JOURNAL OF VIROLOGY, Dec. 1984, p. 991–994
0022-538X/84/120991-04$02.00/0

Vol. 52, No. 3

Sequence Analysis of the Respiratory Syncytial Virus Phosphoprotein Gene

MASANOBU SATAKE, NARAYANSAMY ELANGO, AND SUNDARARAJAN VENKATESAN*

Laboratory of Infectious Diseases, National Institute of Allergy and Infectious Diseases, Bethesda, Maryland 20205

Received 23 April 1984/Accepted 22 August 1984

A recombinant cDNA plasmid ($pRSA_3$) containing an almost full-length copy of the mRNA encoding respiratory syncytial virus phosphoprotein was identified in a cDNA library prepared with mRNA from respiratory syncytial virus-infected cells. The cDNA insert was sequenced, and a protein of 27,150 daltons was deduced from the DNA sequence. The protein is relatively acidic, containing two clusters of acidic amino acids, one in the middle of the molecule and the other at the C-terminus. It is devoid of both cysteine and tryptophan. There was no other potential reading frame within the phosphoprotein gene of respiratory syncytial virus. This situation is unlike that with Sendai virus, a paramyxovirus, which has a nonstructural C protein encoded by a second overlapping reading frame near the 5′ end of the mRNA for phosphoprotein.

Respiratory syncytial (RS) virus, a pleomorphic, enveloped, unsegmented, negative-strand RNA virus, contains an encapsidated genomic RNA of 5×10^6 daltons (16). It is similar in many properties to paramyxoviruses but has been placed in a separate genus, *Pneumovirus*, based on morphological differences and lack of hemagglutinin and neuraminidase activities. It has been shown to encode 9 or 10 polyadenylic acid-containing [poly(A)] mRNAs coding for 9 or 10 virus-associated proteins (5, 6) consistent with the presence of seven or eight nonoverlapping complementation groups based on genetic analysis. The viral proteins include nucleocapsid protein (NP, 51 kilodaltons [kd]), phosphoprotein (P, 33 kd), and a large protein (L, presumably the polymerase of 200 kd) associated with the viral cores and glycoproteins of 84 (G) and 68 (F_o) kd and a matrix protein (M, 28 kd) associated with the viral envelope (1, 21, 25, 29, 32). In addition, two or three smaller viral proteins referred to as nonstructural proteins are present in infected cells (5, 17, 29). Transcriptionally active nucleocapsids of paramyxoviruses contain, in addition to the genomic RNA, NP, P, and L (18). In this respect they resemble vesicular stomatitis virus nucleocapsids, which have the nucleocapsid (N) protein, a phosphorylated nonstructural (NS) protein, and L. The NS protein of vesicular stomatitis virus serves as a transcriptional accessory factor and may be the equivalent of paramyxovirus P (30). In vitro transcriptional studies with paramyxovirus nucleocapsids have tentatively assigned certain roles for the proteins associated with nucleocapsids. In particular, certain putative functional domains of P involved in transcription have been mapped (3, 4, 15). Although RS virus nucleocapsids have not been demonstrated to be transcriptionally active, presumably because optimal conditions have not been worked out, their composition is similar to the paramyxovirus nucleocapsids. During paramyxovirus viral infection certain small nonstructural viral proteins are expressed. The in vivo role of the paramyxovirus nonstructural proteins is not clear presently. By analogy with the suggested roles of influenza virus NS_1 and NS_2 it might be predicted that these proteins have a role in viral replication or cell death or both. Interestingly, the Sendai virus nonstructural C protein is temporally regulated during the life cycle of the virus (20). A similar situation apparently exists for Newcastle disease virus which encodes a nonstructural protein referred to as the S protein (7). Both of these nonstructural proteins are derived from the mRNA encoding the P of the respective viruses (7, 10, 12, 28) as are the nonstructural proteins of mumps virus (14) and measles virus (R. Bellini, personal communication). We wished to determine whether RS virus might have a similar organization of its P. Towards this goal, we have sequenced a cDNA insert of a recombinant plasmid harboring the P gene. The details of the protein sequence deduced from the cDNA sequence and its biological significance are discussed.

In an attempt to analyze the different transcriptional units of RS virus, a library of recombinant cDNA plasmids containing sequences of poly(A)-containing RNAs from infected cells was prepared. Single-stranded cDNAs synthesized with oligodeoxythymidylic acid as a primer were electrophoretically resolved on alkali agarose gels into different size classes. They were used separately for synthesizing double-stranded cDNAs. Our strategy described previously (29) avoided the traditional self-priming reaction for second-strand synthesis that requires S1 nuclease to remove the resulting hairpin. The double-stranded cDNAs were tailed with 20 oligodeoxycytidylic acid residues and inserted into oligodeoxyguanylate tailed pBR322 before transformation of *Escherichia coli* HB101. The resulting cDNA library was screened by several procedures as described previously (29). In particular, recombinants DNAs were used to hybrid select mRNA from the infected cells. Five recombinants ($pRSA_3$, $pRSC_1$, $pRSC_5$, $pRSC_9$, and $pRSC_{13}$) hybrid selected mRNA that yielded on translation a protein which comigrated with authentic viral P and was indistinguishable from the viral P by two-dimensional tryptic fingerprinting (29). They were selected for limited DNA sequencing. Of these, only $pRSA_3$ had a poly(A) tail representing the 3′ end of the template mRNA, whereas the others, although sharing extensive sequence homology with $pRSA_3$, lacked variable DNA stretches corresponding to the 3′ end of the mRNA. Therefore, $pRSA_3$ was chosen for complete DNA sequencing by the chemical method of Maxam and Gilbert (23). The relevant restriction map and sequencing strategy are illustrated in Fig. 1.

A RS viral cDNA sequence of 916 nucleotides is presented (Fig. 2) in the messenger sense. This cDNA insert of 916 nucleotides hybridized to a single viral mRNA species of

* Corresponding author.

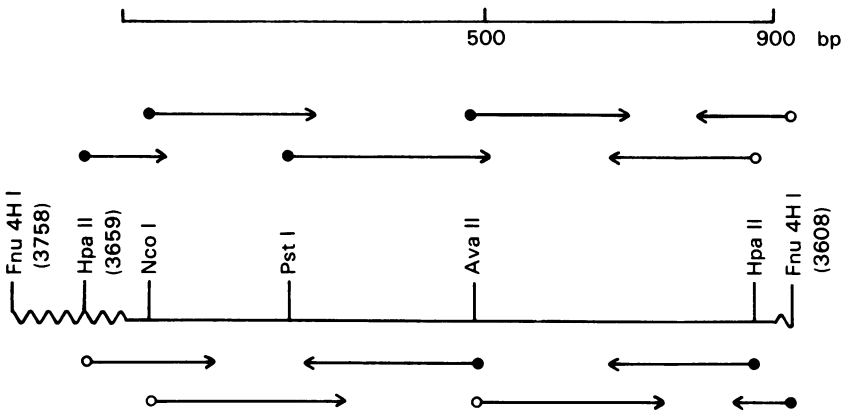

FIG. 1. Restriction map and sequencing strategy. The straight line denotes the cloned RS viral P gene sequence. The wavy lines on either side represent the pBR322 sequence and the oligodeoxyguanylate tails. The length of the sequence in bp is illustrated at the top. The arrows with open or closed circles indicate the different restriction sites of the DNA fragments labeled asymmetrically at the 5′ or 3′ end and used for DNA sequencing. The arrowheads denote the extent of DNA sequence determination.

equivalent size by blotting and filter hybridization (data not shown). Earlier we showed that cDNA recombinants of the viral NP and M genes reacted with discrete mRNA species (9, 27). However, we wished to determine whether the entire 5′ sequence of the mRNA was preserved in $pRSA_3$. A *Hin*fI-*Eco*RI fragment 5′ labeled at the *Hin*fI site (antimessage strand) was hybridized to poly(A) RNA from infected cells. The RNA-DNA hybrid was recovered, and the DNA primer was extended on the RNA template with reverse transcriptase. The primer extension product(s) was analyzed by polyacrylamide gel electrophoresis under denaturing conditions. A DNA sequence ladder derived from the *Hin*fI-

```
CAAATAAATCATC

                                        10                                  20
ATG GAA AAG TTT GCT CCT GAA TTC CAT GGA GAA GAT GCA AAC AAC AGG GCT ACT AAA TTC
MET GLU LYS PHE ALA PRO GLU PHE HIS GLY GLU ASP ALA ASN ASN ARG ALA THR LYS PHE
                                        30                                  40
CTA GAA TCA ATA AAG GGC AAA TTC ACA TCA CCC AAA GAT CCC AAG AAA AAA GAT AGT ATC
LEU GLU SER ILE LYS GLY LYS PHE THR SER PRO LYS ASP PRO LYS LYS LYS ASP SER ILE
                                        50                                  60
ATA TCT GTC AAC TCA ATA GAT ATA GAA GTA ACC AAA GAA AGC CCT ATA ACA TCA AAT TCA
ILE SER VAL ASN SER ILE ASP ILE GLU VAL THR LYS GLU SER PRO ILE THR SER ASN SER
                                        70                                  80
ACT ATT ATC AAC CCA ACA AAT GAG ACA GAT GAT ACT GCA GGG AAC AAG CCC AAT TAT CAA
THR ILE ILE ASN PRO THR ASN GLU THR ASP ASP THR ALA GLY ASN LYS PRO ASN TYR GLN
                                        90                                 100
AGA AAA CCT CTA GTA AGT TTC AAA GAA GAC CCT ACA CCA AGT GAT AAT CCC TTT TCT AAA
ARG LYS PRO LEU VAL SER PHE LYS GLU ASP PRO THR PRO SER ASP ASN PRO PHE SER LYS
                                       110                                 120
CTA TAC AAA GAA ACC ATA GAA ACA TTT GAT AAC AAT GAA GAA GAA TCC AGC TAT TCA TAC
LEU TYR LYS GLU THR ILE GLU THR PHE ASP ASN ASN GLU GLU GLU SER SER TYR SER TYR
                                       130                                 140
GAA GAA ATA AAT GAT CAG ACA AAC GAT AAT ATA ACA GCA AGA TTA GAT AGG ATT GAT GAA
GLU GLU ILE ASN ASP GLN THR ASN ASP ASN ILE THR ALA ARG LEU ASP ARG ILE ASP GLU
                                       150                                 160
AAA TTA AGT GAA ATA CTA GGA ATG CTT CAC ACA TTA GTA GTG GCA AGT GCA GGA CCT ACA
LYS LEU SER GLU ILE LEU GLY MET LEU HIS THR LEU VAL VAL ALA SER ALA GLY PRO THR
                                       170                                 180
TCT GCT CGG GAT GGT ATA AGA GAT GCC ATG ATT GGT TTA AGA GAA GAA ATG ATA GAA AAA
SER ALA ARG ASP GLY ILE ARG ASP ALA MET ILE GLY LEU ARG GLU GLU MET ILE GLU LYS
                                       190                                 200
ATC AGA ACT GAA GCA TTA ATG ACC AAT GAC AGA TTA GAA GCT ATG GCA AGA CTC AGG AAT
ILE ARG THR GLU ALA LEU MET THR ASN ASP ARG LEU GLU ALA MET ALA ARG LEU ARG ASN
                                       210                                 220
GAG GAA AGT GAA AAG ATG GCA AAA GAC ACA TCA GAT GAA GTG TCT CTC AAT CCA ACA TCA
GLU GLU SER GLU LYS MET ALA LYS ASP THR SER ASP GLU VAL SER LEU ASN PRO THR SER
                                       230                                 240
GAG AAA TTG AAC AAC CTA TTG GAA GGG AAT GAT AGT GAC AAT GAT CTA TCA CTT GAA GAT
GLU LYS LEU ASN ASN LEU LEU GLU GLY ASN ASP SER ASP ASN ASP LEU SER LEU GLU ASP

TTC TGA TTAGTTACCA CTCTTCACAT CAACACACAA TACCAACAGA AGACCAACAA ACTAACCAAC
PHE END
CCAATCATCC AACCAAACAT CCATCCGCCA ATCAGCCAAA CAGCCAACAA AACAACCAGC CAATCCAAAA

CTAACCACCC GGAAAAAATC TATAATATAG TTACAAAAAA AAAAAAA

MOLECULAR WEIGHT = 27150
```

FIG. 2. cDNA sequence of RS viral P gene. The DNA sequence is presented in the messenger sense. The derived amino acid sequence of the only available open reading frame is indicated starting with the N-terminal methionine.

*Hin*CII fragment labeled at the *Hin*fI site provided an accurate size marker. The DNA primer was extended five to six nucleotides beyond the *Pst*I cloning site (Fig. 3), implying that this recombinant lacked 5 or 6 nucleotides corresponding to the 5′ end of mRNA.

The cloned cDNA sequence had a single long open reading frame of 241 codons encoding a protein of 27,150 daltons. The calculated mass of 27,150 daltons for P of RS virus deduced from DNA sequence was somewhat smaller than the mass of ca. 33 kd estimated on Laemmli gels. This discrepancy is probably related to the degree of phosphorylation in vivo of this protein or artifacts of sodium dodecyl sulfate-polyacrylamide gel electrophoresis, or both. In addition, it has been shown by others that the mRNA encoding P is smaller than the one encoding M (5, 17). P is relatively acidic (21.1% acidic amino acids) rather than basic (13.2%). This is consistent with an earlier demonstration that a 36-kd viral protein was the second most acidic protein on nonequilibrium pH gradient electrophoresis (8). There are two clusters of acidic amino acids, one in the middle of the molecule and the other at the C terminus. These might represent the domains interacting with basic M. P is devoid of both cysteine and tryptophane, and the lack of cysteine is responsible for the inability of Collins et al. to label the protein in vivo with this amino acid (5). The codon usage reflects a deficiency of the CG dinucleotide similar to other RS viral genes (9, 27). The translational initiator ATG codon is between nucleotides 14 and 16 after the cloning site and the termination codon between positions 737 and 739. There is an untranslated region of 164 nucleotides after the termination codon. Three additional cDNA recombinants ($pRSC_1$, $pRSC_9$, and $pRSC_{13}$) were also sequenced. Although these recombinants lacked a few nucleotides just upstream of the poly(A) tail, their sequence matched base for base that of $pRSA_3$, thus establishing the authenticity of the UGA terminator. Possible cDNA cloning and DNA sequencing frameshift errors were also eliminated in an additional manner. A 70-base-pair (bp) *Hpa*II-*Fok*I fragment asymmetrically 5′-end labeled at the *Hpa*II site on the antimessenger strand was hybridized to mRNA from RS virus-infected cells. This DNA primer is 70 nucleotides downstream of the putative UGA terminator. After hybridization, the primer was partially extended on the template with reverse transcriptase and dideoxynucleoside triphosphates (27). The cDNA sequence derived in this manner (data not shown) was identical to the DNA sequence derived by Maxam-Gilbert sequencing of the cloned DNA. Similar to the RS virus NP and M genes (9, 27), a conserved sequence could not be identified just upstream of the poly(A) tail, unlike the paramyxoviruses and rhabdoviruses which have a conserved AUUC or AUAC sequence present in each gene upstream of the five or seven U residues that are reiteratively transcribed to yield the poly(A) tail of the individual transcripts (13, 22, 26). The initiating ATG codon is embedded within a PXXAUGG sequence found around functional eucaryotic translational codons (19).

P of RS virus is much smaller than the P's of paramyxoviruses. Partial proteolytic cleavage of Sendai virus and Newcastle disease virus nucleocapsids converted the capsid-associated moiety to a more basic form (3, 4); consequent to this treatment, Newcastle disease virus nucleocapsids lost a substantial amount of transcriptase activity, whereas there was minimal effect on similarly treated Sendai virus nucleocapsids (3, 4). Based on these findings Chinchar and Portner have proposed that a certain minimum size (between 26 and 40 kd) of the P must be associated with the viral nucleocapsid for transcription to occur (3, 4). Whether the small size of RS virus P represents the minimum size required for transcriptional activity cannot be determined until optimal in vitro transcriptional assays are established. In this context it

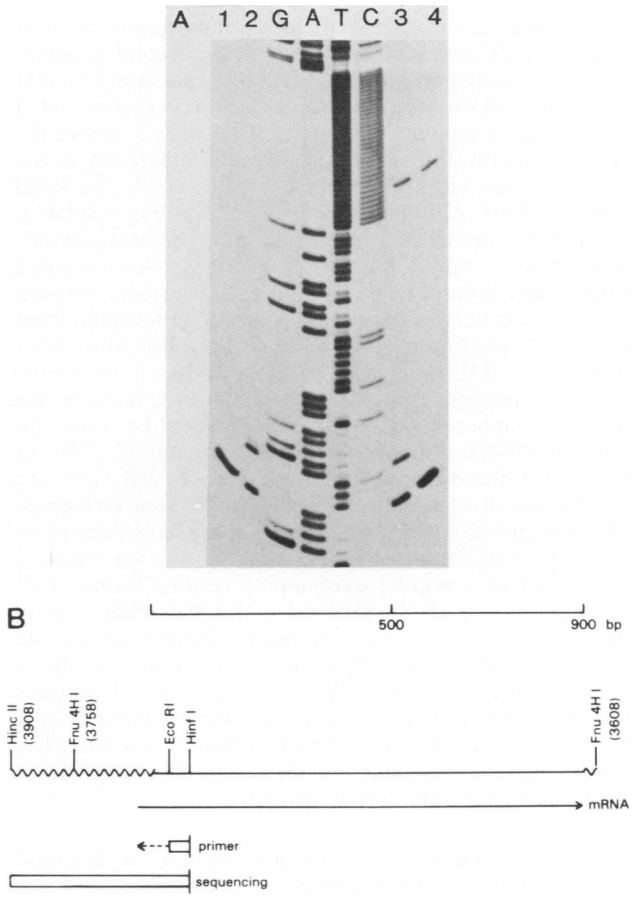

FIG. 3. Location of the 5′ end of mRNA encoding the RS viral P. The map coordinates of the cloned cDNA insert and the relevant restriction fragments with respect to the mRNA are schematically illustrated. The pBR322 sequence including the G:C tails generated during cDNA cloning are denoted by wavy lines on either side of the RS viral sequence. A 1,400-bp *Hin*fI fragment spanning the pBR322-RS cDNA junction was isolated and 5′-end labeled by polynucleotide kinase. The *Hin*fI site in the cloned viral sequence is downstream of the putative 5′ end of the mRNA. A 400-bp *Hin*fI-*Hin*cII fragment asymmetrically labeled at the *Hin*fI site (on the antimessage strand) was used to generate a chemical DNA sequence ladder. One picomole of a 45-bp *Hin*fI-*Eco*RI fragment primer labeled at the *Hin*fI site was denatured and annealed to poly(A) RNA (6 μg) from infected cells under conditions that have been documented before (27). The RNA-DNA hybrids were recovered after several cycles of ethanol precipitation, and the DNA primer was extended on the mRNA template with reverse transcriptase exactly as before (27). The primer extension products were electrophoresed alongside DNA sequencing reactions on an 8% acrylamide–urea gel under denaturing conditions. Lanes 1 and 2 denote the results when enzyme or RNA was not present in the primer extension reaction. Lane 3 illustrates the results in a standard reaction, and lane 4 represents the results when actinomycin D (50 μg/ml) was added during primer extension. The DNA sequence ladders are illustrated in lanes G, A, T, and C. The end of the cloned viral sequence is identified by the presence of the long C ladder generated during G:C tailing of the cloning procedure. The primer extension product is 5 to 6 nucleotides beyond the cloned viral sequence as judged by the C tails.

should be mentioned that NS of vesicular stomatitis virus is a transcriptional regulator of similar size (11).

Computer search failed to reveal local or global homologies (31) of RS virus P with Sendai virus P (12, 28) or NS of vesicular stomatitis virus (11), implying that RS virus P is evolutionarily distinct from these viruses. The other two reading frames are extensively blocked and do not possess an initiator ATG codon. This means that a second polypeptide is not encoded within the P cistron. This is in contrast with Sendai virus whose nonstructural protein C is encoded by a second overlapping reading frame near the 5′ end of the mRNA for P (12, 28). A similar situation apparently exists for the measles virus P gene (R. Bellini et al., personal communication). Although sequence data are not available, NDV and mumps virus also encode nonstructural proteins translated from the P mRNA. Since these nonstructural proteins share tryptic peptides with the N-terminal portion of the P, it has been proposed that they are translated from the same reading frame as the P but with a different initiator methionine (7, 14). Bunyaviruses, notably La Crosse virus, express two proteins from the S segment of the genome. La Crosse virus apparently accomplishes this feat by transcribing two mRNAs that are not 5′ coterminal (2, 24). In contrast, the three nonstructural proteins of RS virus are derived from discrete mRNAs, and three nonoverlapping cDNA plasmids containing these sequences have been identified (5; N. Elango, unpublished data). Earlier we reported the presence of a second overlapping reading frame of 75 amino acids within the M gene (27). An appropriate protein of this size, however, has not been detected among the translation products of hybrid selected mRNAs or in in fected cell extracts. Since RS viral nonstructural proteins are derived from distinct nonoverlapping transcriptional units, it would appear that during evolution RS virus has sacrificed genetic economy by recruiting separate genetic units encoding (specific) viral functions.

This work was performed in the laboratory of R. M. Chanock, whose enthusiasm and constant support for molecular studies with this virus are gratefully acknowledged. E. Camargo provided excellent technical support. The editorial assistance of M. Guiler cannot be overemphasized.

LITERATURE CITED

1. **Bernstein, J. M., and J. F. Hruska.** 1981. Respiratory syncytial virus proteins: identification by immunoprecipitation. J. Virol. **38**:278–285.
2. **Cabradilla, C. D., B. P. Holloway, and J. F. Obijeski.** 1983. Molecular cloning and sequencing of the La Crosse virus S RNA. Virology **128**:463–468.
3. **Chinchar, V. A., and A. Portner.** 1981. Functions of Sendai virus nucleocapsid polypeptides: enzymatic activities in nucleocapsids following cleavage of polypeptide P by *Staphylococcus aureus* protease V8. Virology **109**:59–71.
4. **Chinchar, V. A., and A. Portner.** 1981. Inhibition of RNA synthesis following proteolytic cleavage of Newcastle disease virus P protein. Virology **115**:192–202.
5. **Collins, P. L., Y. T. Huang, and G. W. Wertz.** 1984. Identification of a tenth mRNA of respiratory syncytial virus and assignment of polypeptides to the 10 viral genes. J. Virol. **49**:572–578.
6. **Collins, P. L., and G. W. Wertz.** 1983. cDNA cloning and transcriptional mapping of nine polyadenylated RNAs encoded by the genome of human respiratory syncytial virus. Proc. Natl. Acad. Sci. U.S.A. **80**:3205–3212.
7. **Collins, P. L., G. W. Wertz, L. A. Ball, and L. E. Hightower.** 1982. Coding assignments of the five smaller mRNAs of Newcastle disease virus. J. Virol. **43**:1024–1031.
8. **Dubovi, E. J.** 1982. Analysis of proteins synthesized in respiratory syncytial virus-infected cells. J. Virol. **42**:372–378.
9. **Elango, N., and S. Venkatesan.** 1983. Amino acid sequence of respiratory syncytial virus capsid protein. Nucleic Acids Res. **11**:5941–5951.
10. **Etkind, P. R., R. K. Cross, R. A. Lamb, D. C. Merz, and P. Choppin.** 1980. *In vitro* synthesis of structural and nonstructural proteins of Sendai and SV_5 viruses. Virology **100**:22–33.
11. **Gallione, C. J., J. R. Greene, L. E. Iverson, and J. K. Rose.** 1981. Nucleotide sequences of the mRNA's encoding the vesicular stomatitis virus N and NS proteins. J. Virol. **39**:529–535.
12. **Giorgi, C., B. M. Blumberg, and D. Kolakofsky.** 1983. Sendai virus contains overlapping genes expressed from a single mRNA. Cell **35**:829–836.
13. **Gupta, K. C., and D. W. Kingsbury.** 1982. Conserved polyadenylation signals in two negative strand RNA virus families. Virology **120**:415–421.
14. **Herrler, A., and R. W. Compans.** 1982. Synthesis of mumps virus polypeptides in infected Vero cells. Virology **119**:430–438.
15. **Hsu, C.-H., E. M. Morgan, and D. W. Kingsbury.** 1982. Site-specific phosphorylation regulates the transcriptive activity of Vesicular stomatitis virus NS protein. J. Virol. **43**:104–112.
16. **Huang, Y. T., and G. W. Wertz.** 1982. The genome of respiratory syncytial virus is a negative-strand RNA that codes for at least seven mRNA species. J. Virol. **43**:150–157.
17. **Huang, Y. T., and G. W. Wertz.** 1983. Respiratory syncytial virus mRNA coding assignments. J. Virol. **46**:667–672.
18. **Kingsbury, D. W.** 1977. Paramyxoviruses, p. 367–398. *In* D. B. Nayak (ed.), Molecular Biology of animal viruses. Marcel Dekker, Inc., New York.
19. **Kozak, M.** 1981. Possible role of flanking nucleotides in recognition of the AUG initiator codon by eukaryotic ribosomes. Nucleic Acids Res. **9**:5233–5252.
20. **Lamb, R. A., and P. W. Choppin.** 1978. Determination by peptide mapping of the unique polypeptides in Sendai virions and infected cells. Virology **84**:469–478.
21. **Lambert, D. M., and M. W. Pons.** 1983. Respiratory syncytial virus glycoproteins. Virology **130**:204–214.
22. **Lazzarini, R. A., J. D. Keene, and M. Schubert.** 1981. On the origin of defective interfering particles by the negative strand RNA virus. Cell **26**:145–154.
23. **Maxam, A., and W. Gilbert.** 1980. Sequencing end labeled DNA with base-specific chemical cleavages. Methods Enzymol. **65**:499–560.
24. **Patterson, J. L., C. Cabradilla, B. P. Holloway, J. F. Obijeski, and D. Kolakofsky.** 1983. Multiple leader RNAs and messenger RNAs are transcribed from the La Crosse virus small genome segment. Cell **33**:791–799.
25. **Peeples, M., and S. Levine.** 1979. Respiratory syncytial virus polypeptides: their location in the virion. Virology **95**:137–145.
26. **Rose, J. K.** 1980. Complete intergenic and flanking sequences from the genome of Vesicular stomatitis virus. Cell **19**:415–421.
27. **Satake, M., and S. Venkatesan.** 1984. Nucleotide sequence of the gene encoding respiratory syncytial virus matrix protein. J. Virol. **50**:92–99.
28. **Shioda, T., Y. Hidaka, T. Kanda, H. Shibuta, A. Nomoto, and K. Iwasaki.** 1983. Sequence of 3,687 nucleotides from the 3′ end of Sendai virus genome RNA and the predicted amino acid sequences of viral NP, P, and C proteins. Nucleic Acids Res. **11**:7317–7330.
29. **Venkatesan, S., N. Elango, and R. M. Chanock.** 1983. Construction and characterization of cDNA clones for four respiratory syncytial viral genes. Proc. Natl. Acad. Sci. U.S.A. **80**: 1280–1284.
30. **Wagner, R. R.** 1975. Reproduction of rhabdoviruses, p. 1–93. *In* H. Fraenkel-Conrat and R. R. Wagner (ed.), Comprehensive virology. Plenum Publishing Corp., New York.
31. **Wilbur, W. J., and D. J. Lipman.** 1983. Rapid similarity searches of nucleic acid and protein data banks. Proc. Natl. Acad. Sci. U.S.A. **80**:726–730.
32. **Wunner, W. H., and C. R. Pringle.** 1976. Respiratory syncytial virus proteins. Virology **73**:228–243.