Proc. Natl. Acad. Sci. USA
Vol. 79, pp. 3179–3182, May 1982
Biochemistry

Nucleotide sequence of the transforming gene of simian sarcoma virus

(primate retrovirus/v-*sis* nucleotide sequence/open reading frame)

SUSHILKUMAR G. DEVARE, E. PREMKUMAR REDDY, KEITH C. ROBBINS, PHILIP R. ANDERSEN, STEVEN R. TRONICK, AND STUART A. AARONSON

Laboratory of Cellular and Molecular Biology, National Cancer Institute, Bethesda, Maryland 20205

Communicated by Robert J. Huebner, March 1, 1982

ABSTRACT **The sequence of the transforming region of simian sarcoma virus (SSV) has been determined by using molecularly cloned viral DNA. This region encompassed the 1.0-kilobase pair woolly monkey cell-derived insertion sequence, v-*sis*, and flanking simian sarcoma-associated viral (SSAV) sequences. A 675-nucleotide-long open reading frame commenced 19 nucleotides within the SSAV sequences to the left of the v-*sis* helper viral junction and terminated within v-*sis* itself. Possible promoter and acceptor splice signals were detected in helper viral sequences upstream from this open reading frame, and potential polyadenylylation sites were identified downstream both within v-*sis* and in helper viral sequences beyond v-*sis*. The recombinational event that led to the generation of SSV occurred in the middle of two functional codons, indicating that SSAV provided the regulatory elements for transcription as well as the initiation codon for translation of SSV cell-derived transforming sequences.**

Retroviruses that transform cells in tissue culture and induce solid tumors *in vivo* have been isolated from a number of vertebrate species. The only known sarcoma virus of primate origin was initially isolated from a naturally occurring tumor of a woolly monkey (1). The advent of recombinant DNA techniques has recently made it possible to clone in biologically active form the integrated genome of this virus, designated simian sarcoma virus (SSV) (2). The full-length linear 5.2-kilobase pair (kbp) SSV genome has been found to contain a 1.0-kbp segment of helper virus-unrelated information localized toward the 3′ end with respect to SSV RNA (2, 3). This information, designated v-*sis* according to recent convention (4), is well conserved at low copy number within mammalian cellular DNAs. Moreover, its high extent of hybridization and base pair matching with woolly monkey DNA have established that v-*sis* originated from within the woolly monkey genome (5, 6).

Studies to date have indicated that the cell-derived sequences of transforming retroviruses are responsible for their transforming function(s) (for recent reviews, see refs. 7 and 8). Utilizing molecular clones of subgenomic fragments of SSV DNA, we have shown that the v-*sis* sequences are essential for SSV biologic activity (unpublished results). In an attempt to better understand the structural organization and possible molecular mechanisms involved in transformation by SSV, we have undertaken primary DNA sequence analysis of the transforming region of the molecule. Putative regulatory signals for transcription, RNA processing, and translation of v-*sis* sequences have been identified. Sequence analysis has also demonstrated the occurrence of a long open reading frame within the v-*sis* region that could code for the SSV transforming protein.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

MATERIALS AND METHODS

Molecular Cloning. The integrated proviral SSV was initially cloned in Charon 16A λ phage (2). In the present studies a pBR322 subclone of the 5.8-kbp clone containing SSV DNA and host flanking sequences was utilized. The unintegrated form of helper simian sarcoma-associated virus (SSAV) DNA was isolated by the Hirt procedure (9) from a line of NRK cells infected 48 hr earlier with SSAV. Preliminary restriction enzyme analysis revealed that *Eco*RI cleaved SSAV DNA once. The circular SSAV DNA molecule was thus cut with *Eco*RI, molecularly cloned in Charon 16A λ phage, and subcloned in pBR322. The 5.8-kbp SSV and 8.8-kbp SSAV DNA inserts from plasmid DNAs were purified by agarose gel electrophoresis and DEAE-cellulose (DE-52, Whatman) column chromatography after cleavage with *Eco*RI for use in all subsequent analyses.

Nucleotide Sequence Analysis. Appropriate restriction fragments were labeled either at their 5′ end by using [γ-^{32}P]ATP (Amersham, 3,000 Ci/mmol; 1 Ci = 3.7×10^{10} becquerels) and polynucleotide kinase (P-L Biochemicals) as described by Maxam and Gilbert (10) or at their 3′ end by using cordycepin 5′-[α-^{32}P]triphosphate (Amersham, 3,000 Ci/mmol) and terminal deoxynucleotidyltransferase (P-L Biochemicals) according to Roychoudhury *et al.* (11). End-labeled DNA fragments were digested with appropriate restriction endonucleases (New England BioLabs) and isolated by agarose or polyacrylamide gel electrophoresis. The nucleotide sequence was determined by the procedure of Maxam and Gilbert (10).

RESULTS

Strategy for Determining the Sequence of the SSV Transforming Region. The restriction map of molecularly cloned SSV DNA was constructed by using both double-digestion analysis and the partial digestion technique of Smith and Birnstiel (12). Fig. 1 shows the restriction map of SSV DNA and the localization of its cell-derived sequence (v-*sis*). In the present studies, primary DNA sequence analysis was performed on an SSV subgenomic fragment of about 2.3 kbp, encompassing v-*sis* and its adjacent region (Fig. 1). A molecular clone of this subgenomic fragment contained sufficient information for SSV transforming activity (unpublished observations). The strategy used for determining the sequence of the transforming region is also indicated. The sequences of both DNA strands were determined, and all restriction cleavage sites were confirmed by sequence analysis.

Abbreviations: SSV, simian sarcoma virus; SSAV, simian sarcoma-associated virus; kbp, kilobase pair(s); LTR, long terminal repeat.

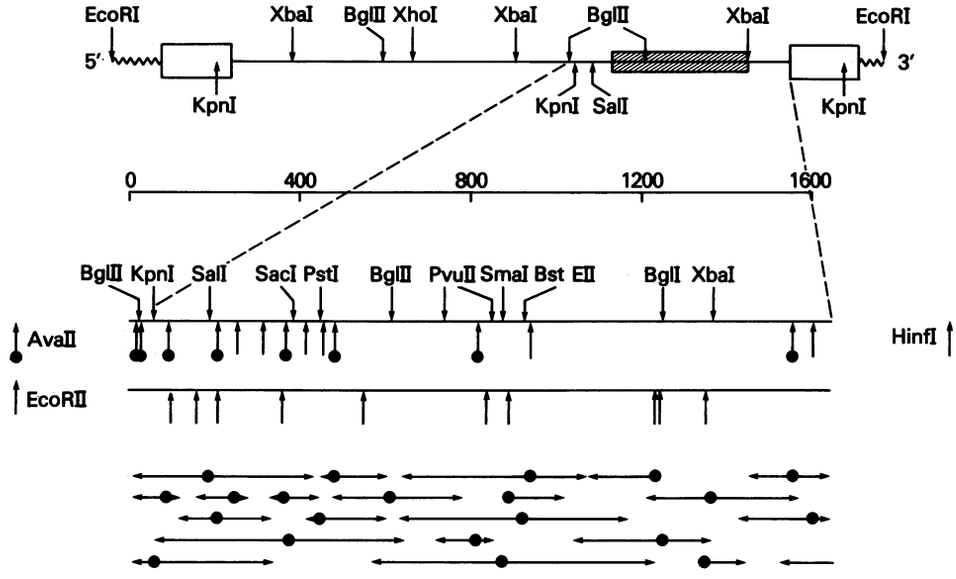

FIG. 1. Restriction enzyme map and strategy for sequencing the transforming region of the SSV genome. The hatched bar indicates v-*sis*, the empty bars indicate long terminal repeats (LTRs), and the wavy line indicates host DNA. The scale is in base pairs and refers to the material below it. The sequence of the genome was determined with the use of the restriction sites indicated on the map. The ^{32}P-labeled end of each fragment is indicated by the filled circle and the extent and direction of sequencing are indicated by arrows.

Heteroduplex analysis has revealed that the SSV genome has undergone a sequence substitution of approximately 1.0 kbp in the region coding for the envelope protein, where *env* gene sequences have been replaced by v-*sis* (2, 3). In order to localize the exact point of this recombinational event, we determined the sequence of the corresponding regions of the SSAV genome. It was, thus, possible to localize the junction points between the v-*sis* and helper viral sequences (Fig. 2). It should be noted that at the right junction, the recombinational event generated an *Xba* I site in SSV DNA. This *Xba* I site was absent from the analogous position within the SSAV genome (data not shown).

Sequence Organization of the v-*sis* Region and Flanking Helper Viral Sequences. Examination of the viral RNA strand of the v-*sis* region of SSV (Fig. 2) revealed an open reading frame starting with the initiation codon ATG at position 352 and terminating with an ochre codon TAA at position 1027. This stretch of 675 nucleotides began with 19 bases of the helper viral information at the left and terminated within the v-*sis* region. Analysis of the two alternative reading frames did not reveal the occurrence of any long open reading frame. To the left of the initiator codon ATG, the open reading frame extended beyond the *Bgl* II site at position 26. However, because a molecularly cloned subgenomic fragment from *Sal* I to *Eco*RI restriction sites had sufficient information for SSV transforming activity (unpublished observations), the ATG at position 352 is the most likely initiator codon for the putative SSV transforming protein. This open reading frame would have a coding capacity for a protein of 225 amino acids (27,000 daltons).

The putative transforming protein contained a high percentage of charged amino acids (26.6%) distributed randomly throughout the molecule, indicating its hydrophilic nature. Moreover, there were disproportionately high numbers of alanine, leucine, and arginine residues. There was no evidence of characteristic transmembrane-specific amino acid sequences (13) at either terminus. The central region, however, did contain a stretch of hydrophobic residues with a chain of 14 amino acids without charge (at position 91–104 from the amino terminus), suggesting that this region may be conformationally buried within the molecule. The sequence Asn-Met-Thr was detected in the molecule at position 48–50 from the amino terminus of the protein. This sequence is known to serve as a recognition site for carbohydrate addition (14) via the dolichol phosphate N-linked pathway (15).

Possible Promoter and Splicing Signals for Transcription of the SSV Transforming Gene. It is known that spliced mRNAs are utilized by retroviruses for synthesis of their envelope glycoproteins (16, 17). Recently, sequence analysis of Moloney murine sarcoma virus and Moloney murine leukemia virus has provided evidence that splicing may be involved in the formation of functional messages for *gag* and *pol* genes as well (18, 19). Upstream from the initiator codon for the putative transforming protein, we identified four putative splice acceptor sites at positions 89, 232, 295, and 347, consisting of a pyrimidine-rich nucleotide track followed by a dinucleotide A-G (20). Thus one or more of these splice acceptor sites may play a role in the generation of the functional mRNA for the putative SSV transforming protein. Moreover, when we determined the sequence of the region of the SSV genome after the LTR, a characteristic donor splice site with the sequence G-A-G-G-T-A-A-G was identified 58 nucleotides downstream from the left LTR (Fig. 3). In addition, analysis of SSAV sequences to the left of v-*sis* revealed an A+T-rich region at position 331–337. The sequence A-A-T-A-A-A-A is similar to the sequence of eukaryotic transcriptional promoters previously identified (21, 22). Whether this is a functional promoter for the transcription of the SSV transforming gene remains to be determined.

Because all known retroviral mRNAs are polyadenylylated, we searched for the occurrence of polyadenylylation signals beyond the open reading frame. This analysis revealed two putative polyadenylylation signals at position 1,160 within v-*sis* and at position 1,533 in helper viral sequence. However, the typical C-A site located 16 nucleotides downstream from polyadenylylation signals was not detected in either case. Nucleotide sequence analysis of the LTR to the right of v-*sis* revealed a polyadenylylation signal (A-A-T-A-A-A) at position 2,057, which was followed by a C-A at position 2,078. Thus, the configuration of the polyadenylylation signal within the LTR suggests that this may be the preferred site for termination of transcription of the mRNA coding for the putative SSV transforming protein.

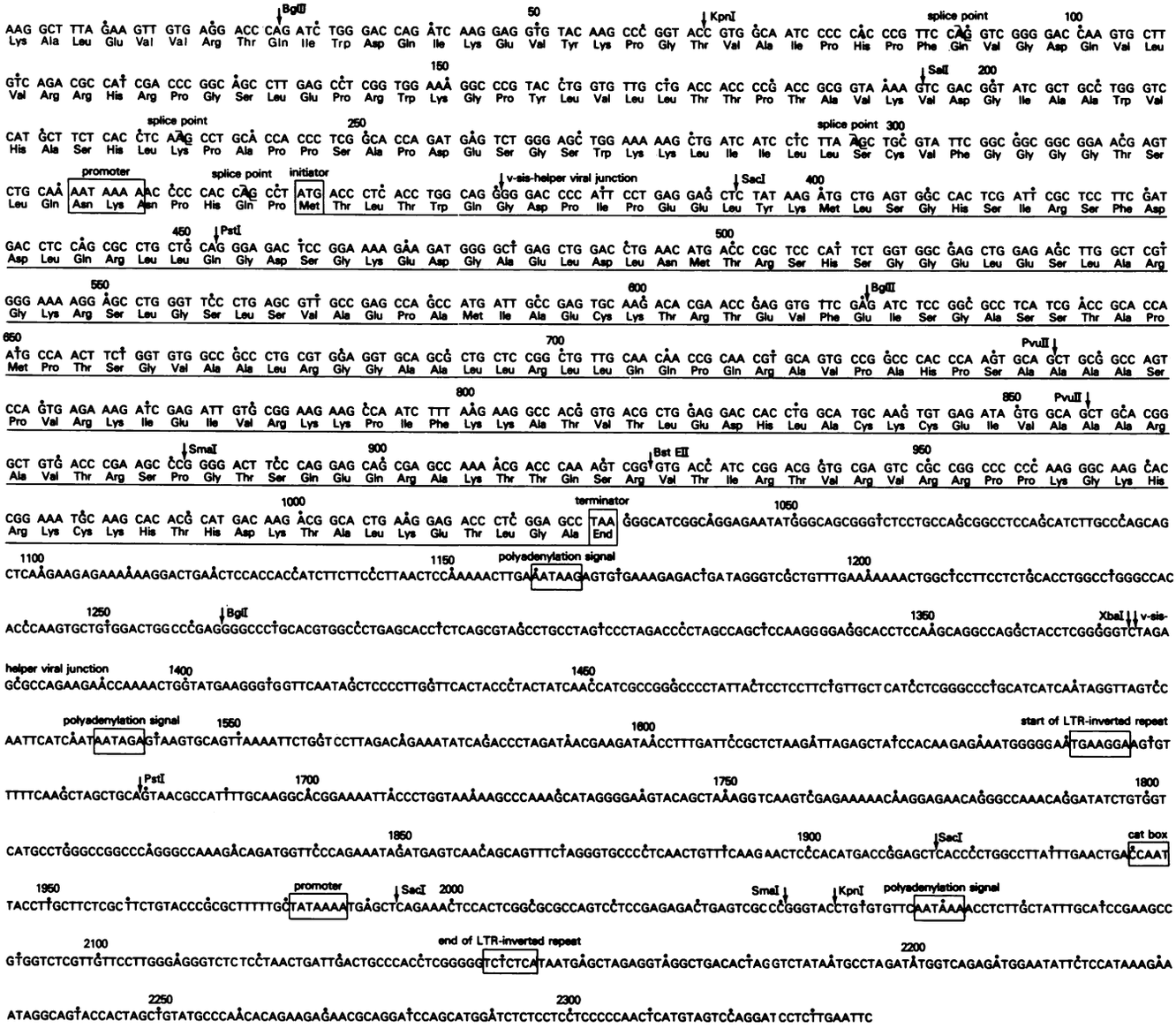

FIG. 2. Nucleotide sequence of the SSV transforming region. The upper line shows the sequence proceeding in the 5′-to-3′ direction and has the same polarity as SSV genomic RNA. Dots mark every 10th nucleotide. The amino acid sequence deduced from the open reading frame is given in the bottom line. The major structural features observed are indicated.

DISCUSSION

Nucleotide sequence analysis of the SSV transforming region has revealed several important features of its molecular organization. Examination of the sequence of the transforming region revealed a single open reading frame on the viral RNA strand. This frame, 675 bases long, could code for a protein of 225 amino acids with a molecular weight of around 27,000.

A number of transforming retroviruses appear to synthesize their transforming proteins by means of a *gag*-X polyprotein, the amino-terminal region of which is composed of helper virus *gag* gene products (4). Thus, these transforming proteins utilize helper viral sequences as initiators for their synthesis. It is not yet known whether the *gag* gene regions of such molecules also play an important role in the transforming functions of these hybrid proteins. Although some SSV variants or DNA molecules do express a portion of the helper viral *gag* gene product, the SSV transforming protein is not coded by means of a *gag*-X polyprotein (ref. 23; unpublished observations.) By comparison of our sequence data for SSV with that of SSAV, the open reading frame for the putative SSV transforming protein was shown to commence 19 bases within helper viral sequences, to the left of the helper viral v-*sis* junction. This region of helper virus corresponds to a location well beyond SSAV *gag* gene coding sequences (2). Analogous findings have recently been reported for Moloney murine sarcoma virus, for which the first five amino acids of its putative transforming protein are contributed by helper viral sequences (18, 24, 25). Thus, it may be a general phenomenon that the transforming gene products of retroviruses are hybrid proteins in which the helper virus contributes sequences that code for the initiation of the protein.

Nucleotide sequence analysis defined possible splicing acceptor sites near the beginning of the large open reading frame that could be used in processing such a message. Alternatively, the promoter-like sequences identified to the left of this open reading frame could serve directly to promote the message. We have been able to show, utilizing molecular clones of subgenomic fragments of SSV DNA, that the SSV transforming gene can function in the absence of the 5′ LTR. However, the trans-

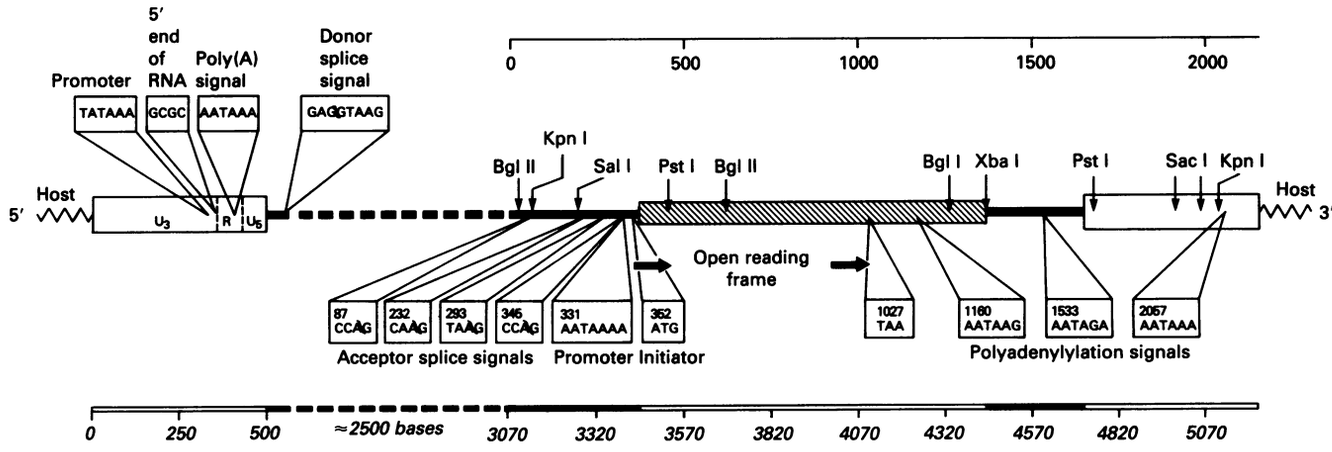

FIG. 3. Summary of the major structural features of the SSV genome. The scales are in base pairs. Important features of SSV genome, including the long open reading frame, possible signals for promoter and polyadenylylation, and donor and acceptor splice signals, are illustrated. Nucleotide numbers obtained from the sequence in Fig. 2 are provided. The number in italics at the bottom represents the approximate nucleotide location in the complete SSV genome.

forming efficiency of such subgenomic fragments is substantially reduced compared to that of the complete molecule (unpublished observations). Thus, further studies will be necessary to elucidate the molecular mechanisms normally involved in transcription of the SSV transforming gene.

Our sequence analysis suggests that the putative SSV transforming protein is hydrophilic in nature with no transmembrane-specific amino acid sequences at either terminus. By molecular hybridization, the SSV transforming gene does not show detectable homology with any of a number of other molecularly cloned retroviral *onc* genes (5, 6). Moreover, comparison of the amino acid sequence of the putative SSV transforming gene product with those of Moloney murine sarcoma virus (18, 24, 25) and Rous sarcoma virus (26) has not yielded evidence of any significant homologies. Comparison of the SSV transforming protein with more than 1,600 known gene products by computer analysis has also failed to show any significant relationships (unpublished observations). Thus, the identity of the SSV transforming protein remains to be determined.

As yet, there have been no reports of antisera capable of detecting a SSV transforming protein. A recent approach for identification of proteins for which nucleotide sequencing data is available involves the use of small peptides synthesized on the basis of the predicted sequence as haptens to elicit immune responses (27, 28). Knowledge of the nucleotide sequence for the putative SSV transforming protein should make it possible to obtain antisera directed against synthetic SSV peptides. Such antisera would be useful in the search for the functional SSV transforming protein in SSV-transformed cells.

We thank Dr. Margaret Dayhoff for help in computer analysis and J. Doria Law and Dace G. Klimanis for excellent technical assistance.

1. Thielen, G. J., Gould, D., Fowler, M. & Dungworth, D. L. (1971) *J. Natl. Cancer Inst.* **47**, 881–889.
2. Robbins, K. C., Devare, S. G. & Aaronson, S. A. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 2918–2922.
3. Gelmann, E. P., Wong-Staal, F., Kramer, R. A. & Gallo, R. C. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 3373–3377.
4. Coffin, J. M., Varmus, H. E., Bishop, J. M., Essex, M., Hardy, W. D., Jr., Martin, G. S., Rosenberg, N. E., Scolnick, E. M., Weinberg, R. A. & Vogt, P. K. (1981) *J. Virol.* **40**, 953–957.
5. Wong-Staal, F., Favera, R. D., Gelmann, E. P., Manzari, V., Szala, S., Josephs, S. F. & Gallo, R. C. (1981) *Nature (London)* **294**, 273–275.
6. Robbins, K. C., Hill, R. L. & Aaronson, S. A. (1982) *J. Virol.* **41**, 721–725.
7. Bishop, J. M. (1978) *Annu. Rev. Biochem.* **47**, 35–88.
8. Shih, T. Y. & Scolnick, E. M. (1980) in *Viral Oncology*, ed. Klein, G. (Raven, New York), pp. 135–160.
9. Hirt, B. (1967) *J. Mol. Biol.* **26**, 365–369.
10. Maxam, A. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 560–564.
11. Roychoudhury, R. & Wu, R. (1980) *Methods Enzymol.* **65**, 43–62.
12. Smith, H. O. & Birnstiel, M. L. (1976) *Nucleic Acids Res.* **3**, 2387–2398.
13. Segrest, J. P., Kahane, I., Jackson, R. L. & Marchesi, V. T. (1973) *Arch. Biochem. Biophys.* **155**, 167–183.
14. Tucker, P. W., Liu, C. P., Mushinski, J. F. & Blattner, F. R. (1980) *Science* **209**, 1353–1360.
15. Robbins, P. W., Hubbard, S. C., Turco, S. J. & Wirth, D. F. (1977) *Cell* **12**, 893–900.
16. Mellon, P. & Duesberg, P. H. (1977) *Nature (London)* **270**, 631–634.
17. Rothenberg, E., Donoghue, D. J. & Baltimore, D. (1978) *Cell* **13**, 435–451.
18. Reddy, E. P., Smith, M. J. & Aaronson, S. A. (1981) *Science* **214**, 445–450.
19. Shinnick, T. M., Lerner, R. A. & Sutcliffe, J. G. (1981) *Nature (London)* **293**, 543–548.
20. Seif, I., Khoury, G. & Dhar, R. (1979) *Nucleic Acids Res.* **6**, 3387–3398.
21. Pribnow, D. (1975) *Proc. Natl. Acad. Sci. USA* **72**, 784–788.
22. Rosenberg, M. & Court, D. (1979) *Annu. Rev. Genet.* **13**, 319–353.
23. Aaronson, S. A., Stephenson, J. R., Hino, S. & Tronick, S. R. (1975) *J. Virol.* **16**, 1117–1123.
24. Reddy, E. P., Smith, M. J., Canaani, E., Robbins, K. C., Tronick, S. R., Zain, S. & Aaronson, S. A. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 5234–5238.
25. Van Beveren, C., Galleshaw, J. A., Jonas, V., Berns, A. J. M., Doolittle, R. F., Donoghue, D. J. & Verma, I. (1981) *Nature (London)* **289**, 258–262.
26. Czernilofsky, A., Levinson, A., Varmus, H., Bishop, J. M., Tischer, E. & Goodman, H. (1980) *Nature (London)* **287**, 198–203.
27. Walter, G., Scheidtmann, K. H., Carbone, A., Laudano, A. P. & Doolittle, R. F. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 5197–5200.
28. Sutcliffe, J. G., Shinnick, T. M., Green, N., Liu, F. T., Niman, H. L. & Lerner, R. A. (1980) *Nature (London)* **287**, 801–805.