The two open reading frames of the 22K mRNA of human respiratory syncytial virus: sequence comparison of antigenic subgroups A and B and expression in vitro

Peter L. Collins, Myron G. Hill and Philip R. Johnson

Laboratory of Infectious Diseases, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland 20892 and Georgetown University School of Medicine and Dentistry, Twinbrook Facility of the National Institute of Allergy and Infectious Diseases, Rockville, Maryland 20852, U.S.A.

The sequence of the 22K mRNA of strain 18537 of antigenic subgroup B of human respiratory syncytial virus (RSV) was determined by sequencing cloned cDNAs of intracellular mRNA. Comparison with the corresponding sequence of the A2 strain of subgroup A showed that there is 78% nucleotide sequence identity overall, that the amino acid sequence of the 22K protein is 92% identical between subgroups and that the 22K mRNA of both subgroups contains a second, internal, overlapping open reading frame (ORF) whose length, nucleotide sequence and potential translational start and stop sites were highly conserved and whose predicted product has 62% amino acid identity between subgroups. Sequence analysis of 36 cDNAs of intracellular 22K mRNA of strain A2 did not detect nucleotide insertions or deletions in the region of overlap between the two ORFs, indicating that the majority of intracellular 22K mRNA is a faithful copy containing the two distinct ORFs. Translation in vitro of mRNAs transcribed from engineered cDNAs showed that an mRNA which contained only the second, internal ORF directed the synthesis of a previously unidentified polypeptide of the predicted size, whereas mRNA representing the complete gene directed the synthesis of both the 22K protein and the product of the internal ORF. This latter species was synthesized in vitro as a discrete, separate protein rather than as a fusion protein. It is not yet known whether this protein is synthesized in RSV-infected cells.

Human respiratory syncytial virus (RSV) is the most important viral agent in serious respiratory tract disease in infants and children worldwide (McIntosh & Chanock, 1985; Collins, 1990). Two major antigenic subgroups of RSV, designated A and B, have been identified based on differences in their reactions with monoclonal and polyclonal antibodies (Coates et al., 1966; Anderson et al., 1985; Muñson et al., 1985; Gimenez et al., 1986). Information on RSV RNA and protein sequence diversity is pertinent to studies on the structure, function and evolution of RSV and its encoded proteins.

Molecular cloning and sequencing studies of strain A2 of subgroup A resulted in an essentially complete picture of the RSV gene map and encoded mRNAs and proteins (Collins et al., 1984; Collins & Wertz, 1985; Elango et al., 1985; Collins, 1990; and references cited therein). Ten RSV proteins have been described to date; these are the F and G glycoproteins, the small integral membrane protein SH (previously called 1A), the non-glycosylated M and 22K (or M2) proteins of the inner envelope, the large nucleocapsid protein L, the major nucleocapsid N protein, the nucleocapsid phosphoprotein P and the putative non-structural proteins 1C (or NS1) and 1B (or NS2) (Collins et al., 1984; Huang et al., 1985; Collins, 1990; and references cited therein). Sequence analysis has also been performed for several of the genes of several other subgroup A strains, and a high degree of intra-subgroup sequence conservation was demonstrated (Lambden, 1985; Baybutt & Pringle, 1987; Johnson et al., 1987b; Lopez et al., 1988).

Cloning and sequencing studies were extended to strain 18537 of antigenic subgroup B and showed that there is a substantial amount of intersubgroup sequence diversity. The percentage identity values between individual genes and proteins of strain 18537 and strain A2, representing subgroup A, are [nucleotide identity (%) / amino acid identity (%)] 1C (78/87), 1B (78/92), N (86/96), P (80/90), G (67/53), SH (78/76) and F (79/91) (Johnson et al., 1987b; Johnson & Collins, 1988a, b, 1989, 1990; Collins et al., 1990).

We have now extended this analysis to the 22K gene of strain 18537. Overlapping cDNA clones representing the complete 22K mRNA of strain 18537 were isolated from a previously described (Johnson et al., 1987a) library of
Fig. 1. (a) The complete nucleotide and deduced amino acid sequences of the 22K mRNA and encoded proteins of strain 18537 of antigenic subgroup B. The nucleotide and amino acid sequences are shown in a sequence alignment with their counterparts from strain A2 of antigenic subgroup A, whose amino acid sequence is also italicized; L ORF, the portion of the mRNA overlap between the 22K and L genes. The nucleotide L gene start sequence is underlined. (b) Schematic diagram of the positions of the ORFs in the 22K mRNAs of strains A2 (top) and 18537 (bottom). The maximum amino acid (aa) lengths of the encoded proteins are indicated in large numbers (194 aa, 90 aa, 195 aa and 93 aa). Nucleotide sequence positions are indicated in smaller numbers for potential translational start and stop sites of the two ORFs, with the numbers referring to the first nucleotide of start codons or the last nucleotide of stop codons.
translational stop codon that terminates the second ORF was exactly conserved between subgroups. The nucleotide sequence of the second ORF (nucleotides 563 to 835) shares 72% identity between subgroups, compared with 84% identity for the major 22K ORF (nucleotides 10 to 597). In contrast, the one area of the 22K gene which does not contain an ORF (nucleotides 836 to 892) has only 34% identity. Given the general observation for RSV (Johnson & Collins, 1988b; Collins, 1990) that the sequences of expressed ORFs are well conserved whereas non-protein-encoding sequences are not, this would suggest that the second ORF of the 22K mRNA encodes a viral protein.

The length of the predicted protein would be 83 to 90 amino acids for strain A2 and 83 to 93 amino acids for strain 18537, depending on the choice of translational start site. The predicted amino acid sequences are 62% identical, a value that is lower than those for all other proteins examined to date except the G protein. The predicted protein would have well conserved N and C termini (88% and 84% identity, respectively, for the N-terminal 24 and C-terminal 19 amino acids, compared with 40% identity for the internal 47 amino acids). We are in the process of preparing antisera against synthetic peptides representing the predicted product in order to investigate its presence in RSV-infected cells and virions.

Recently, several paramyxoviruses were shown (Thomas et al., 1988; Cattaneo et al., 1989; Vidal et al., 1990) to utilize a so-called RNA editing mechanism whereby the viral transcriptase sometimes stutters reiteratively at a single site during transcription of the P/C gene. Stuttering results in the insertion of a variable number of non-templated nucleotides at this one site, giving rise to a mixed population of mRNAs. Some of these species contain inserted nucleotides in the appropriate number to change the register of the reading frame and allow access to an internal ORF, thereby creating a chimeric ORF that encodes a fusion protein.

Given this precedent, it is possible that expression of the 22K gene might involve a similar stuttering mechanism. Specifically, the insertion of two nucleotides [or larger numbers according to the formula (N x 3) + 2, with N being a positive integer] between nucleotides 554 and 591 would create a translational stop codon at nucleotides 527 to 529 (strain A2) or 551 to 553 (strain 18537). To investigate the possibility of stuttering in this region of the 22K mRNA, cDNAs were constructed from three independent preparations of strain A2 mRNA. First-strand synthesis was performed with the synthetic negative-sense primer 883-TGTGTGAATTAATAATGAAAGCTTGGAGTTAGATTAATTGATGTAT (identified by nucleotide sequence position with a HindIII site that was introduced for cloning underlined). The three cDNA preparations were individually amplified during 20 rounds of polymerase chain reaction (PCR) using the same negative-sense oligonucleotide together with the positive-sense oligonucleotide primer 1-GGGGCAATATATGTACGAGGGATCCCTTGCAAGATTGAAATCATGTAATAAGGGTGA. The major products of the PCR were discrete fragments of the appropriate size and these were digested with BamHI and HindIII and cloned in pGem3 (Promega Biotec). Twelve cDNAs from each of the three sets were analysed by dideoxynucleotide sequencing with the positive-sense primer 442-CTATATCCTCCTGTTAAAAGAGA and the negative-sense primer 690-TGTGTGAATTAATAATGAAAGCTTGGAGTTAGATTAATTGATGTAT. The lengths of the predicted protein would be approximately 274 amino acids in length for strain A2. The predicted amino acid sequences are 62% identical, a value that is lower than those for all other proteins examined to date except the G protein. The predicted protein would have well conserved N and C termini (88% and 84% identity, respectively, for the N-terminal 24 and C-terminal 19 amino acids, compared with 40% identity for the internal 47 amino acids). We are in the process of preparing antisera against synthetic peptides representing the predicted product in order to investigate its presence in RSV-infected cells and virions.

Another possible mechanism for expression of the second ORF of the 22K mRNA would be ribosomal frameshifting, as has been shown to be involved in the synthesis of the gag-pol fusion protein of Rous sarcoma virus and in yeast and bacterial systems (Jacks & Varmus, 1985, Craigen & Caskey, 1986; Jacks et al., 1988). To examine this possibility, the previously described cDNA 22K-16, encoding the strain A2 22K protein (Collins & Wertz, 1985), was modified by site-directed mutagenesis to insert a BamHI site preceding nucleotide -3 relative to the translational start site; the 22K cDNA contains a second, naturally occurring BamHI site at nucleotides 904 to 909, downstream of both ORFs. The cDNA was digested with BamHI, placed under the control of the SP6 phage polymerase in
pGem3Zf+ (Promega Biotec) and used to synthesize mRNA in the presence of $m^7G(5')ppp(5')G$ (Pharmacia) using standard protocols (Spriggs & Collins, 1990). For comparison, a cDNA was prepared that contained only the second, internal ORF. This was synthesized by the PCR using full-length cDNA as template and the 51 nucleotide negative-sense primer described above and the positive-sense primer 532-AACCCAAAAGAATCCACTGATTAGGTCCCAATGACCATGCCAAAATAAT (with an introduced BamHI site underlined) as primers. This cDNA was cloned into pGem3Zf+ (Promega Biotec) under the control of the SP6 promoter and the cDNA sequence was confirmed in its entirety, and capped mRNA was synthesized and analysed, in parallel with mRNA synthesized in vitro from the full-length cDNA, by translation in rabbit reticulocyte lysates (Promega Biotec) in the presence of $[^{35}S]$methionine, followed by SDS–PAGE (Fig. 2).

Translation of the truncated mRNA (Fig. 2, lane 1) resulted in a single major product whose estimated $M_r$ was 9.0K, calculated using commercial markers (Amersham), compared to the calculated $M_r$ of 10.666 for the predicted product of the second ORF. This would be consistent with translational initiation at one of the three methionyl codons within nucleotides 563 to 583 at the start of the second ORF. On the other hand, this protein product migrated more rapidly than did the $Sh_0$ protein, which has a calculated $M_r$ of 7.5K. However, it is well known that the migration of small proteins in this gel system is not always proportional to their $M_r$ (Schagger & von Jagow, 1987); for example, the $M_r$ of the $Sh_0$ protein of strain A2 is 3-5K greater than that of its counterpart in strain 18537, whereas its calculated molecular mass is essentially identical (7 Da greater) (Collins et al., 1990). Thus, the exact site of translation initiation of this ORF and the size of the encoded protein remain to be determined experimentally.

The products synthesized in vitro in response to the complete mRNA (Fig. 2, lane 3) included the 22K protein, a series of smaller species which could have been generated by premature translational termination or translational initiation at internal methionine residues, and the protein which corresponded to the second ORF. This protein was synthesized in less abundance than the 22K protein, although the molar ratio cannot be calculated reliably until the exact translational start site of the second ORF is determined so that the number of methionine residues is known (for strain A2, the 22K protein contains four methionine residues compared with up to five for the product of the second ORF). We assume that, in the case of the complete mRNA, the product of the second ORF was synthesized from intact mRNA, although we cannot yet eliminate the possibility that mRNA fragmentation was required for ribosomes to gain access to the internal ORF. Importantly, there was no significant accumulation of translation products larger than the 22K protein which could possibly have resulted from frameshifting, suggesting that that mechanism is not involved here.

Recently, Horvath et al. (1990) showed that a second, internal ORF in RNA segment 7 of influenza B virus encodes a previously undescribed protein called BM2. In the segment 7 mRNA, the upstream ORF encoding the M protein and the downstream ORF encoding the BM2 protein overlap by a single nucleotide at the junction TAATG and their translation is coupled. In comparison, for the 22K mRNA, the closest translational start sites in the second ORF precede or follow the translational stop sites of the upstream ORF by 14 or 46 nucleotides,
respectively. Additional information will be required for the 22K mRNA to determine whether it is expressed by a mechanism similar to that of segment 7 of influenza B virus. That the segment 7 and 22K gene mRNAs both encode a non-glycosylated inner membrane protein as the major translation product might be a coincidence or might be evidence of a distant evolutionary relationship.

Finally, another possibility is that the internal ORF is expressed by direct internal ribosomal initiation independent of the upstream ORF.

The high degree of conservation of the 22K gene between subgroups A and B is of particular interest because it has recently been shown to encode the major target antigen for RSV-specific murine cytotoxic T lymphocytes (Openshaw et al., 1990; Nicholas et al., 1990). This is likely to be an important factor in the high degree of cross-strain reactivity observed for RSV-specific murine cytotoxic T lymphocytes (Bangham & Askonas, 1986). However, the recombinant vaccinia-22K viruses used for the target analyses in the studies cited above contained both ORFs of the 22K gene and potentially could have directed the synthesis of both proteins. It will be necessary to test individually the activities of the two ORFs in encoding target antigens.

This work was performed in the laboratory of Dr Robert M. Chanock and Dr Brian R. Murphy. We thank them for their interest and support. We also thank Dr Robert A. Lamb for communicating results on the influenza B virus BM2 protein prior to publication.

References


JOHNSON, P. R. & COLLINS, P. L. (1989). The 1B (NS2), 1C (NS1) and N proteins of human respiratory syncytial virus (RSV) of antigenic subgroups A and B: sequence conservation and divergence within RSV genomic RNA. Journal of General Virology 70, 1539-1547.


(Received 11 June 1990; Accepted 28 August 1990)