JOURNAL OF VIROLOGY, Oct. 1982, p. 1–11
0022-538X/82/100001-11$02.00/0

Vol. 44, No. 1

# DNA Sequence of the Viral and Cellular *src* Gene of Chickens

## I. Complete Nucleotide Sequence of an *Eco*RI Fragment of Recovered Avian Sarcoma Virus Which Codes for gp37 and pp60$^{src}$

TATSUO TAKEYA,* RICARDO A. FELDMAN, AND HIDESABURO HANAFUSA

*The Rockefeller University, New York, New York 10021*

Received 11 March 1982/Accepted 11 June 1982

Recovered avian sarcoma virus is a class of virus obtained from chicken tumors induced by mutants of Rous sarcoma virus which have a deletion in the *src* gene. We have determined the entire nucleotide sequence of a 3.1-kilobase *Eco*RI DNA fragment of molecularly cloned recovered avian sarcoma virus DNA. This DNA fragment contains part of the *env* gene and the entire *src* gene. Amino acid sequences of both gene products were deduced from the DNA sequences; the predicted amino acid sequences were verified by protein studies. An *env* protein (gp37) was found to be composed of 205 amino acids with three glycosylation sites. gp37 had a long stretch of hydrophobic residues near the carboxyl terminus. The *src* gene product, pp60$^{src}$, was composed of 526 amino acids and contained the possible sites for tyrosine and serine phosphorylation. The amino acid sequences predicted in this study differ significantly from the amino acid sequence predicted previously for the Schmidt-Ruppin strain of Rous sarcoma virus.

Recovered avian sarcoma viruses (rASVs) are a group of sarcoma viruses obtained from chicken tumors induced by transformation-defective (*td*) mutants of Rous sarcoma virus (RSV) (10–12, 40). Since the *td* mutants have a deletion within the *src* gene of RSV (19), rASVs have been considered to have acquired their *src* sequences from cellular-*src* (c-*src*), the cellular homolog of the viral *src* gene (v-*src*). The rASV *src* gene has been extensively analyzed by fingerprinting of RNase $T_1$-resistant oligonucleotides (42–44), and its product, pp60$^{src}$, has been examined by tryptic peptide analysis (17, 18). Results of these studies showed that the *src* gene of rASV is very similar to the *src* gene of its progenitor, the Schmidt-Ruppin strain of RSV, subgroup A (SR-A). However, some specific differences between SR-A and rASV have been detected in both the RNA and the protein. These differences seem to reflect the difference between the *src* sequences of SR-A and the c-*src* gene (17, 42–44). Therefore, these data strongly suggest that rASVs are recombinants between *td* mutants and c-*src*.

We have cloned DNAs of the interacting elements in this system: SR-A, rASV1441 (one of the rASVs), and c-*src* (38). The sequencing of these DNAs should provide information about the structure of the *src* gene and its product pp60$^{src}$, the origin of rASV, and possibly the mechanism of rASV generation. In a previous study (38), we showed that digestion of rASV1441-cloned DNA with several restriction endonucleases produced restriction patterns almost identical to those of standard SR-A DNA. This indicates that no significant rearrangement, deletion, or insertion is present in the rASV1441 genome.

In this paper, we report the entire nucleotide sequence of the 3.1-kilobase (kb) *Eco*RI fragment of rASV1441 containing a part of the *env* gene and the entire *src* gene. We describe here full details of various characteristics of the DNA and protein sequences of the two genes. The data were derived from rASV1441; however, the conclusions are, for the most part, applicable to those of standard SR-A. As we show in the accompanying paper (37), the DNA sequence and deduced protein sequence of the *src* gene are nearly identical in rASV1441 and our SR-A virus. However, significant differences were found in the *env* and *src* DNA sequences and predicted protein sequences of rASV1441 compared with SR-A previously reported by Czernilofsky et al. (5). Therefore, to confirm our sequence data, examination of the peptides of the glycoprotein gp37 and pp60$^{src}$ was carried out.

## MATERIALS AND METHODS

**Viruses and molecular cloning.** rASV1441 was derived from a *td* mutant of SR-A, *td*108 (19). The isolation and characterization of rASV1441 have been described (10, 18, 42). The preparation of our SR-A virus and Rous-associated virus 2 (RAV-2) was previously described (20). rASV1441 has been molecularly cloned into the vector λgtWES · B. The 3.1-kb *Eco*RI fragment which contains the *src* gene and a part of the *env* gene was subcloned into pBR322 (pTT108) (38). DNA sequencing has been carried out on this plasmid DNA.

**DNA sequencing.** The DNA sequence was determined by the chemical method described by Maxam and Gilbert (27). For pyrimidine-specific reactions, the modification by Rubin and Schmid (31) was used. Restriction enzymes purchased from Bethesda Research Laboratories and New England Biolabs were used according to the suppliers' instructions. Purified DNA fragments were labeled by using either T4 polynucleotide kinase and [$\gamma$-$^{32}$P]ATP (for the 5′ termini) or terminal deoxynucleotidyl transferase and [$\alpha$-$^{32}$P]cordycepin 5′-triphosphates (for the 3′ termini) (39).

**Isotopic labeling of cells and virus purification. (i) Amino acid labeling.** rASV1441- or SR-A-transformed cells grown in 100-mm tissue culture plates were incubated with 3 ml of minimum essential medium lacking an amino acid, which was to be added later as labeled amino acid (prepared from the minimum essential medium select-amine kit [GIBCO Laboratories] and supplemented with 1% calf serum). After 1 h, the medium was changed and replaced with 3 ml of fresh medium containing 200 $\mu$Ci of labeled amino acid per ml: L-[$^{35}$S]methionine (900 to 1,200 Ci/mmol; Amersham Corp.), L-[$^{3}$H]tyrosine (525 Ci/mmol; New England Nuclear Corp.), or L-[$^{3}$H]phenylalanine (60 Ci/mmol; New England Nuclear). Incubation was continued for 8 h. At the end of this period cells and viruses were harvested.

**(ii) Glucosamine labeling.** RAV-2-infected cells grown in 60-mm culture plates were incubated in 3 ml of F-10 medium containing 300 $\mu$Ci of D-[$^{3}$H]glucosamine (30 Ci/mmol; New England Nuclear) for 24 h, and cell lysates were prepared.

Purification of radiolabeled rASV1441 from the culture medium was carried out as described previously (30).

Preparation of cell extracts, immunoprecipitation, and sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) were performed as described before (7).

**V8 protease digestion analysis.** Protein samples from excised gel bands were digested with V8 protease and analyzed by SDS-PAGE as described previously (17, 25).

**Cyanogen bromide cleavage analysis.** Excised pp60$^{src}$ gel bands were washed five times in 10% methanol and then lyophilized to dryness. Gel bands were then incubated in either 0.5 ml of 70% formic acid alone or 0.5 ml of 70% formic acid (vol/vol) containing 50 mg of cyanogen bromide per ml at room temperature for 30 min. The gel slices were washed five times with water for 5 min each time and then lyophilized to dryness. Dried slices were loaded onto a 12% polyacrylamide

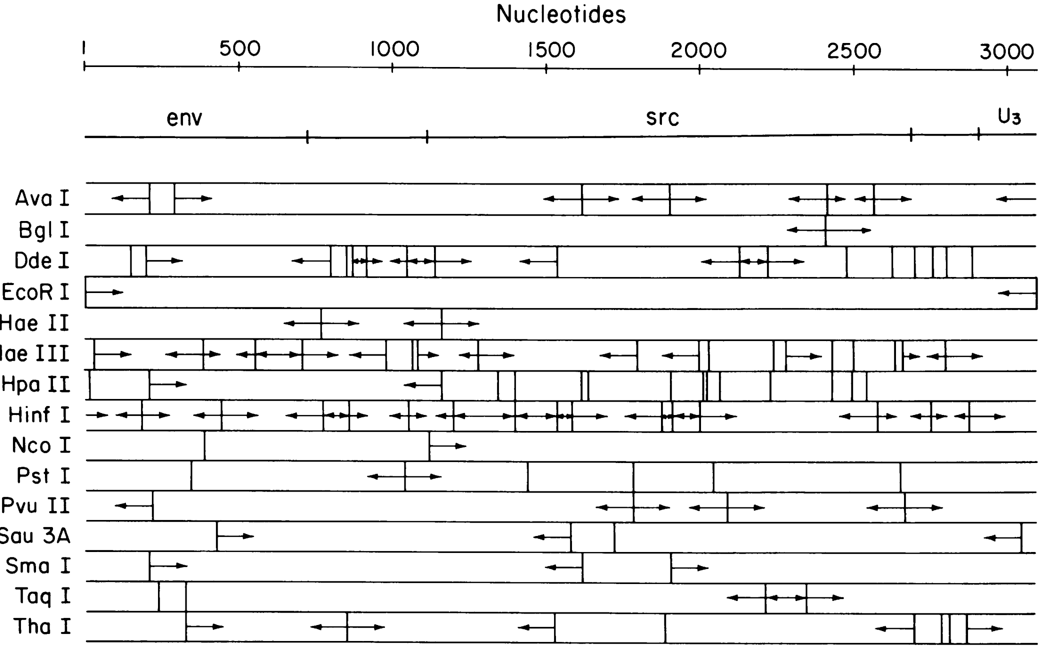


FIG. 1. Restriction enzyme sites on the 3.1-kb *Eco*RI fragment of rASV1441 and strategy for DNA sequence analysis. Restriction sites of several enzymes which have been used for DNA sequencing are shown. Some of these sites were described previously (38). Arrows indicate the starting sites and the orientation of each sequencing reaction.

gel and analyzed by SDS-PAGE as described for the V8 protease analysis.

**Antiserum.** The following specific antisera were used: serum from rabbits bearing tumors induced by the Schmidt-Ruppin strain of RSV, subgroup D (TBR serum) (1, 18), and a rabbit antiserum against the subgroup E glycoprotein of RAV-60 (30). The latter antiserum is able to immunoprecipitate viral glycoproteins of subgroups A through F.

## RESULTS AND DISCUSSION

**Nucleotide sequence of the 3.1-kb *Eco*RI fragment.** Detailed cleavage maps of the pTT108 insert, containing the *Eco*RI 3.1-kb fragment of rASV1441 DNA, were constructed for a number of restriction enzymes (Fig. 1). Based on these maps, DNA sequencing was carried out by the method of Maxam and Gilbert (27). Arrows in Fig. 1 indicate the orientation and the starting sites for each sequencing reaction. The complete nucleotide sequence numbered from the first nucleotide of the 5′ end, is shown in Fig. 2. The total length of the fragment was 3,111 nucleotides.

One of the *Eco*RI sites in RSV DNA is present in the U3 region of the long terminal repeat. The sequence of the last 189 nucleotides of the fragment starting with nucleotide 2,926 (AATGTA–) matches the published sequences of the 5′ end of the long terminal repeat (16, 35).

Czernilofsky et al. reported the presence of direct repeats before and after the *src* region (5). We also see homologous base sequences in the same orientation at positions 783 to 894 and 2,797 to 2,915; the homology of the two sequences is about 81%.

Czernilofsky et al. (5) reported the entire nucleotide sequence of the same *Eco*RI fragment of another strain of RSV [SR-A; this strain will be referred to in this paper as SR-A(SF) to distinguish from our SR-A(NY) virus]. The size of the fragment of SR-A(SF) was reported to be 3,092 nucleotides. Comparison of the two sequences shows differences in 159 nucleotides; these include base changes, additions, deletions, and substitutions. These differences appeared throughout the entire fragment, but about 70% of them are located in the middle of the fragment (nucleotides 640 to 1,801). At one stretch (within the region 1,031 and 1,067 according to our numbering), a sequence of 24 nucleotides is deleted in SR-A(SF). The differences may be partially due to the difference in the source of the viruses [rASV versus SR-A and SR-A(SF) versus SR-A(NY)]. However, in the accompanying paper (37), we show that rASV1441 and our strain of SR-A(NY) differ in only 17 nucleotides within the *src* gene.

**Reading frames and identification of coding sequences.** The possible translational products from three reading frames on this DNA fragment are shown in Fig. 3. In frame 1, there is no long open reading frame.

Frame 2 in Fig. 3 shows two large open reading frames (nucleotides 2 to 739 and 1,121 to 2,698). It is well established that the 5′ terminus of the 3.1-kb *Eco*RI fragment maps in the middle of the *env* gene (33). Therefore, the polypeptide encoded in the nucleotide sequence 2 to 739 must be part of the *env* protein. The second open reading frame is from nucleotides 1,121 to 2,698. We presume this region to be the *src* gene of rASV1441 for the following reasons. (i) This region encodes 526 amino acids of total predicted molecular weight of $58.8 \times 10^3$, which is close to the known size of pp60$^{src}$ (18, 29). (ii) The location of this open reading frame corresponds to the map location of the *src* gene within the genome of RSV (15, 41). (iii) This region contains the nucleotide sequences which match with the composition of oligonucleotides identified as *src* specific (42–44).

In frame 3, there is one large open reading frame (nucleotides 894 to 1,715) which would encode 274 amino acids. We do not know whether this polypeptide is synthesized in infected cells as a nonstructural protein. We learned, however, that the sequence of our standard SR-A(NY) has a termination codon at 1,049 (37), so SR-A(NY) cannot encode a large protein from this region. If it is produced by rASV1441, this polypeptide is unique to recovered virus and would not be essential for virus propagation or cell transformation.

Recently, Swanstrom et al. (36) suggested that the 5′-terminal leader sequence of RSV RNA, which is spliced to the 5′ ends of subgenomic mRNAs, consists of 389 nucleotides and includes 18 nucleotides within the coding region for the *gag* protein. Within the intercistronic region upstream of the *src* sequence, we detect three possible consensus sequences for splice acceptor sites (23): CCTTAGG at nucleotide 864, CCTTAGA at 919, and CTGCAGG at 1,039. It is formally possible that the leader sequence is spliced to one of these sites and the initiation codon in the leader sequence is used in the translation of the *src* protein. However, splicing to the first acceptor site would be in phase with the second open reading frame in frame 3. This protein cannot be *src* protein because this is unique to rASV1441 as discussed above. Splicing at the two other acceptor sites would be in phase with frame 1, but this frame contains termination codons at 949 and 1,054. We concluded that the initiation codon within the leader sequence does not serve to initiate synthesis of the *src* protein; *src* must therefore be initiated at position 1,121.

From these analyses, we defined the nucleo-

```
  1                                              61
  IleProSerArgProValGlyGlyProCysTyrLeuGlyLysLeuThrMetLeuAlaProAsnHisThrAspIle
 AATTCCCAGTCGTCCGGTAGGGGGCCCCTGCTATTTAGGCAAGCTCACCATGTTAGCACCCAACCATACAGATATT

                                  121 ↓— gp37 —→
LeuLysIleLeuAlaAsnSerSerArgThrGlyIleArgArgLysArgSerValSerHisLeuAspAspThrCysSer
CTCAAAATTCTTGCTAATTCATCACGGACAGGAATAAGACGTAAACGAAGCGTCTCACACCTGGATGATACATGCTCA
                    (20)
                    181
AspGluValGlnLeuTrpGlyProThrAlaArgIlePheAlaSerIleLeuAlaProGlyValAlaAlaAlaGlnAla
GATGAAGTACAGCTTTGGGGTCCTACAGCAAGAATCTTTGCGTCTATCTTAGCCCCGGGGGTAGCAGCTGCACAAGCC
      (40)                                            (60)
      241                            □  □  □          301
LeuLysGluIleGluArgLeuAlaCysTrpSerValLysGlnAlaAsnLeuThrThrSerLeuLeuGlyAspLeuLeu
TTAAAAGAAATCGAGAGACTAGCCTGTTGGTCCGTTAAACAGGCTAACTTGACAACATCACTCCTCGGGGACTTATTG
                                      (80)
                                      361
AspAspValThrSerIleArgHisAlaValLeuGlnAsnArgAlaAlaIleAspPheLeuLeuLeuAlaHisGlyHis
GATGATGTCACGAGTATTCGACACGCGGTCCTGCAGAACCGAGCGGCTATTGATTTCTTGCTCCTAGCTCACGGCCAT
                         (100)
                         421□ □  □
GlyCysGluAspValAlaGlyMetCysCysPheAsnLeuSerAspHisSerGluSerIleGlnLysLysPheGlnLeu
GGCTGTGAGGACGTTGCTGGAATGTGCTGTTTCAATTTGAGTGATCACAGTGAGTCTATACAGAAGAAGTTCCAGCTA
           (120)                                             (140)
           481                                               541
MetLysGluHisValAsnLysIleGlyValAspSerAspProIleGlySerTrpLeuArgGlyLeuPheGlyGlyIle
ATGAAGGAACATGTCAATAAGATCGGCGTGGACAGCGACCCAATTGGAAGTTGGCTGCGAGGACTATTCGGGGGAATA
                                           (160)
                                           601
GlyGluTrpAlaValHisLeuLeuLysGlyLeuLeuLeuGlyLeuValValIleLeuLeuLeuValValCysLeuPro
GGAGAATGGGCCGTTCATTTGCTGAAAGGACTGCTTTTGGGGCTTGTAGTTATTTTGTTGCTAGTAGTGTGCCTGCCT
                             (180)
                             661  □  □  □
CysLeuLeuGlnIleValCysGlyAsnIleArgLysMetIleAsnAsnSerIleSerTyrHisThrGluTyrLysLys
TGCCTTTTGCAAATCGTATGCGGTAACATCAGAAAGATGATTAATAACTCCATCAGCTACCACACGGAATATAAGAAG
               (200)
               721
LeuGlnLysAlaTyrGlyGlnProGluSerArgIleVal
CTACAAAAGGCCTATGGGCAGCCTGAAAGCAGAATAGTATAAGGCAGTACATGGGTGGTGGTATAGCGCTTGTGAGTC

  781                                          841
GGGTTGTAACGGGGCATGGCTTAACTAAGGGGACTATGGCATGTATAGGCGCAAAGCGGGGTTACGGTACGCGACTTA
    ----------------------------------------------------------------------------

                                  901
GGAGTCCCCTTAGGATATAGTAGACACGCTTTTGCATATGTTACATAACTTCCCTGTTTTGCCCTTAGACTATTCAAG
-------------------------------------

                    961
TTGCCTCTGTGGATTAGGGCTGGAGGCAGCTCGGATGGTCGGACGGCCAGATAAGGCAGGAAAGACAGCTATTGGTAA

      1021                                               1081
TTGTGAAATACGCTTTTGTCTGTGTGCTGCAGGAGCTGAGCTGACTCTACGTAGTGGCCTCACGTACCACTGTGGCCA

                       (1)             1141
                       MetGlySerSerLysSerLysProLysAspProSerGlnArgArgCys
GGCGGTAGCTGGGACGTGCAGCCCACCACCATGGGGAGCAGCAAGAGCAAGCCTAAGGACCCCAGCCAGCGCCGGTGC

 ●     (20)               1201                            (40)
SerLeuGluProProAspSerThrHisHisGlyGlyPheProAlaSerGlnThrProAsnLysThrAlaAlaProAsp
AGCCTGGAGCCACCCGACAGCACCCACCACGGGGGATTCCCAGCCTCGCAGACCCCCAACAAGACAGCAGCCCCCGAC

           1261                            (60)
ThrHisArgThrProSerArgSerPheGlyThrValAlaThrGluProLysLeuPheGlyGlyPheAsnThrSerAsp
ACGCACCGCACCCCCAGCCGCTCCTTTGGGACCGTGGCCACCGAGCCCAAGCTCTTCGGGGGCTTCAACACTTCTGAC

                         (80)              1381
ThrValThrSerProGlnArgAlaGlyAlaLeuAlaGlyGlyValThrThrPheValAlaLeuTyrAspTyrGluSer
ACCGTCACGTCGCCGCAGCGTGCCGGGGCACTGGCTGGCGGCGTCACCACTTTCGTGGCTCTCTACGACTACGAGTCC

           (100)                 1441                        (120)
ArgThrGluThrAspLeuSerPheLysLysGlyGluArgLeuGlnIleValAsnAsnThrGluGlyAspTrpTrpLeu
CGGACTGAAACGGACTTGTCCTTCAAGAAAGGAGAACGCCTGCAGATTGTCAACAACACGGAAGGTGACTGGTGGCTG

               1501                        (140)
AlaHisSerLeuThrThrGlyGlnThrGlyTyrIleProSerAsnTyrValAlaProSerAspSerIleGlnAlaGlu
GCTCATTCCCTCACTACAGGACAGACGGGCTACATCCCCAGTAACTATGTCGCGCCCTCAGACTCCATCCAGGCTGAA
```

```
                                       V8
  1561                           ●     ↓(160)                1621
GluTrpTyrPheGlyLysIleThrArgArgGluSerGluArgLeuLeuLeuAsnProGluAsnProArgGlyThrPhe
GAGTGGTACTTTGGGAAGATCACTCGTCGGGAGTCCGAGCGGCTGCTGCTCAACCCCGAAAACCCCCGGGGAACCTTC

                 V8
                  ↓ (180)                   1681
LeuValArgGluSerGluThrThrLysGlyAlaTyrCysLeuSerValSerAspPheAspAsnAlaLysGlyLeuAsn
TTGGTCCGGGAGAGCGAGACGACAAAAGGTGCCTATTGCCTCTCCGTTTCTGACTTTGACAACGCCAAGGGGCTCAAT

  (200)                   1741                                 (220)
ValLysHisTyrLysIleArgLysLeuAspSerGlyGlyPheTyrIleThrSerArgThrGlnPheSerSerLeuGln
GTGAAGCACTACAAGATCCGCAAGCTGGACAGCGGCGGCTTCTACATCACCTCACGCACACAGTTCAGCAGCCTGCAG

        1801                                (240)                  1861
GlnLeuValAlaTyrTyrSerLysHisAlaAspGlyLeuCysHisArgLeuThrAsnValCysProThrSerLysPro
CAGCTGGTGGCCTACTACTCCAAACACGCTGATGGCTTGTGCCACCGCCTGACCAACGTCTGCCCCACGTCCAAGCCC

                          (260)                   1921
GlnThrGlnGlyLeuAlaLysAspAlaTrpGluIleProArgGluSerLeuArgLeuGluValLysLeuGlyGlnGly
CAGACCCAGGGACTCGCCAAGGACGCGTGGGAAATCCCCCGGGAGTCGCTGCGGCTGGAGGTGAAGCTGGGGCAGGGC

         (280)                  1981                                (300)
CysPheGlyGluValTrpMetGlyThrTrpAsnGlyThrThrArgValAlaIleLysThrLeuLysProGlyThrMet
TGCTTTGGAGAGGTCTGGATGGGGACCTGGAACGGCACCACCAGAGTGGCCATAAAGACTCTGAAGCCCGGCACCATG

              2041                                (320)
SerProGluAlaPheLeuGlnGluAlaGlnValMetLysLysLeuArgHisGluLysLeuValGlnLeuTyrAlaVal
TCCCCGGAGGCCTTCCTGCAGGAAGCCCAAGTGATGAAGAAGCTCCGGCATGAGAAGCTGGTACAGCTGTACGCAGTG
        V8
         ↓                       (340)                  2161
ValSerGluGluProIleTyrIleValThrGluTyrMetSerLysGlySerLeuLeuAspPheLeuLysGlyGluMet
GTGTCGGAAGAGCCCATCTACATCGTCACTGAGTACATGAGCAAGGGGAGCCTCCTGGATTTCCTGAAGGGAGAGATG

               (360)                  2221                                 (380)
GlyLysTyrLeuArgLeuProGlnLeuValAspMetAlaAlaGlnIleAlaSerGlyMetAlaTyrValGluArgMet
GGCAAGTACCTGCGGCTGCCACAGCTCGTCGATATGGCTGCTCAGATTGCATCCGGCATGGCCTATGTGGAGAGGATG

                    2281                                 (400)
AsnTyrValHisArgAspLeuArgAlaAlaAsnIleLeuValGlyGluAsnLeuValCysLysValAlaAspPheGly
AACTACGTGCACCGAGACCTGCGGGCGGCCAACATCCTGGTGGGGGAGAACCTGGTGTGCAAGGTGGCTGACTTTGGG

  2341                       ■         (420)                 2401
LeuAlaArgLeuIleGluAspAsnGluTyrThrAlaArgGlnGlyAlaLysPheProIleLysTrpThrAlaProGlu
CTGGCACGCCTCATCGAGGACAACGAGTACACAGCACGGCAAGGTGCCAAGTTCCCCATCAAGTGGACAGCCCCCGAG

                     (440)                  2461
AlaAlaLeuTyrGlyArgPheThrIleLysSerAspValTrpSerPheGlyIleLeuLeuThrGluLeuThrThrLys
GCAGCCCTCTATGGCCGGTTCACCATCAAGTCGGATGTCTGGTCCTTCGGCATCCTGCTGACTGAGCTGACCACCAAG

   (460)                  2521                                 (480)
GlyArgValProTyrProGlyMetGlyAsnGlyGluValLeuAspArgValGluArgGlyTyrArgMetProCysPro
GGCCGGGTGCCATACCCAGGGATGGGCAACGGGGAGGTGCTGGACCGGGTGGAGAGGGGCTACCGCATGCCCTGCCCG

         2581                                (500)                  2641
ProGluCysProGluSerLeuHisAspLeuMetCysGlnCysTrpArgArgAspProGluGluArgProThrPheGlu
CCCGAGTGCCCCGAGTCGCTGCATGACCTTATGTGCCAGTGCTGGCGGAGGGACCCTGAGGAGCGGCCCACTTTTGAG

                           (520)                  2701
TyrLeuGlnAlaGlnLeuLeuProAlaCysValLeuGluValAlaGlu
TACCTGCAGGCCCAGCTGCTTCCTGCTTGTGTGTTGGAGGTCGCTGAGTAGTGCGCGAGCAAAATTTAAGCTACAACA

                                2761
AGGCAAGGCTTGGCCGACAATTGCATGAAGAATCTGCTTAGGGTTAGGCGCTTTGCGCTGCTTCGCGATGTACGGCCA
                                                                    ----------
              2821
GATATACGCGTATCTGAGGGGACTAGGGTGTGTTTAGGCGAAAAGCGGGGCTTCGGTTGTACGCGGTTAGGAGTCCCC
------------------------------------------------------------------------------
                                                        2941
CCTCAGGATATAGTAGTTTCGCTTTTGCATAGGGAAGGGGAAATGTAGTCTTATGCAATACTCTTGTAGTCTTGCAAC
-------------------------------
                                       3001
ATGCTTATGTAACGATGAGTTAGCAACATGCCTTACTTGGAGAGAAAAAGCACCGTGCATGCCGATTGGTGGAAGTAA

                    3061
GGTGGTACGATCGTGCCTTATTAGGAAGGCAACAGACGGGTCTGACATGGATTGGACAAACCACCGAATT
```

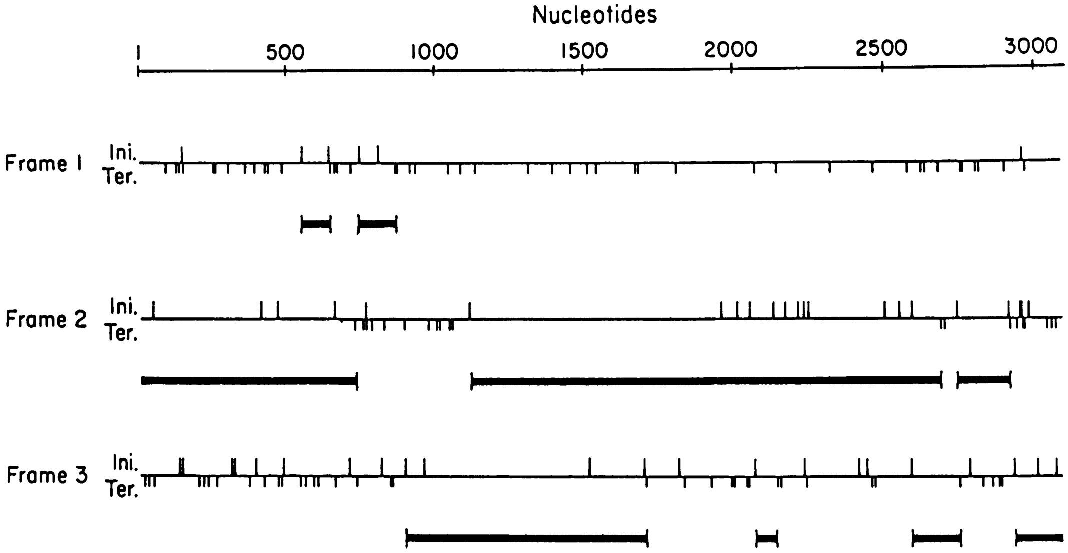


FIG. 3. Location of initiation and termination codons for protein synthesis and possible translational products. Ini. indicates the initiation codon (AUG), and Ter. indicates the termination codons (UAG, UAA, and UGA). Solid bars show possible open reading frames. Open reading frames shorter than 70 nucleotides are not shown.

tide sequences 1 to 734 as the *env* region, 735 to 1,120 as an intercistronic region, 1,121 to 2,698 as the *src* region, 2,699 to 2,925 as another noncoding region, and 2,926 to 3,111 as a part of the long terminal repeat. The deduced amino acid sequences of the *env* and *src* regions are shown in Fig. 2. The five amino acids (Ser-Val-Ser-His-Leu) encoded by nucleotides 125 to 139 are identical to the N-terminal amino acids of gp37 of the Prague strain of RSV (E. Hunter, personal communication), making Arg-Ser at position 124 to 125 the possible cleavage site on the precursor gp85-gp37 protein. We presume that this sequence defines the amino-terminal end of the rASV1441 gp37 and that the protein then consists of a total of 205 amino acids. The positions of amino acids in both *env* and *src* proteins are given in parentheses in Fig. 2.

After the boundaries of coding and noncoding regions are defined as described above, the 159-nucleotide difference between rASV1441 and SR-A(SF) is divided as follows: 26 in gp37, 59 in the intercistronic region, and 48 in the *src* region. These differences in the nucleotide sequence significantly affect the predicted amino acid sequences. Particularly, addition or deletion of one or two nucleotides results in a shift of the reading frame. In the gp37 region, a total of 40 amino acids are different; among them, the amino acids in positions 27 to 34 and 184 to 205 differ because of changes in the reading frame. Likewise, a total of 227 amino acids are different in pp60$^{src}$; the changes in amino acid positions 20 to 187, 220 to 224, 266 to 305, 367 to 373, and 504 to 510 are due to the reading of different frames.

**Analysis of gp37 and pp60$^{src}$.** Because substantial differences were found in the amino acid sequences deduced from the nucleotide sequences of rASV1441 shown in Fig. 2 and those of SR-A(SF) described by Czernilofsky et al. (5), we attempted to verify our results by analyzing proteins gp37 and pp60$^{src}$.

The rASV1441 sequence contains three tyrosyl residues in gp37, whereas the predicted sequence of SR-A(SF) gp37 contains none. Cells infected with rASV1441 were therefore labeled with [$^3$H]tyrosine, and labeled virions were purified. Labeled proteins in the virions were analyzed by SDS-PAGE. As a control, immunopre-

FIG. 2. Entire nucleotide sequence of the 3.1-kb *Eco*RI fragment of rASV1441. Nucleotides are numbered from 1 to 3,111 in the same polarity as the viral genomic RNA. Numbers in parentheses represent the amino acid positions in gp37 and pp60$^{src}$. These two gene products were localized as described in the text. Open squares indicate the possible glycosylation sites (32). Closed squares and closed circles indicate the sites for tyrosine phosphorylation (34) and the possible site for serine phosphorylation (21), respectively. Putative V8 protease cleavage sites are shown as V8. Dashed lines indicate the direct repeat sequences which were originally found by Czernilofsky et al. (5) in SR-A(SF).

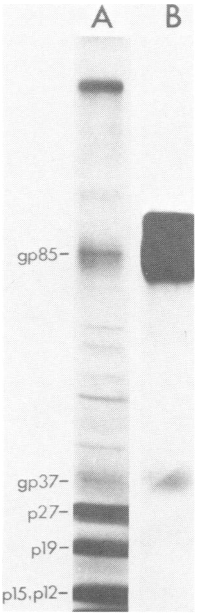


FIG. 4. Labeling of rASV1441 gp37 by [$^3$H]tyrosine. rASV1441 labeled with [$^3$H]tyrosine was purified from infected cell cultures as described in the text. RAV-2-infected cells were labeled with [$^3$H]glucosamine, and cell lysates were prepared. Total virion proteins labeled with [$^3$H]tyrosine (A) or immunoprecipitates of [$^3$H]glucosamine-labeled viral proteins by antiserum against viral glycoproteins (B) were analyzed by SDS-PAGE in a 5 to 15% gradient gel followed by fluorography.

cipitated viral glycoproteins from RAV-2-infected cells labeled with [$^3$H]glucosamine were also analyzed by SDS-PAGE. In virions prepared from rASV1441-infected cells, [$^3$H]tyrosine-labeled gp37 was detectable (Fig. 4), indicating that there are tyrosyl residues in rASV1441 gp37. The diffuse appearance of the gp85 and gp37 bands in Fig. 4 is due to the glycosylation of the proteins. gp85 and gp37 obtained from rASV1441 virions were also immunoprecipitable by the specific antiserum (data not shown).

To confirm the amino acid sequence of pp60$^{src}$ deduced from the DNA sequence, we used two approaches. The first approach was to examine the size of peptides generated by treatment of pp60$^{src}$ with cyanogen bromide, which cleaves peptides at methionine residues (8, 9). The amino acid sequence of rASV1441 pp60$^{src}$ (Fig. 2) suggests that cleavage at methionine residues would produce two long peptides, one containing 282 amino acids (positions 2 to 283; expected molecular weight, 29.9 × 10$^3$) and the other containing 86 amino acids (positions 381 to 466; expected molecular weight, 10.0 × 10$^3$). Cultures infected with rASV1441 were labeled with [$^3$H]tyrosine, and pp60$^{src}$ was isolated by SDS-PAGE after immunoprecipitation of cell lysates with TBR serum. Gel bands containing pp60$^{src}$ were then subjected to cyanogen bromide treatment. The sizes of peptides generated by this treatment were approximately 30 and 10 kilodaltons (kd) (Fig. 5), as expected from the predicted sequences. The results are not consistent, however, with the sequence of pp60$^{src}$ deduced for SR-A(SF), since it contains methionine residues at amino acid positions 125, 137, and 317 in the N-terminal regions. Two major fragments of about 20 and 14 kd would have been expected (5).

The second approach was to examine the distribution of certain amino acids within pp60$^{src}$. The *Staphylococcus aureus* V8 protease is known to cleave pp60$^{src}$ into two major fragments, an N-terminal 34-kd fragment (V1) and a C-terminal 26-kd fragment (V2) (3). The V1 fragment is further cleaved into subsets of 18

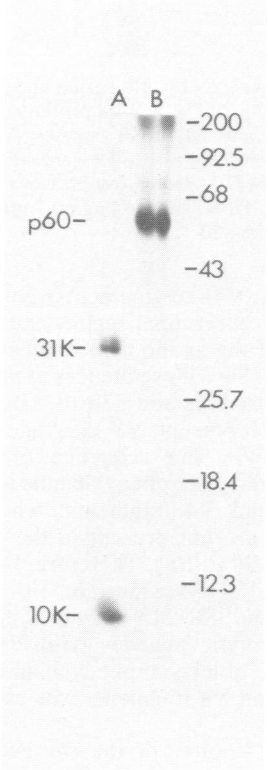


FIG. 5. Size of cyanogen bromide fragments generated from rASV1441 pp60$^{src}$. rASV1441 pp60$^{src}$ labeled with [$^3$H]tyrosine was subjected to cyanogen bromide treatment and analyzed by SDS-PAGE as described in the text. (A) Cyanogen bromide treated; (B) untreated. The sizes are indicated on the left. The positions of molecular weight markers are indicated on the right.

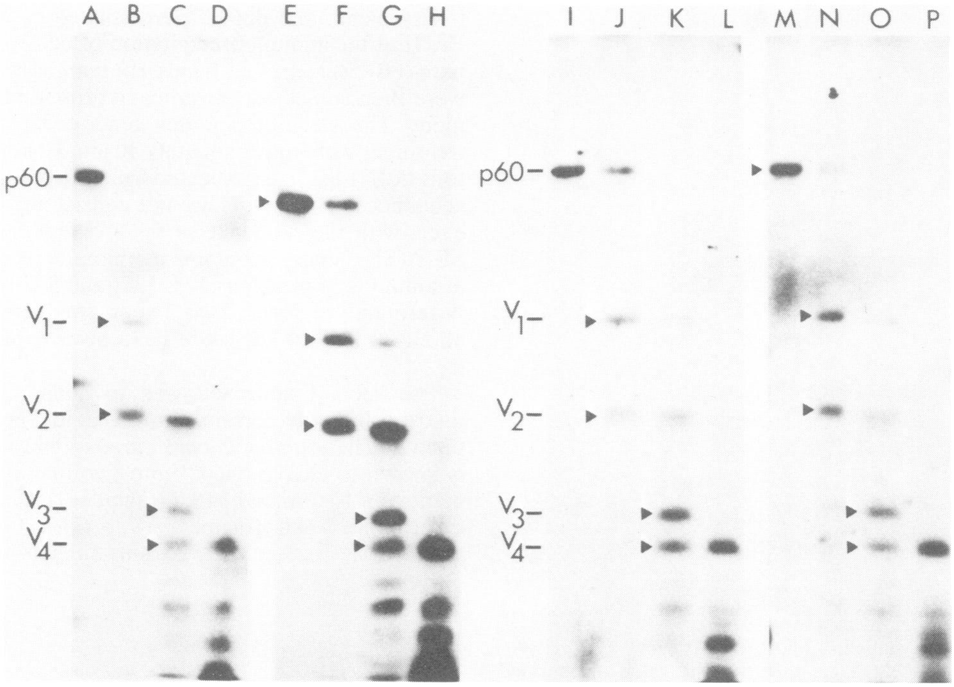


FIG. 6. Presence of [$^3$H]tyrosine and [$^3$H]phenylalanine in V3 and V4 fragments generated from rASV1441- and SR-A(NY)-pp60$^{src}$ after V8 protease digestion. Excised gel bands of pp60$^{src}$ from rASV1441 (lanes E to H and M to P) and SR-A(NY) (lanes A to D and I to L) labeled with [$^3$H]tyrosine (lanes A to H) or [$^3$H]phenylalanine (lanes I to P) were subjected to *S. aureus* V8 protease analysis as described in the text. The concentrations of V8 protease used were as follows: lanes A, E, I, and M, no enzyme; lanes B, F, J, and N, 0.2 μg/ml; lanes C, G, K, and O, 4 μg/ml; lanes D, H, L, and P, 80 μg/ml. The positions of V8 fragments V1, V2, V3, and V4 are indicated.

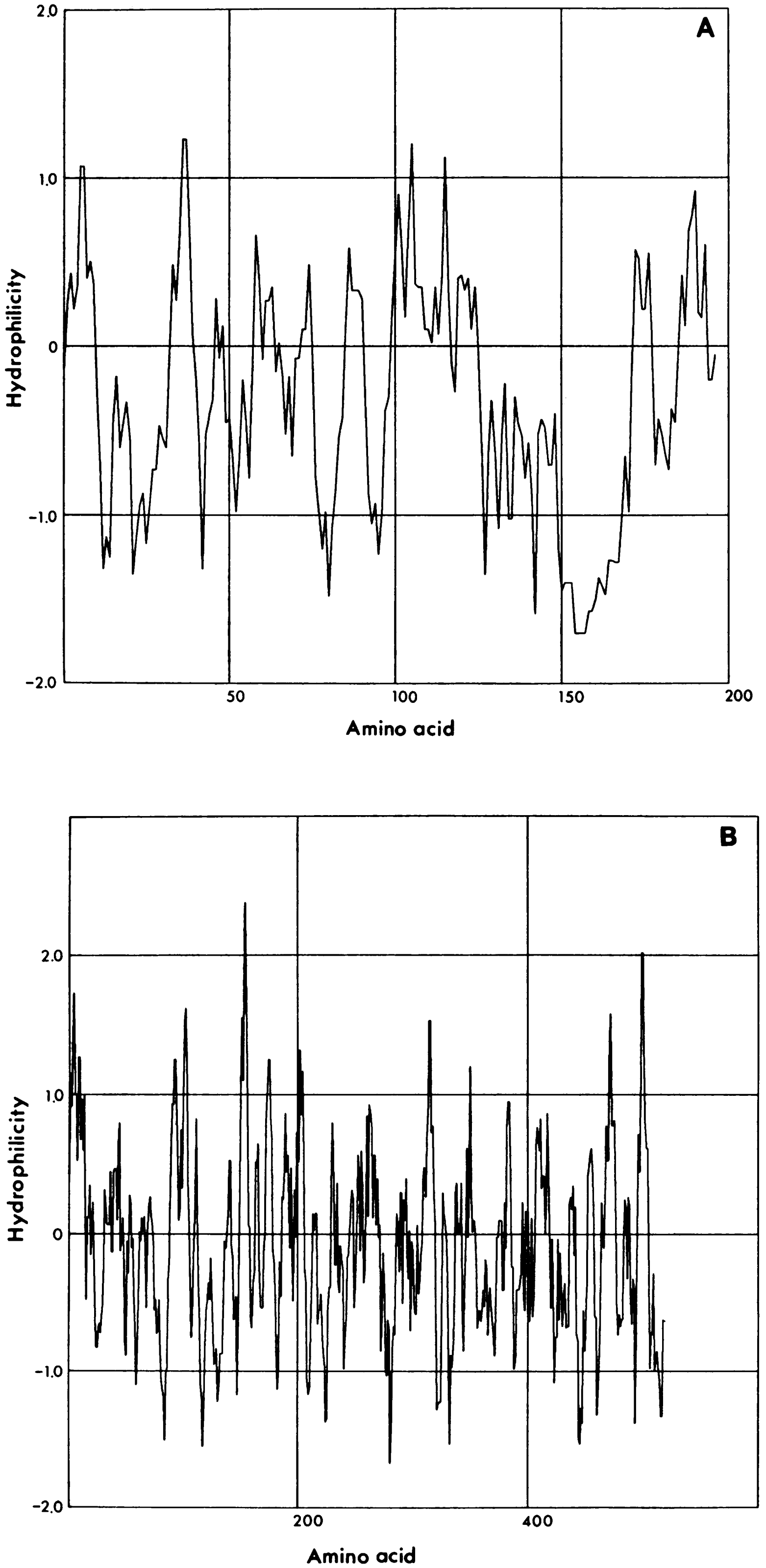
(V3)- and 16 (V4)-kd fragments, both of which maintain the N-terminal region of pp60$^{src}$ (17). Inspection of the amino acid sequence in Fig. 2 showed three Ser-Glu sequences at positions 158 to 159, 177 to 178, and 330 to 331, which we consider to represent V8 cleavage sites (discussed below). The sequence of rASV1441 pp60$^{src}$ contains both phenylalanine and tyrosine in the V3 and V4 fragments, whereas these amino acids are not present in the V3 and V4 fragments of SR-A(SF). Therefore, [$^3$H]phenylalanine- or [$^3$H]tyrosine-labeled pp60$^{src}$ of rASV1441 and SR-A(NY) were subjected to partial proteolytic analysis with V8 protease. The presence of labeled phenylalanine and tyrosine in V3 and V4 fragments was easily detectable (Fig. 6).

**Structural features of the *env* gene product.** The predicted structure of gp37 derived from the DNA sequence (Fig. 2) has several features characteristic of this envelope glycoprotein. The molecular weight of the nonglycosylated form of this protein was calculated to be 22,983. Overall, the protein is quite hydrophobic. Figure 7A shows the profile of hydrophilicity of the various regions of the protein. The most unique portion is the uninterrupted stretch of 27 hydrophobic amino acids from positions 150 Gly to 176 Ile. This structure is likely to be involved in the embedding of gp37 into the viral lipid membrane (13, 28). This hydrophobic region is followed by a relatively hydrophilic region at the C terminus (Fig. 2 and 7A).

The sequence of tripeptides, Asn-X-Thr or Asn-X-Ser, which are believed to be the sites for glycosylation (32), can be found at three places, positions 52 to 54, 100 to 102, and 181 to 183.

**Structural features of the *src* gene product. (i) General properties.** The predicted *src* protein (Fig. 2) has 526 amino acids, and its calculated molecular weight is 58,810. With few exceptions, each amino acid is distributed rather evenly in the pp60$^{src}$ molecule. As a notable exception, most of the methionine residues (11 of 12)

FIG. 7. Hydrophilicity analysis of gp37 and pp60$^{src}$. Hydrophilicity analysis of proteins was carried out on the predicted amino acid sequences of gp37 (A) and pp60$^{src}$ (B) by using the program described by Hopp and Woods (14). Hydrophilicity values for each amino acid obtained by Levitt (26) were used.


A
2.0
1.0
0
-1.0
-2.0
Hydrophilicity
50
100
150
200
Amino acid
B
2.0
1.0
0
-1.0
-2.0
Hydrophilicity
200
400
Amino acid

are located in the C-terminal half. About 65% of valine and 59% of leucine residues are also in the C-terminal half. On the other hand, 75% of serine and 68% of threonine residues are present in the N-terminal half. Overall, the protein contains more hydrophobic than hydrophilic amino acids. The hydrophilicity profile of $pp60^{src}$ plotted in the same way as for gp37 is shown in Fig. 7B. The profile shows that small stretches of hydrophobic and hydrophilic regions alternate along the sequence with no significant long stretch of either type. It is known that a peptide of about 8 kd located at the N-terminal end of $pp60^{src}$ is often cleaved off when this protein is immunoprecipitated from cell lysates (4, 22). This N-terminal domain has been implicated in the reported association of $pp60^{src}$ with cellular membranes (22, 24). The profile at the N-terminal end (Fig. 7B) shows some hydrophobic regions, but there is no long hydrophobic stretch typical of membrane proteins such as gp37. A 15-amino acid peptide from region 76 to 90 is consecutively hydrophobic. Whether this structure has any relevance to the association with membranes or to the derivation of a 52-kd peptide remains to be seen.

**(ii) Sites of cleavage by V8 protease.** As described above, *S. aureus* V8 protease cleaves $pp60^{src}$ into discrete fragments (3, 17). The V8 enzyme has been shown to cleave specifically peptide bonds at the carboxy-terminal side of either aspartic or glutamic acid residues (6). There are many Glu and Asp residues in $pp60^{src}$; however, only three Ser-Glu sites appear in the molecule (positions 159, 178, and 331). Cleavage at these positions would produce 34 (V1)-, 26 (V2)-, 18 (V3)-, and 16 (V4)-kd fragments. We have no further evidence that these sites are involved in the cleavage, but it seems unlikely that the above correlation is coincidental.

**(iii) Sites for phosphorylation.** The site for tyrosine phosphorylation of $pp60^{src}$ of SR-A(SF) has been identified at position 419 by analysis of the amino acid sequence of the phosphotyrosine-containing peptides (34). rASV1441 has the same amino acid sequence at that region, although the tyrosyl residue is at position 416. It is known that, when labeled in vivo, the tryptic peptide containing phosphotyrosine in $pp60^{v\text{-}src}$ and $pp60^{c\text{-}src}$ is different (2, 17, 34) and that the phosphotyrosine-containing peptide from rASV1441 shows the same mobility as that from SR-A(NY). This suggests that the 3′ end of the rASV1441 *src* gene, probably including the phosphotyrosine region, is derived from *td* virus rather than from c-*src*.

$pp60^{src}$ is phosphorylated at one serine residue by cyclic AMP-dependent protein kinase (3), and this site is located within the V4 fragment (3, 17). The amino acid sequences at such phosphorylation sites have been classified into two categories (21): (i) -Lys-Arg-X-X-Ser-X- and (ii) -Arg-Arg-X-Ser-X- (X indicates any amino acid, but usually hydrophobic amino acids are found at the sites next to serine). There are two -Arg-Arg-X-Ser-X- sequences in the V4 region of rASV1441 (at positions 17 and 158).

The phosphopeptides of $pp60^{src}$ have been analyzed by two-dimensional separation of the products of tryptic digestion (2, 17). By chromatography, the mobility of the phosphoserine peptide derived from SR-A was slower than that of the phosphotyrosine-containing peptide (17). However, Karess and Hanafusa (17) also found that a phosphoserine-containing peptide obtained from rASV1702, which was derived from another *td* mutant of SR-A, was less labeled in vivo and migrated faster than a phosphotyrosine-containing peptide. This particular rASV1702 was shown to encode a $pp60^{src}$, which is about 4,000 daltons shorter than the standard RSV $pp60^{src}$ due to a deletion near the amino terminus of the molecule. The difference in mobility of its phosphoserine-containing peptide was attributed to the alteration in the amino acid sequence at the amino terminus. These results seem to be compatible with the idea that position 17 is the phosphoserine site.

## ACKNOWLEDGMENTS

We thank P. Model, The Rockefeller University, for computer analysis of the hydrophilicity of gp37 and $pp60^{src}$; E. Hunter, The University of Alabama in Birmingham, for providing information concerning the amino acid sequence of gp37 before publication; G. Ju and A. M. Skalka, The Roche Institute of Molecular Biology, for providing the cloned viral DNAs; and T. Lerner, The Rockefeller University, for help in the glucosamine labeling experiment. We are indebted to L.-H. Wang, F. Cross, and T. Lerner for reading the manuscript.

This work was supported by Public Health Service grants CA14935 and CA18213 from the National Cancer Institute.

## LITERATURE CITED

1. **Brugge, J. S., and R. L. Erikson.** 1977. Identification of a transformation specific antigen induced by an avian sarcoma virus. Nature (London) **269:**346–348.
2. **Collett, M. S., J. S. Brugge, and R. L. Erikson.** 1978. Characterization of a normal avian cell protein related to the avian sarcoma virus transforming gene product. Cell **15:**1363–1369.
3. **Collett, M. S., E. Erikson, and R. L. Erikson.** 1979. Structural analysis of the avian sarcoma virus transforming protein: sites of phosphorylation. J. Virol. **29:**770–781.
4. **Courtneidge, S. A., A. D. Levinson, and J. M. Bishop.** 1980. The protein encoded by the transforming gene of avian sarcoma virus ($pp60^{src}$) and a homologous protein in normal cells ($pp60^{proto\text{-}src}$) are associated with the plasma membrane. Proc. Natl. Acad. Sci. U.S.A. **77:**3783–3787.
5. **Czernilofsky, A., A. Levinson, H. Varmus, J. M. Bishop, E. Tisher, and H. Goodman.** 1980. Nucleotide sequence of an avian sarcoma virus oncogene (*src*) and proposed amino acid sequence for gene product. Nature (London) **287:**198–203.
6. **Drapeau, G. R., Y. Boily, and J. Houmard.** 1972. Purification and properties of an extracellular protease of *Staphy-*

*lococcus aureus*. J. Biol. Chem. 247:6720–6726.

7. **Feldman, R. A., T. Hanafusa, and H. Hanafusa.** 1980. Characterization of protein kinase activity associated with the transforming gene product of Fujinami sarcoma virus. Cell 22:757–765.
8. **Gross, E.** 1967. The cyanogen bromide reaction. Methods Enzymol. 11:238–255.
9. **Gross, E., and B. Witkop.** 1961. Selective cleavage of the methionyl peptide bonds in ribonuclease with cyanogen bromide. J. Am. Chem. Soc. 83:1510–1511.
10. **Halpern, C. C., W. S. Hayward, and H. Hanafusa.** 1979. Characterization of some isolates of newly recovered avian sarcoma virus. J. Virol. 29:91–101.
11. **Hanafusa, H.** 1981. Cellular origin of transforming genes of RNA tumor viruses. Harvey Lect. 75:255–275.
12. **Hanafusa, H., C. C. Halpern, D. L. Buchhagen, and S. Kawai.** 1977. Recovery of avian sarcoma virus from tumors induced by transformation-defective mutants. J. Exp. Med. 146:1735–1747.
13. **Hardwick, J. M., and E. Hunter.** 1981. Rous sarcoma virus mutant LA3382 is defective in virion glycoprotein assembly. J. Virol. 40:752–761.
14. **Hopp, T. P., and K. R. Woods.** 1981. Prediction of protein antigenic determinants from amino acid sequences. Proc. Natl. Acad. Sci. U.S.A. 78:3824–3828.
15. **Joho, R. H., M. A. Billeter, and C. Weissmann.** 1975. Mapping and biological functions on RNA of avian tumor viruses: location of regions required for transformation and determination of host range. Proc. Natl. Acad. Sci. U.S.A. 72:4772–4776.
16. **Ju, G. L., and A. M. Skalka.** 1980. Nucleotide sequence analysis of the long terminal repeat (LTR) of avian retroviruses: structural similarities with transposable elements. Cell 22:379–386.
17. **Karess, R. E., and H. Hanafusa.** 1981. Viral and cellular *src* genes contribute to the structure of recovered avian sarcoma virus transforming protein. Cell 24:155–164.
18. **Karess, R. E., W. S. Hayward, and H. Hanafusa.** 1979. Cellular information in the genome of recovered avian sarcoma virus directs the synthesis of transforming protein. Proc. Natl. Acad. Sci. U.S.A. 76:3154–3158.
19. **Kawai, S., P. H. Duesberg, and H. Hanafusa.** 1977. Transformation-defective mutants of Rous sarcoma virus with *src* gene deletions of varying length. J. Virol. 24:910–914.
20. **Kawai, S., and H. Hanafusa.** 1972. Genetic recombination with avian tumor virus. Virology 49:37–44.
21. **Krebs, E. G., and J. A. Beavo.** 1979. Phosphorylation-dephosphorylation of enzymes. Annu. Rev. Biochem. 49:923–959.
22. **Krueger, J. G., E. Wang, and A. R. Goldberg.** 1980. Evidence that the *src* gene product of Rous sarcoma virus is membrane associated. Virology 101:25–40.
23. **Lerner, M. R., J. A. Boyle, S. M. Mount, S. L. Wolin, and J. A. Steitz.** 1980. Are snRNPs involved in splicing? Nature (London) 283:220–224.
24. **Levinson, A. D., S. A. Courtneidge, and J. M. Bishop.** 1981. Structural and functional domains of the Rous sarcoma virus transforming protein (pp60$^{src}$). Proc. Natl. Acad. Sci. U.S.A. 78:1624–1628.
25. **Levinson, A. D., and A. J. Levine.** 1977. Group C adenovirus tumor antigens: identification in infected and transformed cells and a peptide map analysis. Cell 11:871–879.
26. **Levitt, M.** 1976. A simplified representation of protein conformations for rapid simulation of protein folding. J. Mol. Biol. 104:59–107.
27. **Maxam, A. M., and W. Gilbert.** 1977. A new method for DNA sequence analysis. Proc. Natl. Acad. Sci. U.S.A. 74:560–564.
28. **Montelaro, R. C., S. J. Sullivan, and D. P. Bolognesi.** 1978. An analysis of type C retrovirus polypeptides and their associations in the virion. Virology 84:19–31.
29. **Purchio, A. F., E. Erikson, J. S. Brugge, and R. L. Erikson.** 1978. Identification of a polypeptide encoded by the avian sarcoma virus *src* gene. Proc. Natl. Acad. Sci. U.S.A. 75:1567–1571.
30. **Rettenmier, C. W., S. M. Anderson, M. W. Riemen, and H. Hanafusa.** 1979. *gag*-related polypeptides encoded by replication-defective avian oncoviruses. J. Virol. 32:749–761.
31. **Rubin, C. M., and C. W. Schmid.** 1980. Pyrimidine-specific chemical reactions useful for DNA sequencing. Nucleic Acids Res. 8:4613–4619.
32. **Schultz, A. M., S. M. Lockhart, E. M. Rabin and S. Oroszlan.** 1981. Structure of glycosylated and unglycosylated *gag* polyproteins of Rauscher murine leukemia virus: carbohydrate attachment sites. J. Virol. 38:581–592.
33. **Shank, P. R., S. H. Hughes, H. J. Kung, J. E. Majors, N. Quintrell, R. V. Guntaka, J. M. Bishop, and H. E. Varmus.** 1978. Mapping unintegrated avian sarcoma virus DNA: termini of linear DNA bear 300 nucleotides present once or twice in two species of circular DNA. Cell 15:1383–1395.
34. **Smart, J. E., H. Oppermann, A. P. Czernilofsky, A. F. Purchio, R. L. Erikson, and J. M. Bishop.** 1981. Characterization of sites for tyrosine phosphorylation in the transforming protein of Rous sarcoma virus (pp60$^{v\text{-}src}$) and its normal cellular homologue (pp60$^{c\text{-}src}$). Proc. Natl. Acad. Sci. U.S.A. 78:6013–6017.
35. **Swanstrom, R., W. J. DeLorbe, J. M. Bishop, and H. E. Varmus.** 1981. Nucleotide sequence of cloned unintegrated avian sarcoma virus DNA: viral DNA contains direct and inverted repeats similar to those in transposable elements. Proc. Natl. Acad. Sci. U.S.A. 78:124–128.
36. **Swanstrom, R., H. E. Varmus, and J. M. Bishop.** 1982. Nucleotide sequence of the 5′ noncoding region and part of the *gag* gene of Rous sarcoma virus. J. Virol. 41:535–541.
37. **Takeya, T., and H. Hanafusa.** 1982. DNA sequence of the viral and cellular *src* genes of chickens. II. Comparison of the *src* genes of two strains of avian sarcoma virus and of the cellular homolog. J. Virol. 44:12–18.
38. **Takeya, T., H. Hanafusa, R. P. Junghans, G. Ju, and A. M. Skalka.** 1981. Comparison between the viral transforming gene (*src*) of recovered avian sarcoma virus and its cellular homolog. Mol. Cell. Biol. 1:1024–1037.
39. **Tu, C.-P. D., and S. N. Cohen.** 1980. 3′-End labeling of DNA with [$\alpha$-$^{32}$P] cordycepin-5′-triphosphates. Gene 10:177–183.
40. **Vigne, R., M. L. Breitman, C. Moscovici, and P. K. Vogt.** 1979. Restitution of fibroblast-transforming ability in *src* deletion mutants of avian sarcoma virus during animal passage. Virology 93:413–426.
41. **Wang, L.-H., P. H. Duesberg, K. Beemon, and P. K. Vogt.** 1975. Mapping RNase T$_1$-resistant oligonucleotides of avian tumor virus RNAs: sarcoma-specific oligonucleotides are near the poly(A) end and oligonucleotides common to sarcoma and transformation-defective viruses are at the poly(A) end. J. Virol. 16:1051–1070.
42. **Wang, L.-H., C. C. Halpern, M. Nadel, and H. Hanafusa.** 1978. Recombination between viral and cellular sequences generates transforming sarcoma viruses. Proc. Natl. Acad. Sci. U.S.A. 75:5812–5816.
43. **Wang, L.-H., C. Moscovici, R. E. Karess, and H. Hanafusa.** 1979. Analysis of the *src* gene of sarcoma viruses generated by recombination between transformation-defective mutants and quail cellular sequences. J. Virol. 32:546–556.
44. **Wang, L.-H., P. Snyder, T. Hanafusa, and H. Hanafusa.** 1980. Evidence for the common origin of viral and cellular sequences involved in sarcomagenic transformation. J. Virol. 35:52–64.