M. (1982) *Nucleic Acids Res.*, **10**, 459-472.
Neuberger,A., Gottschalk,A., Marshall,R.D. and Spiro,R.G. (1972) in Gottschalk,A. (ed.), *The Glycoproteins, Their Composition, Structure and Function*, Elsevier, Amsterdam, pp. 450-490.
Ohno,T., Mesa-Tejada,R., Keydar,I., Ramanarayanan,M., Bausch,J. and Spiegelman,S. (1979) *Proc. Natl. Acad. Sci. USA*, **76**, 2460-2464.
Parks,W.P., Howk,R.S., Scolnick,E.M., Oroszlan,S. and Gilden,R.V. (1974) *J. Virol.*, **13**, 1200-1210.
Pinter,A. and Fleissner,E. (1977) *Virology*, **83**, 417-422.
Pinter,A., Lieman-Hurwitz,J. and Fleissner,E. (1978) *Virology*, **91**, 345-351.
Porter,A.G., Barber,C., Carey,N.H., Hallewell,R.A., Threlfall,G. and Emtage,J.S. (1979) *Nature*, **282**, 471-477.
Racevskis,J. and Sarkar,N.H. (1980) *J. Virol.*, **35**, 937-948.
Ringold,G.M., Yamamoto,K.R., Tomkins,G.M., Bishop,J.M. and Varmus, H.E. (1975) *Cell*, **6**, 299-305.
Robertson,D.L. and Varmus,H.E. (1979) *J. Virol.*, **30**, 576-589.
Robertson,D.L. and Varmus,H.E. (1981) *J. Virol.*, **40**, 673-682..
Sanger,F., Nicklen,S. and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5463-5467.
Sanger,F. and Coulson,A.R. (1978) *FEBS Lett.*, **87**, 107-110.
Sarkar,N.H., Taraschi,N.E., Pomenti,A.A. and Dion,A.S. (1976) *Virology*, **69**, 677-690.
Schwartz,D., Tizard,R. and Gilbert,W. (1982) in Weiss,R., Teich,N., Varmus,H.E. and Coffin,J. (eds.), *The Molecular Biology of Tumour Viruses, RNA Tumour Viruses*, 2nd Edn., Cold Spring Harbor Laboratory Press, NY, pp. 1338-1348.
Sen,G.C., Smith,S.W., Marcus,S.L. and Sarkar,N.H. (1979) *Proc. Natl. Acad. Sci. USA*, **76**, 1736-1740.
Shinnick,T.M., Lerner,R.A. and Sutcliffe,J.G. (1981) *Nature*, **293**, 543-548.
Ullrich,A., Shine,J., Churgurin,J., Pictet,R., Tischer,E., Rutter,W.J. and Goodman,H.M. (1977) *Science (Wash.)*, **196**, 1313-1319.
Waterfield,M.D., Gething,M.-J., Scrace,G. and Skehel,J.J. (1980) in Laver, G. and Air,G. (eds.), *Structure and Variation in Influenza Virus*, Elsevier, North Holland, pp. 11-20.
Weiss,R., Teich,N., Varmus,H.E. and Coffin,J. (eds.) (1982), *The Molecular Biology of Tumour Viruses: RNA Tumour Viruses*, 2nd Edn., published by Cold Spring Harbor Laboratory Press, NY.
Wheeler,D., Butel,J., Medina,D., Cardiff,R.D. and Hager,G. (1983) *J. Virol.*, in press.
White,J., Matlin,K. and Helenius,A. (1981) *J. Cell Biol.*, **89**, 674-679.
Winter,G. and Fields,S. (1980) *Nucleic Acids Res.*, **8**, 1965-1974.
Witte,O.N., Weissmann,I.L. and Kaplan,H.S. (1973) *Proc. Natl. Acad. Sci. USA*, **70**, 36-40.

Note added in proof

During the preparation of this manuscript we received a copy of the C3H strain MMTV *env* sequence from J. Majors. Comparison of the two sequences showed that they are virtually identical with only minor single base changes.