Influenza C Virus Hemagglutinin: Comparison with Influenza A and B Virus Hemagglutinins

SUSUMU NAKADA, RICHARD S. CREAGER, MARK KRYSTAL, ROBERT P. AARONSON, AND PETER PALESE*

Department of Microbiology, Mount Sinai School of Medicine of The City University of New York, New York, New York 10029

Received 10 September 1983/Accepted 12 December 1983

The complete nucleotide sequence of the influenza C/California/78 virus RNA 4 was obtained by using cloned cDNA derived from the RNA segment. This gene is 2,071 nucleotides long and can code for a polypeptide of 654 amino acids. Although there are no convincing sequence homologies between RNA 4 and the hemagglutinin genes of influenza A and B viruses, we suggest, on the basis of structural features, that RNA 4 of the influenza C virus codes for the hemagglutinin. The structural features which are common to the hemagglutinins of influenza A, B, and C viruses include (i) a hydrophobic signal peptide, (ii) an arginine cleavage site between the hemagglutinin 1 and 2 subunits, (iii) hydrophobic regions at the amino and carboxyl termini of the hemagglutinin 2 subunit, and (iv) several conserved cysteine residues. Additional evidence that RNA 4 of influenza C virus codes for the hemagglutinin is that the tripeptide Ile-Phe-Gly, known to be present at the amino terminus of the hemagglutinin 2 subunit of influenza C virus, is encoded by RNA 4 at a point immediately adjacent to the presumptive arginine cleavage site. The lack of primary sequence homology between the influenza C virus hemagglutinin and the influenza A or B virus hemagglutinins, which all have similar functions, might be attributed to convergent rather than divergent evolution. However, the structural similarities among the influenza A, B, and C virus hemagglutinins strongly suggest that the three hemagglutinin genes have diverged from a common precursor.

Influenza C viruses share many biochemical properties with the other orthomyxoviruses. For example, all three influenza virus types (A, B, and C) possess segmented RNA genomes of negative polarity. They also show morphological similarity, and features of their replication cycles are similar (1). On the other hand, there exist significant differences between influenza C viruses and the other orthomyxoviruses in the number of RNA segments and protein components. The influenza C virus genome appears to consist of only seven RNA segments (6, 23), whereas influenza A and B viruses each possess eight RNAs. Influenza C virions lack alpha-neuraminidase (11, 20), but both influenza A and B viruses carry such activity on the surface of the particles.

Coding assignments for the genomes of influenza A and B viruses have been established, but no definitive information has been obtained regarding a genetic map of influenza C viruses. It has been demonstrated that the exchange of RNA 4 and RNA 5 in influenza C/JHG/66 virus results in a change of its plaquing phenotype (24), but a precise coding assignment of these RNAs has not been achieved.

In this paper, we report the complete nucleotide sequence of the influenza C/California/78 (C/Cal/78) virus RNA 4, which appears to code for a gene product with structural properties similar to those of influenza A and B virus hemagglutinins (HAs). Thus, it is likely that RNA 4 codes for the glycoprotein which has been shown to be present in intact influenza C viruses and to make up the surface spikes of the virions (4, 9).

MATERIALS AND METHODS

Virus. Influenza C virus C/Cal/78 was grown in embryonated chicken eggs. Virus purification and RNA extraction have been described previously (25).

Cloning of influenza C virus-specific DNA and identification

of clones derived from RNA 4. Double-stranded cDNA was synthesized from influenza C/Cal/78 virus RNA by using reverse transcriptase and the synthetic dodecamer nucleotide primers d(AGCAAAAGCAG)rG and d(AGTAGAAACAAG) (Collaborative Research, Inc., Waltham, Mass.), which had been similarly utilized for the cloning of influenza A and B virus RNAs (3, 14). Unless otherwise indicated, synthetic *EcoRI* linkers (Collaborative Research, Inc.) were added to the double-stranded DNA, and insertion into the *EcoRI* site of plasmid pBR322 was achieved through ligation and transformation of *Escherichia coli* C600 cells (3). Several virus-specific clones were obtained, and a gene library was established, through hybridization of inserts by the Southern blotting technique (28). One group of clones was shown to derive from RNA 4 (see below), and the largest clone, pC316, was found to contain the 3'-terminal nucleotide sequence of influenza C virus RNAs (6). However, partial end sequencing revealed that the cloned DNA lacked the sequence present at the 5' end of influenza C virion RNAs (6). Extensive analysis of all recombinant plasmids did not result in the identification of a clone covering additional regions of RNA 4. Therefore, a restriction enzyme fragment (corresponding to nucleotides 856 to 1126) isolated from the pC316 clone was then used as a primer to obtain cDNA transcripts of the 5' end of the virion RNA. Second-strand cDNA synthesis was achieved by using a new primer, d(AGCAGGAGCAAG) (New England Biolabs, Beverly, Mass.), specific for the 5' end of influenza C virion RNAs. Since the primer used was known to contain an *NcoI* restriction enzyme site, the double-stranded DNA was treated with *NcoI* and subsequently inserted between the *NcoI* and *NruI* sites of a pBR322 recombinant plasmid (submitted for publication), which contains a full-length copy of the influenza A/Alaska/6/77 virus NS gene. This plasmid was used because it has a convenient *NcoI*-*NruI* window, permitting direct insertion of the virus-specific double-stranded

* Corresponding author.

FIG. 1. Identification of a pBR322 recombinant plasmid containing cDNA derived from influenza C/Cal/78 virus RNA. Glyoxalated total viral RNA was separated by electrophoresis on polyacrylamide gels. (A) Hybridization done with radioactive labeled cDNA prepared from total viral RNA. (B) Hybridization done with nick-translated insert DNA from recombinant plasmid pC316.

DNA into the cloning vector without the addition of synthetic linkers. This procedure resulted in the isolation of a plasmid (pC402) containing the 5'-terminal sequence of influenza C virus RNAs (6). It should be noted that a primer specific for the 5' terminus of influenza C virus RNAs was used in this cloning experiment.

Sequencing and computer analysis. The sequence of both clones was determined by the Maxam and Gilbert chemical modification procedure (19). DNA fragments suitable for sequence analysis were obtained either by strand separation or by secondary cleavage of 5'-end-labeled or 3'-end-filled restriction enzyme fragments.

Nucleotide and amino acid sequence data were stored and edited with an Amdahl 470/V6 computer at the University Computer Center of the City University of New York by using published programs (29-31). The nucleotide and amino acid sequences of the C/Cal/78 RNA 4 were compared with those of the A/PR/8/34 and B/Lee/40 HA genes by using an algorithm, performed on an IBM 3033 computer, in which the window length and the stringency (number of identical residues) were independently adjustable and the comparison matrices were plotted via a Tektronix 4052 microcomputer coupled to either a 4061 screen copier or a 4662 X-Y plotter as described previously (14).

For the analysis of the relative hydrophilicity of the influenza C/Cal/78 virus RNA 4 protein and the A and B virus HAs, a window of 10 amino acids was used to generate a relative hydrophilicity value based on published hydrophil-

icity values for each amino acid (16) (relative hydrophilicity = total hydrophilicity of decapeptide/10).

RESULTS AND DISCUSSION

Identification of clones derived from influenza C virus RNA

4. Preliminary cloning experiments of the RNAs of influenza C virus C/Cal/78 resulted in the isolation of several virus-specific clones. One of the recombinant plasmids (pC316) was shown by the Northern blotting technique to be derived from RNA 4 (2, 5). This was accomplished by hybridizing nick-translated insert DNA to total viral RNA which was electrophoretically separated on polyacrylamide gels (Fig. 1). Figure 1 indicates that plasmid pC316 contains sequences corresponding to RNA 4. Since clone pC402 contains sequences from the 5' end of influenza C virus RNA and from the insert of plasmid pC316, it was assumed that these two clones span the entire sequence of RNA 4. In addition, size estimates of the inserts suggested that these two plasmids contained approximately 2,000 nucleotides of unique virus-derived sequences. This was in good agreement with the results obtained by polyacrylamide gel electrophoresis of glyoxalated influenza C virus RNA 4 (23).

Nucleotide sequences. The strategy for determining the nucleotide sequence of the inserts of plasmids (pC316 and pC402) is shown in Fig. 2. All restriction enzyme sites used for the preparation of restriction enzyme fragments were covered by overlapping sequences. The entire nucleotide sequence of RNA 4 of the influenza C/Cal/78 virus is shown in Fig. 3. The segment contains 2,071 nucleotides. After a 5' noncoding region of 21 nucleotides, there is an open reading frame that extends from the first possible initiation codon for protein synthesis at nucleotides 22 through 24 to a termination codon at nucleotides 1984 through 1986. This open reading frame could code for a protein of 654 amino acids. The adenine-rich string at nucleotides 2043 through 2056 probably represents the polyadenylation site of the viral mRNA in analogy to those found in influenza A and B viral mRNAs (14, 26). No additional open reading frames longer than 90 amino acids were detected on either the plus- or the minus-sense strand.

Identification of influenza C virus RNA 4 protein. The deduced amino acid sequence of the influenza C/Cal/78 virus RNA 4 is shown in Fig. 3. Preliminary analysis and comparison of this sequence with those of known influenza A and B virus proteins did not permit an unequivocal assignment to a particular protein such as the nucleoprotein or the hemagglutinin. However, several lines of evidence suggest that this sequence is the equivalent of an HA in influenza C viruses.

FIG. 2. Strategy for obtaining the nucleotide sequence of the influenza C/Cal/78 virus RNA 4. The numbering system corresponds to the direction of mRNA transcription. The arrows indicate the direction and length of sequence analysis performed on each of the restriction enzyme fragments. Two pBR322 recombinant plasmids (pC316 and pC402) containing overlapping virus-specific sequences were used in this analysis.

AGC	AAA	AGC	AGG	GGT	TTA	ATA	MET	PHE	PHE	SER	LEU	LEU	LEU	MET	LEU	GLY	LEU	THR	GLU	ALA	GLU	LYS	ILE	LYS	ILE	CYS	LEU	GLN	LYS	22	
GLN	VAL	ASN	SER	SER	PHE	SER	CTA	HIS	ASN	GLY	PHE	GLY	GLY	ASN	LEU	TYR	ALA	THR	GLU	GLU	GLY	ARG	MET	PHE	GLU	LEU	VAL	LYS	PRO	530	
CAA	GTC	AAC	AGT	AGC	TTC	AGC	CTA	CAC	AAT	GGC	TTC	GGA	GGA	AAT	TTG	TAT	GCC	ACA	GAA	GAA	AAA	AGA	ATG	TTT	GAG	CTT	GTT	AAG	CCC	180	
LYS	ALA	GLY	ALA	SER	VAL	LEU	ASN	GLN	SER	THR	TRP	ILE	GLY	PHE	GLY	ASP	SER	ARG	THR	ASP	GLN	SER	ASN	SER	ALA	PHE	PRO	ARG	SER	270	
AAA	GCT	GGA	GCC	TCT	GTC	TTG	AAT	CAA	AGC	ACA	TGG	ATT	GGC	TTT	GGA	GAT	TCA	AGA	ACT	GAC	CAA	AGC	AAT	TCA	GCT	TTT	CCT	AGG	TCG	83	
LEU	MET	SER	ALA	LYS	THR	ALA	ASP	LYS	PHE	ARG	SER	LEU	SER	GLY	GLY	SER	LEU	MET	LEU	SER	MET	PHE	GLY	PRO	PRO	GLY	LYS	VAL	ASP	113	
CTG	ATG	TCA	GCA	AAA	ACT	GCT	GAT	AAA	TTT	CGT	TCT	TTG	TCT	GCT	GGA	TCC	TTG	ATG	TTG	AGT	ATG	TTT	GGC	CCA	CCT	GGG	AAG	GTA	GAT	360	
TYR	LEU	TYR	GLN	GLY	CYS	GLY	LYS	HIS	LYS	VAL	PHE	TYR	GLU	GLY	VAL	ASN	TRP	SER	PRO	HIS	ALA	ALA	ILE	ASP	CYS	TYR	ARG	LYS	ASN	143	
TAC	CTT	TAC	CAA	GGA	TGT	GGA	AGC	CAT	AAA	GTT	TTT	TAT	GAA	GGA	GTC	AAC	TGG	AGT	CCA	CAT	GCT	GCT	ATA	GAT	TGT	TAC	AGA	AAA	AAT	450	
TRP	THR	ASP	ILE	LYS	LEU	ASN	PHE	GLN	LYS	SER	ILE	TYR	GLU	LEU	ALA	SER	GLN	SER	HIS	CYS	MET	SER	LEU	VAL	ASN	ALA	LEU	ASP	LYS	173	
TGG	ACT	GAC	ATC	AAA	CTG	AAT	TTC	CAG	AAA	AGC	ATT	TAT	GAA	TTG	GCT	TCA	CAA	TCA	CAT	TGC	ATG	AGC	TTG	GTG	AAT	GCC	TTG	GAC	AAA	540	
THR	ILE	PRO	LEU	GLN	VAL	THR	LYS	GLY	VAL	ALA	LYS	ASN	CYS	ASN	ASN	SER	PHE	LEU	LYS	ASN	PRO	ALA	LEU	TYR	THR	GLN	GLU	VAL	LYS	203	
ACT	ATT	CCT	TTA	CAA	GTC	ACT	AAA	GGA	GTT	GCA	AAA	AAT	TGC	AAC	AAC	AGC	TTC	TTA	AAA	AAT	CCA	GCA	TTG	TAC	ACA	CAA	GAA	GTC	AAA	630	
PRO	LEU	GLU	GLN	ILE	CYS	GLY	GLU	GLU	ASN	LEU	ALA	PHE	PHE	THR	LEU	PRO	THR	GLN	PHE	GLY	THR	TYR	GLU	CYS	LYS	LEU	HIS	LEU	VAL	233	
CCT	TTA	GAG	CAA	ATA	TGT	GGG	GAA	GAA	AAT	CTT	GCT	TTT	TTC	ACA	CTT	CCA	ACC	CAA	TTT	GGA	ACC	TAT	GAG	TGC	AAA	CTG	CAT	CTT	GTG	720	
ALA	SER	CYS	TYR	PHE	ILE	TYR	ASP	SER	LYS	GLU	VAL	TYR	ASN	LYS	ARG	GLY	CYS	GLY	ASN	TYR	PHE	GLN	VAL	ILE	TYR	ASP	SER	GLY	263		
GCT	TCT	TGC	TAT	TTC	ATC	TAT	GAT	AGC	AAA	GAA	GTG	TAC	AAT	AAA	AGA	GGA	TGT	GGC	AAC	TAC	TTT	CAA	GTG	ATC	TAT	GAT	TCA	TCT	GGA	810	
LYS	VAL	VAL	GLY	GLY	LEU	ASP	ASN	ARG	VAL	SER	PRO	TYR	THR	GLY	ASN	SER	GLY	ASP	THR	PRO	THR	MET	GLN	CYS	ASP	MET	LEU	GLN	LEU	283	
AAA	GTT	GTT	GGA	GGG	CTA	GAT	AAC	AGG	GTA	TCA	CCT	TAC	ACA	GGG	AAT	TCT	GGA	GAC	ACT	CCA	ACA	ATG	CAA	TGT	GAC	ATG	CTC	CAG	CTG	900	
LYS	PRO	GLY	ARG	TYR	SER	VAL	ARG	SER	SER	PRO	ARG	PHE	LEU	LEU	MET	PRO	GLU	ARG	SER	TYR	CYS	PHE	ASP	MET	LYS	GLU	LYS	GLY	PRO	323	
AAA	CCT	GGA	AGA	TAT	TCA	GTA	AGA	AGC	TCT	CCA	AGA	TCT	CTT	TTA	ATG	CCT	GAA	AGG	AGT	TAT	TGC	TTT	GAC	ATG	AAA	GAA	AAA	GGA	CCA	990	
VAL	THR	ALA	VAL	GLN	SER	ILE	TRP	GLY	LYS	GLY	ARG	LYS	SER	ASP	TYR	ALA	VAL	ASP	GLN	ALA	CYS	LEU	SER	THR	PRO	GLY	CYS	MET	LEU	353	
GTC	ACT	GCT	GTC	CAA	TCC	ATC	TGG	GGA	AAA	GGC	AGA	AAA	TCT	GAC	TAT	GCA	GTA	GAT	CAG	GCT	TGC	TTG	AGC	ACT	CCA	GGG	TGC	ATG	TTG	1080	
ILE	GLN	LYS	GLN	LYS	PRO	TYR	ILE	GLY	GLU	ALA	ASP	ASP	HIS	HIS	GLY	ASP	GLN	GLU	MET	ARG	GLU	LEU	LEU	SER	GLY	LEU	ASP	TYR	GLU	383	
ATC	CAA	AAG	CAA	AAG	CCA	TAC	ATT	GGA	GAG	GCT	GAT	GAT	CAC	CAT	GCA	GTA	GAT	CAA	GAA	ATG	AGG	GAG	TTG	CTG	TCA	GGA	CTG	GAC	TAT	GAA	1170
ALA	ARG	CYS	ILE	SER	GLN	SER	GLY	TRP	VAL	ASN	GLU	THR	SER	PRO	PHE	THR	GLU	GLU	TYR	LEU	LEU	PRO	PRO	LYS	PHE	GLY	ARG	CYS	PRO	413	
GCT	AGA	TGC	ATA	TCA	CAA	TCA	GGG	TGG	GTG	AAT	GAA	ACC	AGT	CCT	TTT	ACG	GAA	GAA	TAC	CTC	CTT	CCT	CCC	AAA	TTT	GGG	AGA	TGT	CCC	1260	
LEU	ALA	ALA	LYS	GLU	GLU	SER	ILE	PRO	LYS	ILE	PRO	ASP	GLY	LEU	LEU	ILE	PRO	THR	SER	GLY	THR	ASP	THR	THR	VAL	THR	LYS	PRO	LYS	443	
TTG	GCC	GCA	AAG	GAA	TCC	ATT	CCA	AAA	ATC	CCA	GAT	GGA	CTT	CTA	ATT	CCC	ACC	AGT	GGA	ACT	GAT	ACC	ACT	GTA	ACC	AAA	CCT	AAA	1350		
SER	ARG	ILE	PHE	GLY	ILE	ASP	LEU	ILE	ILE	GLY	LEU	PHE	VAL	ALA	ILE	VAL	GLU	ALA	GLY	ILE	GLY	GLY	TYR	LEU	GLY	SER	LEU	GLY	SER	473	
AGC	AGA	ATT	TTT	GGA	ATC	GAT	GAC	CTT	ATT	ATT	GGT	CTA	CTA	TTT	GTT	GCA	ATT	GTT	GAA	GCA	GGA	ATT	GGA	GGC	TAT	CTG	CTT	GGA	AGT	1440	
ARG	LYS	GLU	SER	GLY	GLY	VAL	THR	LYS	GLU	SER	ALA	GLU	LYS	PHE	GLU	LYS	ILE	GLY	ASN	ASP	ILE	GLN	ILE	ARG	SER	SER	SER	SER	SER	503	
AGA	AAA	GAA	TCA	GGA	GGA	GGT	GTG	ACA	AAA	GAA	TCA	GCT	GAA	AAA	GGG	TTT	GAA	AAA	ATT	GGA	AAT	GAC	ATA	CAA	ATC	TTA	AGA	TCT	TCT	1530	
THR	ASN	ILE	ALA	ILE	GLU	LYS	LEU	ASN	ASP	ARG	ILE	SER	HIS	ASP	GLU	GLN	ALA	ILE	ARG	ASP	LEU	THR	LEU	GLU	ILE	GLU	ASN	ALA	ARG	533	
ACA	AAT	ATT	GCA	ATA	GAA	AAA	CTG	AAC	GAC	AGA	ATT	TCT	CAT	GAT	GAG	CAA	GCC	ATC	AGA	GAT	CTA	ACT	TTA	GAA	ATT	GAA	AAT	GCA	AGA	1620	
SER	GLU	ALA	LEU	LEU	GLY	GLU	LEU	GLY	ILE	ILE	ARG	ALA	LEU	LEU	VAL	GLY	ASN	ILE	SER	ILE	GLY	LEU	GLN	GLU	SER	LEU	TRP	GLU	LEU	563	
TCT	GAA	GCT	CTA	TTA	GGA	GAA	TTG	GGA	ATA	ATA	AGC	GCC	TTG	CTG	TTA	GGA	AAT	ATA	AGC	ATA	GGA	TTA	CAA	GAA	TCT	TTA	TGG	GAA	CTA	1710	
ALA	SER	GLU	ILE	THR	ASN	ARG	ALA	GLY	ASP	LEU	ALA	VAL	GLU	VAL	SER	PRO	GLY	CYS	TRP	ILE	ILE	ASP	ASN	ASN	ILE	CYS	ASP	GLN	SER	593	
GCT	TCA	GAA	ATA	ACA	AAT	AGA	GCA	GGA	GAC	CTG	GCA	GTC	GAA	GTC	TCT	CCA	GGT	TGC	TGG	ATA	ATC	GAC	AAT	AAC	ATT	TGT	GAT	CAA	AGT	1800	
CYS	GLN	ASN	PHE	ILE	PHE	LYS	PHE	ASN	GLU	THR	ALA	PRO	VAL	PRO	THR	ILE	PRO	PRO	LEU	ASP	THR	LYS	ILE	ASP	LEU	GLN	SER	ASP	PRO	623	
TGT	CAA	AAC	TTT	ATT	TTC	TAG	TCT	AAC	GAA	ACT	GGC	CCT	GTT	CCA	ACC	ATT	CCC	CCT	CTT	GAC	ACA	AAA	ATT	GAT	CTG	CAA	TCA	GAT	CCT	1890	
PHE	TYR	TRP	GLY	SER	SER	LEU	GLY	LEU	ALA	ILE	THR	ALA	ALA	ASN	LEU	MET	ALA	ALA	LEU	VAL	ILE	SER	GLY	ILE	ALA	ILE	CYS	ARG	THR	653	
TTT	TAC	TGG	GGA	AGC	AGC	TTG	GGC	TTA	GCA	ATA	ACT	GCT	GCT	AAT	CTA	ATG	GCA	GCT	TTG	GTG	ATC	TCT	GGG	ATC	GCC	ATC	TGC	AGA	ACT	1980	
LYS	***																												654		
AAA	TGA	TCA	GGA	CAA	TTT	TGA	AAA	ATG	GAT	AAT	ATA	TTA	GTC	AAT	ATT	TTG	TAC	AGC	TTT	ATA	AAA	AAA	CAA	AAA	ACC	CCT	TGC	TAC	TGC	2070	
T																													2070		

FIG. 3. Complete nucleotide and deduced amino acid sequence of the influenza C/Cal/78 virus RNA 4. The predicted cleavage sites in the protein precursor are indicated by arrows.

First, Figure 4 shows the hydrophilicity plot of influenza A/PR/8/34 and B/Lee/40 virus HAs as well as that of the deduced amino acid sequence of RNA 4 of influenza C/Cal/78 virus. In all three proteins there are hydrophobic amino and carboxyl termini which, in the case of influenza A and B virus HAs, correspond to the signal peptides and to the membrane-bound portions of the HA2 subunit. Furthermore, the influenza A and B virus HAs show a hydrophobic region at the amino terminus of the HA2. A similar hydrophobic region is found at amino acid position 446 to 466 of the C/Cal/78 virus RNA 4 protein (Fig. 3 and 4). Clearly, the RNA 4 protein has a hydrophobicity pattern which is compatible with the protein structure of an influenza A or B virus HA possessing a cell fusion function and a signal peptide in its precursor.

Previously, Herrler et al. reported that the smaller cleavage product of the influenza C viral glycoprotein has an Ile-Phe-Gly amino terminus (9). This tripeptide corresponds to amino acid position 446 through 448, which is located at the beginning of the above-mentioned internal hydrophobic region (Fig. 3). This tripeptide is preceded by an arginine residue which may represent a protease cleavage site (Fig. 3).

Herrler et al. also reported that the amino terminus of the large surface glycoprotein of influenza C virus is blocked to the Edman degradation procedure (9). If we assume that the first 14 amino acids of RNA 4 protein represent a signal peptide sequence, cleavage at the alanine residue would result in an amino-terminal glutamic acid residue of the protein (Fig. 3). It is known that the amino-terminal glutamic

acid residues of proteins may be blocked (22, 33). This could explain the previous failure to determine the amino terminus of the HA1 subunit (9). Although the amino acid sequence analysis was performed on the glycoprotein derived from a different influenza C virus strain (9), it is likely that these regions are conserved among different influenza C virus proteins. There are eight potential glycosylation sites (Asn-X-Thr/Ser) in the RNA 4 protein, which suggests that this protein is glycosylated in vivo (see Fig. 6).

Based on these findings, we postulate that RNA 4 in influenza C/Cal/78 virus codes for the glycoprotein precursor (HA) of the virus. It is suggested that processing of the precursor leads to removal of the hydrophobic signal peptide sequence, resulting in a glycoprotein (gpI) described earlier (9). Secondary post-translational cleavage at position 445 (Fig. 3) leads to at least two subunits, HA1 and HA2 (see Table 1), comparable in size to the HA subunits of influenza A and B viruses. The estimated molecular weights of the two subunits are compatible with those (gp65 and gp30) reported by Herrler et al. (9). However, in the absence of additional amino acid sequence data and of studies on the glycosylation levels, it is not possible to define precisely the lengths of the amino acid backbones of the two subunits. Additional data are also necessary for elucidating the structure of the second form of the large glycoprotein, gpII or gp88, which has been previously described (9).

Comparison of influenza A, B, and C virus HAs. Previously, we have reported on the evolutionary relationship of influenza A and B virus HAs (14, 15). Structural similarities as well as nucleotide and amino acid sequence homologies

FIG. 4. Relative hydrophilicity plots of the influenza A/PR/8/34 and B/Lee/40 virus HAs and of the influenza C/Cal/78 virus RNA 4 protein. The plots were generated as described in the text. In plots of the A and B virus HAs, the arrows at amino acid positions 17 (A/PR/8/34) and 15 (B/Lee/40) indicate the cleavage sites of the signal peptides, and the arrows at positions 344 (A/PR/8/34) and 361 (B/Lee/40) designate the HA1-HA2 cleavage sites. Brackets near the right of the figures for influenza A and B virus HAs correspond to the hydrophobic membrane-bound portions of the HA2s. In the case of the influenza C virus RNA 4 protein, the presumptive signal peptide and HA1-HA2 cleavage sites and the hydrophobic membrane-bound region of the HA2 are indicated with the same symbols as in the A and B virus HAs. The sequence and the relative hydrophilicity plots of the influenza A/PR/8/34 and B/Lee/40 virus HAs are from references 3 and 14 and are included for purposes of comparison.

were detected between these molecules. Specifically, computer-generated comparison matrices analyzing the influenza B/Lee/40 and A/PR/8/34 virus HA sequences revealed long regions of homology in the HA2 subunit and in the amino-terminal portion of HA1 (14). Similar analysis comparing the influenza C virus HA gene with that of the

TABLE 1. Comparison of human type A, type B, and type C influenza virus HA genes

Virus	Total length (nucleotides)	Subunit lengths (amino acids)	
		HA1	HA2
A/PR/8/34	1,778	326	222
B/Lee/40	1,882	346 ^a	223 ^a
C/Cal/78	2,071	431 ^a	209 ^a

^a Deduced length of subunits without taking into account any secondary processing.

influenza A or B virus HA genes failed to reveal homology. These analyses employed stringency conditions to detect homologies of 80, 50, and 40% in windows 10, 50, and 200 nucleotides long, respectively.

In comparing protein sequences, a window of five amino acids was used, and a dot was generated when the amino acids in three out of five positions were identical. Although such a low-stringency condition was used, no significant sequence homology was detected between the B and C viral HAs (Fig. 5) or A and C viral HAs (data not shown). The only short stretch of amino acid homology between the B and C viral HAs was a match of six out of seven amino acids in positions 359 through 365 of the C virus HA corresponding to positions 302 through 308 in the B virus HA (Fig. 5).

Based on available computer programs, it was thus not possible to align the A and C or B and C viral HAs either by their nucleotide or by their amino acid sequences. However, a preliminary alignment of the A, B, and C virus HAs was obtained emphasizing structural criteria, i.e., similarities among the hydrophilicity patterns and the known amino termini of the HA2s. Three domains were distinguished and were aligned separately: the signal peptides and the HA1 and the HA2 portions of the molecules (see Fig. 6).

Examining the HA2 domain first (Fig. 6), it is possible to align three internal cysteine residues (at positions 582, 590, and 594) with those of the previously aligned A and B virus HA2 subunits (at positions 481, 488, and 492 of the A virus HA and 498, 506, and 510 of the B virus HA [14]). This alignment requires the introduction (Fig. 6) of only two single amino acid deletions (at positions 463 and 589) and one

FIG. 5. Computer-generated graphic comparison of the deduced amino acid sequence of the influenza C/Cal/78 virus RNA 4 and the influenza B/Lee/40 virus HA gene. Each dot represents a sequence of five amino acids in which there is exact homology at three or more positions. The amino acid sequence of the influenza B/Lee/40 virus HA is from reference 14.

FIG. 6. Tentative alignment of the deduced amino acid sequences coded for the influenza A/PR/8/34, B/Lee/40, and C/Cal/78 virus HA genes. Alignment was done independently for the signal peptide, the HA1 region, and the HA2 region. This alignment emphasizes structural features (hydrophobic regions, conserved cysteine residues, known cleavage sites) and leaves the previously published alignment of the A and B virus HAs intact (14); the numbering system refers only to the C virus sequence. For example, it is assumed that the alanine residue (position 14) before the glutamic acid residue is the cleavage site of the signal peptide, and we have therefore introduced a gap three amino acids long into the C virus HA to allow alignment with the A and B virus HAs. It should be noted that the alignment shown represents only one of several possible alignments of the three sequences. The aligned cysteine residues in the three sequences are shown with open circles, and nonconserved cysteine residues are indicated with open diamonds. All other conserved residues are enclosed within rectangles. The closed triangles above the sequences indicate potential glycosylation sites. The open triangles above the sequences denote the proposed sialic acid binding sites in the type A and B virus HAs. The deletions and insertions which allow alignment of the three sequences are indicated by bars. Arrows show (possible) cleavage sites in the HA precursors. Solid lines indicate disulfide bonds which appear to be conserved in all three HAs.

15-amino acid insertion (between residues 627 and 628). An additional cysteine residue, conserved in the carboxyl terminus of the A and B virus HA2s, is also present in the carboxyl terminus of the C virus HA2 (position 651). It is likely that the conservation of cysteine residues in the HA2s is of structural importance for the HAs. It should be noted, however, that it is difficult to detect further amino acid homology among the A and C or B and C HA2 portions of these HAs. In particular, the highly conserved amino terminal sequence of the A and B virus HA2s, responsible for cell fusion activity, is not conserved in the C virus HA2. Nevertheless, the presumptive amino terminus of the C virus HA2 is relatively hydrophobic, and the virus is able to fuse cells (13, 21).

Similar findings were obtained regarding structural homologies in the HA1 portions of the three molecules. Close to

the amino terminus of the HA1, a cysteine residue (position 20) is found in all three HAs. This amino-terminal cysteine residue was found by three-dimensional X-ray crystallography to be cross-linked with the first cysteine residue of the HA2 (position 582) in the A virus HA (32). We would like to speculate that these cysteine residues have a similar function in the C virus HAs. With respect to the carboxyl termini of the HA1s, a single insertion of seven amino acids into the C virus HA1 (positions 436 through 442) permits alignment of the two carboxyl-terminal cysteine residues (positions 386 and 412) of all three HA1s. These two cysteine residues were previously shown to form a disulfide bridge (32). Again, we would postulate that these cysteine residues provide a framework for the C virus HA, which is similar to that of the A and B virus HAs. It should be noted, however, that no other sequence homologies can be detected among the three

HA molecules in this region. By introducing several insertions and deletions, it is possible to align additional cysteine residues (positions 119, 139, 164, and 209) of the C virus HA1 with those of the A and B virus HAs. These four cysteine residues have been shown to form internal disulfide bridges in the A virus HA (32).

Finally, examination of the third domain, containing the presumptive signal peptide of the C virus HA precursor, again reveals apparent conservation of a hydrophobic sequence but not of the precise amino acid sequences found in the signal peptides of the A and B virus HA precursors. In summary, the alignment of the amino acid sequence of the C virus HA remains tentative since it is based solely on structural homologies and thus may have to be revised once three-dimensional structural data have been obtained for the influenza C virus HA.

Glycosylation sites. The predicted amino acid sequence contains six potential glycosylation sites (Asn-X-Thr/Ser) in the HA1 portion and two in the HA2 portion of the C virus HA (Fig. 6). However, a conservation of the glycosylation sites of the influenza A, B, and C virus HAs is not obvious. This is not surprising because there is variation in glycosylation sites, even among A virus HAs (7, 10). It is not known which of the potential glycosylation sites in the C virus HA is used in vivo.

Sialic acid binding sites. We have reported previously that the proposed sialic acid binding sites in the HA1 subunit of the influenza A virus (7, 10) are conserved in the B virus HAs except for minor changes (15). These residues are, however, not found in the C virus HA1 molecule (Fig. 6). This is consistent with the fact that sialic acid is not a component of the cell receptor for influenza C virus (11).

Divergent versus convergent evolution of the influenza virus HAs. Although it is difficult to detect nucleotide and amino acid sequence homologies between the influenza C/Cal/78 virus HA and the A and B virus HAs, there is significant structural homology among the three molecules as a consequence of conserved hydrophobic domains and cysteine residues. Furthermore, cell fusion is promoted by influenza C viruses, and this activity is most likely mediated by the HA, as is agglutination of erythrocytes. Thus, there are structural and functional similarities among the HAs of the influenza A, B, and C viruses.

The question then arises whether the influenza C virus HA arose by divergent evolution from a common ancestral gene or by convergent evolution to provide for similar activities. The lack of sequence homology between the influenza C virus HA and the A or B virus HAs argues for convergent evolution. For example, it is assumed that subtilisin and alpha-chymotrypsin, which lack sequence homologies, have similar catalytic functions as a result of convergent evolution (27). Similarly, thermolysin and carboxypeptidase A also lack sequence homologies but share enzymatic activity patterns (12). On the other hand, lysozymes from chicken egg white, goose egg white, and bacteriophage T4 also lack detectable sequence homologies. It has been postulated, however, that these latter proteins have a common ancestral gene because different elements of the protein structure have been retained in each species (8, 17, 18). The lysozyme data suggest that the amino acid sequence of a protein may change more rapidly than its tertiary structure during evolution. Accordingly, the influenza virus HAs also appear to have arisen from a common ancestor because structural similarities are still present, although sequence homology may be absent.

We believe that, in addition to the lysozyme system, the

influenza A, B, and C virus HAs now represent a second example, among proteins lacking extensive sequence homology, of divergent evolution.

ACKNOWLEDGMENTS

This work was supported in part by Public Health Service grants AI 11823 and AI 18998 from the National Institutes of Health. P.P. is a recipient of an I. T. Hirschl Career Research Award. M.K. is the recipient of an American Lung Association postdoctoral fellowship. R.S.C. is the recipient of a Parker B. Francis Fellowship in pulmonary research.

LITERATURE CITED

1. Air, G. M., and R. W. Compans. 1983. Influenza B and influenza C viruses, p. 280-304. In P. Palese and D. W. Kingsbury (ed.), *Genetics of influenza viruses*. Springer-Verlag KG, Vienna.
2. Alwine, J. C., D. J. Kemp, B. A. Parker, J. Reiser, J. Renart, G. R. Stark, and G. M. Wahl. 1979. Detection of specific RNAs or specific fragments of DNA by fractionation in gels and transfer to diazobenzylloxymethyl paper. *Methods Enzymol.* 68:220-242.
3. Baez, M., R. Taussig, J. J. Zazra, J. F. Young, and P. Palese. 1980. Complete nucleotide sequence of the influenza A/PR/8/34 virus NS gene and comparison with the NS genes of the A/Udm/72 and A/FPV/Rostock/34 strains. *Nucleic Acids Res.* 8:5845-5858.
4. Compans, R. W., D. H. L. Bishop, and H. Meier-Ewert. 1977. Structural components of influenza C virions. *J. Virol.* 21:658-665.
5. Desselberger, U., and P. Palese. 1978. Molecular weights of RNA segments of influenza A and B viruses. *Virology* 88:394-399.
6. Desselberger, U., V. R. Racaniello, J. J. Zazra, and P. Palese. 1980. The 3' and 5'-terminal sequences of influenza A, B, and C virus RNA segments are highly conserved and show partial inverted complementarity. *Gene* 8:315-328.
7. Gething, M.-J., J. Bye, J. Skehel, and M. Waterfield. 1980. Cloning and DNA sequence of double-stranded copies of haemagglutinin genes from H2 and H3 strains elucidates antigenic shift and drift in human influenza virus. *Nature (London)* 287:301-306.
8. Grutter, M. G., L. H. Weaver, and B. W. Matthews. 1983. Goose lysozyme structure: an evolutionary link between hen and bacteriophage lysozymes? *Nature (London)* 303:828-831.
9. Herrier, G., A. Nagele, H. Meier-Ewert, A. S. Bhowan, and R. W. Compans. 1981. Isolation and structural analysis of influenza C virion glycoproteins. *Virology* 113:439-451.
10. Hiti, A. L., A. R. Davis, and D. P. Nayak. 1981. Complete sequence analysis shows that the hemagglutinins of the H0 and H2 subtypes of human influenza virus are closely related. *Virology* 111:113-124.
11. Kendal, A. P. 1975. A comparison of "influenza C" with prototype myxoviruses: receptor-destroying activity (neuraminidase) and structural polypeptides. *Virology* 65:87-99.
12. Kester, W. R., and B. W. Matthews. 1977. Comparison of the structures of carboxypeptidase A and thermolysin. *J. Biol. Chem.* 252:7704-7710.
13. Kitame, F., K. Sugawara, K. Ohwada, and M. Homma. 1982. Proteolytic activation of hemolysis and fusion by influenza C virus. *Arch. Virol.* 73:357-361.
14. Krystal, M., R. M. Elliott, E. W. Benz, Jr., J. F. Young, and P. Palese. 1982. Evolution of influenza A and B viruses: conservation of structural features in the hemagglutinin genes. *Proc. Natl. Acad. Sci. U.S.A.* 79:4800-4804.
15. Krystal, M., J. F. Young, P. Palese, I. A. Wilson, J. J. Skehel, and D. C. Wiley. 1983. Sequential mutations in hemagglutinins of influenza B virus isolates: definition of antigenic domains. *Proc. Natl. Acad. Sci. U.S.A.* 80:4527-4531.
16. Levitt, M. 1976. A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.*

- 104:59–107.
17. **Matthews, B. M., M. G. Grutter, W. F. Anderson, and S. J. Remington.** 1981. Common precursor of lysozymes of hen egg-white bacteriophage 4. *Nature (London)* **290**:334–335.
18. **Matthews, B. M., S. J. Remington, M. G. Grutter, and W. F. Anderson.** 1981. Relation between hen egg white lysozyme and bacteriophage T4 lysozyme: evolutionary implications. *J. Mol. Biol.* **147**:545–558.
19. **Maxam, A. M., and W. Gilbert.** 1980. Sequencing end-labeled DNA with base-specific chemical cleavages. *Methods Enzymol.* **65**:499–560.
20. **Nerome, K., M. Ishida, and M. Nakayama.** 1976. Absence of neuraminidase from influenza C virus. *Arch. Virol.* **50**:241–244.
21. **Ohuchi, M., R. Ohuchi, and K. Mifune.** 1982. Demonstration of hemolytic and fusion activities of influenza C virus. *J. Virol.* **42**:1076–1079.
22. **Ovchinnikov, Y. A., N. G. Abdulaev, M. Y. Feigina, A. V. Kiselev, and N. A. Lobanov.** 1979. The structural basis of the functioning of bacteriorhodopsin: an overview. *FEBS Lett.* **100**:219–224.
23. **Palese, P., V. R. Racaniello, U. Desselberger, J. F. Young, and M. Baez.** 1980. Genetic structure and genetic variation of influenza viruses. *Phil. Trans. R. Soc. Lond. Sect. B.* **288**:299–305.
24. **Racaniello, V. R., and P. Palese.** 1979. Isolation of influenza C virus recombinants. *J. Virol.* **32**:1006–1014.
25. **Ritchey, M. B., P. Palese, and E. D. Kilbourne.** 1976. RNAs of influenza A, B, and C viruses. *J. Virol.* **18**:738–744.
26. **Robertson, J. S., M. Schubert, and R. A. Lazzarini.** 1981. Polyadenylation sites for influenza virus mRNA. *J. Virol.* **38**:157–163.
27. **Robertus, J. D., R. A. Alden, J. J. Birktoft, J. Kraut, J. C. Powers, and P. E. Wilcox.** 1972. An X-ray crystallographic study of the binding of peptide chloromethyl ketone inhibitors to subtilisin BPN'. *Biochemistry* **11**:2439–2449.
28. **Southern, E. M.** 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J. Mol. Biol.* **98**:503–517.
29. **Staden, R.** 1977. Sequence data handling by computer. *Nucleic Acids Res.* **4**:4037–4051.
30. **Staden, R.** 1978. Further procedures for sequence analysis by computer. *Nucleic Acids Res.* **5**:1013–1015.
31. **Staden, R.** 1979. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res.* **6**:2601–2610.
32. **Wilson, I. A., J. J. Skehel, and D. C. Wiley.** 1981. Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3 Å resolution. *Nature (London)* **289**:366–373.
33. **Wong, R. S. C., T. Hofmann, and A. Bennick.** 1979. The complete primary structure of a proline-rich phosphoprotein from human saliva. *J. Biol. Chem.* **254**:4800–4808.