
Complete nucleotide sequence of the haemagglutinin gene from a human influenza virus of the Hong Kong subtype

G.W.Both and M.J.Sleigh

CSIRO, Molecular and Cellular Biology Unit, P.O. Box 184, North Ryde, N.S.W. 2113, Australia

Received 19 May 1980

ABSTRACT

The complete nucleotide sequence has been determined for a cloned double-stranded DNA copy of the haemagglutinin gene from the human influenza strain A/NT/60/68/29C, a laboratory-isolated variant of A/NT/60/68, an early strain of the Hong Kong subtype. The gene is 1765 nucleotides long and contains information sufficient to code for a protein of 566 amino acids, which includes a hydrophobic leader peptide (16 residues), HA1 (328), HA2 (221) and an arginine residue which joins the HA subunits. Comparison of the predicted amino acid sequence for 29C haemagglutinin with protein sequence data available for HA from other influenza strains shows that no potential coding information is lost by processing of the mRNA.

A comparison of the amino acid sequences predicted from the gene sequences for 29C and fowl plague virus haemagglutinins, (1) indicates the extent to which changes can occur in the primary sequence of different regions of the protein, while maintaining essential structure and function.

INTRODUCTION

The genome of influenza A virus is segmented and consists of eight single stranded RNA species of negative polarity. The fourth largest segment codes for the viral haemagglutinin (HA) and the sixth for neuraminidase (2-7). The virus is notable for the frequency with which alterations in these two surface proteins are observed, changes in their structure resulting in changes in viral antigenic character. Antigenic shift occurs when there is a radical change in the antigenicity of the surface proteins leading to the appearance of a new viral subtype, while antigenic drift results from smaller, progressive changes in antigenicity within a subtype (8).

In an attempt to relate changes in viral antigenicity to changes in the primary structure of the major antigenic protein, haemagglutinin, peptide maps and amino acid sequences of this protein prepared from different viral

strains have been compared (9,10). However, the development of techniques for cloning double-stranded (ds) DNA copies of RNA genes and for rapid nucleotide sequencing has made it easier to study antigenic variation at the level of the nucleic acid. As a prelude to comparative sequence analysis of influenza HA genes, we synthesized a dsDNA copy of the HA gene and cloned it by insertion into the plasmid pBR322, amplified in *E. coli* RRI (7,11). Here we report the complete sequence of the HA gene from influenza strain A/NT/60/68/29C, a laboratory-derived mutant produced from A/NT/60/68, an early field isolate in the Hong Kong subtype (12,13).

MATERIALS AND METHODS

Growth and Purification of Virus. The virus strain A/NT/60/68/29C, supplied by Dr. C. Hannoun was grown and purified by Drs. V. Bender and B. Moss, as previously described (11).

Synthesis, cloning and characterisation of a dsDNA copy of the HA gene.

Procedures for the extraction of viral RNA, the synthesis of a dsDNA copy of the HA gene, its insertion into pBR322 and amplification in *E. coli* RRI have been described (7,11). (All recombinant DNA experiments were carried out under CII-EKI conditions as prescribed by the Recombinant DNA Committee of the Australian Academy of Science). The sequence inserted into pBR322 in clone C89 was previously identified as an authentic copy of the HA gene by comparing the nucleotide sequence of a small section (7) with the amino acid sequence determined for the corresponding region of the HA protein of the influenza strain A/Mem/102/72 (14).

Preparation of labelled restriction fragments. Plasmid DNA prepared from clone C89 (7,11) was digested for two hours with restriction enzymes in 10 μ l of buffer containing Tris-HCl, pH7.4 (6mM), NaCl (20mM), MgCl₂ (6mM), 2-mercaptoethanol (6mM) and 0.1 mg/ml bovine serum albumin. After digestion, the mixture was adjusted to give a concentration of Tris-HCl, pH 8.0 (55mM), KCl (40mM) and three unlabelled deoxynucleoside triphosphates (each 40 μ M). This solution was incubated for 15 min. at 37 $^{\circ}$ with 10-20 μ Ci of the fourth deoxynucleoside triphosphate, α ³²P-labelled, and 1 μ l (approx. 8 units) of AMV reverse transcriptase (kindly supplied by Dr. J.W. Beard, Life Sciences, Inc., St. Petersburg, Fla.). Restriction enzymes used for digestion were chosen such that only one end of the required DNA fragment could be labelled under the above conditions. Alternatively, after labelling, the digestion mixtures were heated to inactivate reverse transcriptase (70 $^{\circ}$, 15 min) and an unlabelled excess (1mM) of the radioactive deoxynucleoside triphosphate was

added. A second restriction enzyme digestion was then carried out. Labeled fragments were separated by electrophoresis on a 4% polyacrylamide gel (11) together with labeled restriction fragments of known size as markers. Appropriate fragments were extracted from the gel and sequenced by the method of Maxam and Gilbert (15).

Determination of gene sequence directly from viral RNA. The sequence at the 5' end of the HA gene, not represented in C89, was determined by the method of Sanger *et al.*, (16) using a denatured restriction fragment from C89 to prime DNA synthesis, with viral genome RNA as template (17).

Compilation and analysis of sequence data. Nucleotide sequence data were stored and analysed in a Digital PDP 11/10 computer, using programmes devised by Staden (18,19), kindly adapted for our system by Caroline Bucholtz and Dr. Alex Reisner. The HA proteins from fowl plague virus (FPV) and the Hong Kong subtype were compared using the hydrophobicity values for amino acids (20,21) as described by Bigelow (22) and computer programmes devised by Dr. Alex Reisner.

RESULTS

Characterisation of the cloned ds DNA copy of the HA gene from influenza strain A/NT/60/68/29C (7) included the derivation of a restriction map. This information was used to prepare suitable restriction fragments for nucleotide sequence analysis, resulting in the sequencing strategy shown in Fig. 1. Since data were available on the amino acid sequence of areas of the HA protein from another Hong Kong-type virus, A/Mem/102/72 (14), approximately 60% of the

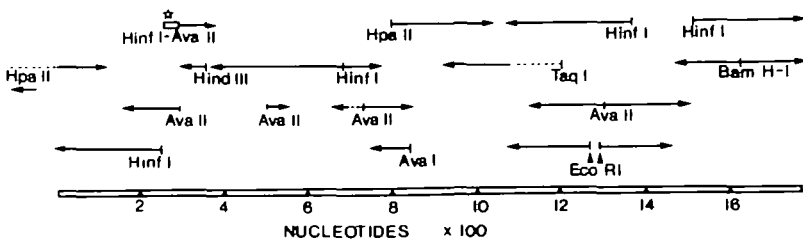


Figure 1. Strategy for sequencing a cloned dsDNA copy of the HA gene from strain 29C. The arrow shows the amount and the direction of the composite sequence information obtained from multiple experiments. (☆) The sequence of bases 300-370 was obtained using the Sanger chain termination method (16) copying the HA gene RNA into cDNA using the Hinf I - Ava II fragment as a primer for reverse transcriptase (17).

vRNA 3' -UCGUURUCGUCDCDCUUAUUAAGAUAUAUAG UAC UUC UGG UAG UAA CGA AAC UCG AUG UAA AAC ACA GAC CGA GAG CCG GUU CUG
 cRNA 5' -AGCAAAAGCAGGGGAUUAUUCUAUUAUUC MboI AAC ACC ADC AUU GCU UUG AGC UAC AUU UUC UGU CUG GCU CUC GGC CAA GAC HaeIII
 met lys thr ile ile ala leu ser tyr ile phe cys leu ala leu gly gln asp-2
 Precursor peptide

 GAA GCU CCU UUA CUG UUG UUG UGU CGU UGC GAC AGC GAC CCU GUA GUA CGC CAC GCU UUG CCU UGU GAU CAC UUU UGU UAG UGU
 100 EcoRII 150
 CUU CCA GGA AAU GAC AAC AAC ACA GCA AGC CUG UGC CUG GGA CAU CAU GCG GUG CCA AAC GGA ACA CUA GUC AAA ACA AUC ACA
 leu pro gly asp asp PstI PstI PstI ala thr leu cys leu gly his his ala val pro PstI PstI PstI leu val lys thr ile thr-30

 CUA CUA GUC UAA CUU CAC UGA UUA CGA UGA CUC GAU CAA GUC UCG AGG AGU UGC CCG UUU UAU AGC UUG UUA GGA GUA GCU UAG
MboI 200 Taq, HinfI
 GAU GAU CAG AUU GAA GUG ACU AAU GCU ACU GAG CUA CUU CAG AGC UCC UCA AGC GGG AAA AUA UGC AAC AAU CCU CAU CGA AUC
 asp asp gln ile glu val thr PstI PstI PstI glu leu val gln ser ser ser thr gly lys ile cys asp asp pro his arg ile-58

 GAA CUA CCU UAU CUG ACC UGU GAC UAU CUA CGA GAU AAC CCC CUG GGA GUA ACA CUA CAA AAA GUU UUA CUC UGU ACC CUG GAA
AvaiI 300 AvaiI
 CUU GAU GGA AUA GAC UGC ACA CUG AUA GAU GCU CUA UUG GGC GAC CCU CAU UGU GAU GUU UUU CAA AAU GAG ACA UGC GAC GCU
 leu asp gly ile asp cys thr leu ile asp ala leu leu gly asp pro his cys asp val phe gln PstI PstI PstI trp asp leu-86

 AAC CAA CUU CCG UCG UUU CGA AAG UCG UUG ACA AUG GGA AUA CUA CAC GCU CUA AUA CCG AGC GAA UCC AGU GAU CAA CGC AGC
 350 HindIII 400
 UUC GUU GAA CGC AGC AAA GCU UUC AGC AAC UGU UAC CCU UAU GAU GUG CCA GAU UAU GGC UCC CUU AGC UCA CUA GUU GCC UGC
 phe val glu arg ser lys ala phe ser asp cys tyr pro tyr asp val pro asp tyr ala ser leu arg ser leu val ala ser-114

 AGU CCG UGA GAC CUC AAA UAG UGA CUC CCA AAG UGA ACC UGA CCC CAG UGA GUC UUA CCC CCU UCG UUA CGA ACC UUU UCC CCU
 450 500
 UCA GGC ACU CUG CAG UUU AUC ACU GAG GCU UUG ACU UGC ACU GGG GUC ACU CAG AAU GGG GGA ACC AAU GCU UGC AAA AGC GGA
 ser gly thr leu glu phe ile thr glu gly phe thr trp thr gly val thr gln asp gly gly ser asp ala cys lys arg gly-142

 GGA CUA UCG CCA AAA AAG UCA UCU GAC UUG ACC AAC UGG UUU AGU CCU UCG UGU AUA GGU CAC GAA UUG CAC UGA UAC GCU UUG
AvaiI HindIII 550
 CUU GAU AGC GGU UUU UUC AGU AGA CUG AAC UUG UGC ACC AAA UCA GGA AGC ACA UAU CCA GUG CUU AAU GUG ACU AUG CCA AAC
 pro asp ser gly phe phe ser arg leu asp trp leu thr lys ser gly ser thr tyr pro val leu PstI PstI PstI met pro asp-170

 UUA CUG UUA AAA CUG UUU GAU AUG UAA ACC CCC CAA CUG GUG GGC UCC UGC UUG GUU CUU GUU UGG UCG GAC AUA CAA CUU CGU
 600 AvaiI 650
 AAU GAC AAU UUU GAC AAA CUA UAC AUU UGC GGG GUU CAC CAC CCG AGC ACG AAC CAA GAA CAA ACC AGC CUG UAU GUU CAA GCA
 asp asp asp phe asp lys leu tyr ile trp gly val his his pro ser thr asp gln glu gln thr ser leu tyr val gln ala-198

 AGU CCC UCU CAG UCU CAG AGA UGC UCC UCU UGC GUC GUR UGA UAU UAG GGC UUA UAG CCC AGC UCU GGG ACC CAU UCC CCA GUC
HinfI EcoRII 700 AvaiI EcoRII 750
 UCA GGC AGA GUC ACA CUC UCU ACC AGG AGA AGC CAG CAA ACU AUA AUC CCC AAU AUC GCG UCC ACA CCG UGG GUA AGC GGU CAC
 ser gly arg val thr val ser thr arg arg ser gln gln thr ile ile pro asp ile gly ser arg pro trp val arg gly gln-226

 AGA UCA UCU UAU UCG UAG AUA ACC UGU UAU CAA UUC GGC CCU CUG CAU GAC CAU UAA UUA UCA UUA CCC UUG GAU UAG CGA GGA
HpaII 800
 UCU ACU AGA AUA AGC AUC UAU UGC ACA AUA GUU AAG CCG GCA GAC GUA CUG GUA AUU AAU AGU AAU GGC AAC CUA AUC CCU CCU
 ser ser arg ile ser ile tyr trp thr ile val lys pro gly asp val leu val ile asp ser asp gly asp leu ile ala pro-254

 GCC CCA AUA AAG UUU UAC GGC UGA CCC UUU UCG AGL UAU UAG UCC AGU CUA CCU GGA UAA CUA UUG ACA UAA AGA CUU ACC UAG
AvaiI 850 HhaI 900
 CGC GGU UAU UUC AAA ADC CCC ACU GGC AAA AGC UCA AUA AUG AGG UCA GAU CCA CCU AUU GAA ACC UGU AUU UCU CAA UGC AUC
 arg gly tyr phe lys met arg thr gly lys ser ser ile met arg ser asp ala pro ile asp thr cys ile ser glu cys ile-282

 UGA GGU UUA CCU UCG UAA GGG UUA CUC UCC GGC AAA GUL UUC CAU UUG UUC UAG UGU AUA CCU CCU ACC GGC UUC AUA CAA UUC
 950 MboI
 ACU CCA AAU GGA AGC AUU CCC AAU GAC AAC CCC UUU CAA AAC GUA AAC AAG AUC ACA UAU GGA GCA UCC CCC AAG UAU GUU AAG
 thr pro PstI PstI PstI ile pro asp asp lys pro phe gln asp val asp lys ile thr tyr gly ala cys pro lys tyr val lys-310

 GUU UUC UGG GAC UUC AAC CGU UGU CCC UAC GCC UUA CAU GCU UUC UUU GUU UGA [UCL]
 CAA AAC ACC CUC AAC UUC CCA ACA GGC AUG CCG AAU GUA CCA GAG AAA CAA ACU [ACA]
 gln asp thr leu lys leu ala thr gly met arg asp val pro glu lys gln thr [arg]

Figure 2a. For legend see over page.

vRNA 3' CGC GAG AAC CCG CGU UAU CGU CCA AAG UAU CUU UUA CCA ACC CUC CCU UAC UAU CUC CCA ACC AUC CCA AAG UCC GUA
HaeIII HheI 1100
cRNA 5' UCC UUA UUC GGC GCA AHA CCA CGU UUC AHA GAA AAU CGU UGC GAG CCA AUC AHA GAC CGU UGC UAC CGU UUC AGG CAU
gly leu phe gly ala ile ala gly phe ile glu am gly trp glu gly met ile asp gly trp tyr gly phe arg his-26

CGU UUA AGA CUC CCG UGU CCU GGU CGU CGU CUA GAA UUU UCC UGA CUU CGU CCG UAG CUG GUU UAG UUA CCC UUU AAC UUC UCC
1150 MboI TaqI
CAA AAU UCU GAC CGC ACA GGA CAA GCA GCA CAU CUU AAA AGC ACU CAA CCA CCC AUC GAC CAA AUC AAU GGC AAA UUC AAC AGG
gln am ser glu gly thr gly gln ala ala asp leu lys ser thr gln ala ala ile asp gln ile am gly lys leu am arg-54

CAU UAG CUC UUC UGC UUG CUC UUU AAC GUA GUU UAG CUU UUC CUU AAG AGU CUU CAU CUU CCC UCU UAA GGC CUG GAC CUC UUU
TaqI MboII 1250 TaqI EcoRI
GUA AUC GAC AAC ACC AAC GAC AAA UUC CAU CAA AUC GAA AAG CAA UUC UCA CAA CUA GAA CCG AGA AUC CAC GAC CUC GAC AAA
val ile glu lys thr am glu lys phe his gln ile glu lys glu phe ser glu val glu gly arg ile gln asp leu glu lys-82

AUC CAA CUU CUG UGA UUU UAU CUA GAG ACC AGA AUC UUA CUC CUC GAA GAA CAG CCA GAC CUC UUA GUU GUA UGU UAA CUC GAC
MboII MboI 1350 HinfI
UAC GUU UCA GAC ACU AAA AHA CAU CUC UGC UCU UAC AAU CCG GAG CUC CUU CUC CCU CUG GAC AAU CAA CAU ACA AUC GAC CUC
tyr val glu asp thr lys ile asp leu trp ser tyr am ala glu leu leu val ala leu glu am gln his thr ile asp leu-110

UGA CUC AGC CUU UAC UUG UUC GAC AAA CUU UUU UGU UCC UCC GGU GAC UCC CUU UUA CCA CUU CUG UAC CCG UUA CCA ACC AAG
HinfI MboII 1450
ACU GAC UGC GAA AUC AAC AAC CUC UUU GAA AAA ACA AGC AGC CAA CUC AGC GAA AAU CCU GAA GAC AUC GGC AAU CGU UGC UUC
thr asp ser glu met am lys leu phe glu lys thr arg arg gln leu arg glu am ala glu asp met gly am gly cys phe-138

UUU UAU AUC CUC UUU ACA CUC UUG CCA AGC UAU CUC AGU UAG UCU UUA CCC UGA AHA CUC GUA CUA CAU AUC UCU CUC CUU CGU
1500 HinfI 1550
AAA AHA UAC CAC AAA UGU GAC AAC CUC UGC UHA GAG UCA AUC AGA AAU CCG ACU UAU GAC CAU CAU GUA UAC AGA GAC GAA GCA
lys ile tyr his lys cys asp am ala cys ile glu ser ile arg app gly tyr asp his asp val tyr arg asp glu ala-166

AAU UUG UUG GCC AAA CUC UAG UUU CCA CAA CUU GAC UUC AGA CCU AUC UUU CUG ACC UAG GAC ACC UAA AGC AAA CCG UAU AGU
HpaII MboI 1600 BamHI, MboI 1650
UUA AAC AAC CCG UUU CAG AUC AAA CGU GUU GAA CUC AAC UCU CGA UAC AAA GAC UGC AUC CUC UGC APU UCC UUU GCC AHA UCA
leu am am arg phe gln ile lys gly val glu leu lys ser gly tyr lys asp trp ile leu trp ile ser phe ala ile ser-194

ACG AAA AAC GAA ACA CAU CAA AAC GAC CCC AAG UAG UAC ACC CCG AGC CUC UCU CCG UGC UAA UCC ACG UGC UAA ACG UAA ACU
HaeIII 1700
UCC UUU UUC CUU UGU GUA CUU UUC CUC CCG UUC AUC AUC UGC CCG UCC CAG AGA CGC AAC APU ACG UGC AAC APU UCC APU UCA
cys phe leu leu cys val val leu leu gly phe ile met trp ala cys gln arg gly am ile arg cys am ile cys ile

|
CACAATAADCAUAAAUUUUUGCGGCAACAAGADCA -5'
1750
GUCUUAUACAAAUUAAAAACACCCUUCUUCACU -3'

Fig 2b

Figure 2. Nucleotide sequence of the HA gene from Hong Kong influenza strain 29C and the amino acid sequence predicted from it. The RNA sequence (-) strand is shown from 3' to 5' below it, the complementary (+) strand representing the mRNA sequence. Initiation and termination codons are boxed and the arginine residue which connects HA1 (Fig. 2a) and HA2 (Fig. 2b) is bracketed. Possible glycosylation sites in the protein are underlined with dots. The end of the clone is indicated by the vertical line to the right of the termination codon. Restriction sites in the plasmid DNA are indicated in the equivalent position on the mRNA sequence.

gene copy was sequenced on one DNA strand only. Adjoining sections of sequence overlapped by a minimum of 15 nucleotides, except in the region of the Hind III site (base 353), where the sequence was confirmed from the viral RNA itself, using the chain termination sequencing method (16). A denatured 51-base DNA fragment, obtained by digestion of C89 DNA with Hinf I and Ava II, was used as

Downloaded from https://academic.oup.com/nar/article/8/12/2561/12381014 by guest on 17 April 2024

a primer for DNA synthesis (17). A similar technique was used in an attempt to obtain the 5' terminal gene sequence, which was not represented in the cloned gene (7).

Figure 2 shows the nucleotide sequence determined for the cloned dsDNA copy of the HA gene from strain 29C and the amino acid sequence predicted for its protein. The cloned gene copy contains 1739 nucleotides, commencing from the 3' terminal base of the gene, with the first 12 bases identical to the common sequence found at the 3' termini of other influenza genome segments (23,24). The cloned sequence extends nine bases beyond a termination codon in the same phase as the only reading frame that is continuous for the length of the gene. Part of the sequence shown for the 5' terminal region of the gene beyond the end of the clone must be regarded as tentative. The sequence shown is identical to that obtained from a cloned copy of this section of the HA gene from the 29C parent strain, A/NT/60/68 (25). Attempts to determine the sequence in this region directly from the 29C viral RNA gave clear results between bases 1734-1744 and 1752-1763, the latter segment lying within a sequence common to the 5' termini of all influenza genes so far examined (23,24). This leaves in doubt a section of 7 nucleotides, whose sequence appeared to be the same as that in A/NT/60/68, but for which unequivocal data could not be obtained (data not shown).

Possible deletion of a base during cloning of a gene copy

The amino acid sequence data of Ward and Dopheide (14) enabled us to determine the correct reading frame for the nucleic acid sequence of the ds DNA copy of 29C HA. However, reading backwards in this frame towards the N-terminus of HA1, our initial sequence for 29C contained an in-phase termination codon at bases 95-97 (Fig. 3a) and no in-phase ATG codon. The sequence of both strands of the cloned insert agreed in this respect (data not shown). We therefore attempted to confirm the sequence of this region directly by using a MboII/Hae III fragment (bases 45-76 of the cloned insert) as a primer for cDNA synthesis, with 29C genome RNA as a template (17). The sequence of the HA gene thus derived included an extra A residue at position 107 in the plus strand (Fig. 3b) which provided a continuous reading frame back to the ATG codon at bases 30-32 and yielded an amino acid sequence compatible with that determined for the N-terminus of mature HA from A/Mem/102/72 (26). We also determined the nucleotide sequence in this region for C55, another plasmid containing a dsDNA copy of the HA gene from 29C, isolated with C89 from the same *E. coli* RRI transformation. Unlike C89, this gene insert contained the A/T base pair at position 107 (data not shown).

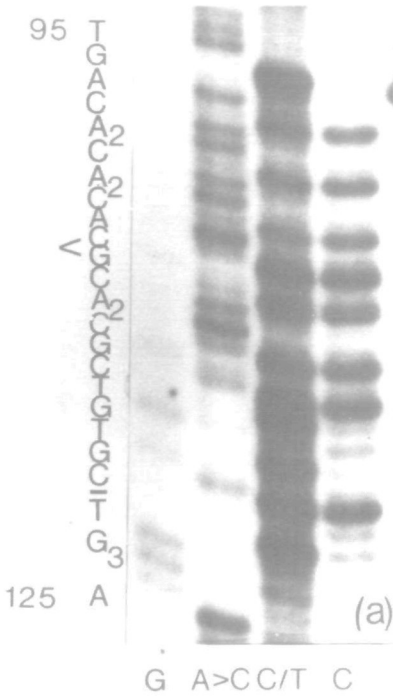
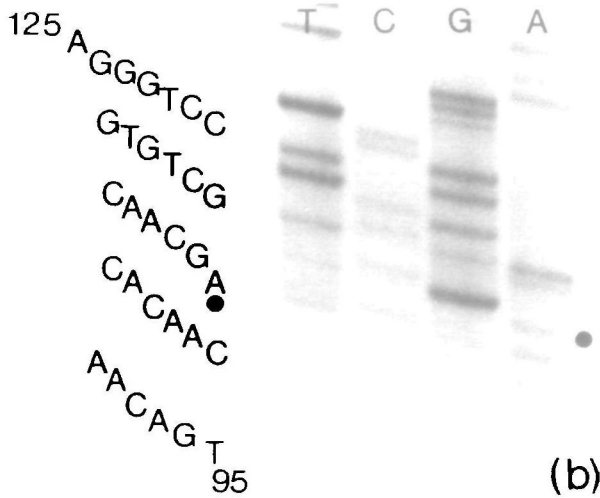


Figure 3. Comparison of (+) strand DNA sequences between bases 95-125. (a) A *Hinf* I/*Hae* III fragment labelled at the *Hinf* I site was sequenced by the Maxam and Gilbert procedure (15). The position of the missing base is indicated (<). The base marked (-) at position 120 is a C residue and is part of an *Eco* RII restriction endonuclease site which is methylated when the hybrid plasmid is grown in *E. coli* RRI. (b) 29C genome RNA was used as a template for cDNA synthesis by reverse transcriptase using a *Mbo*II/*Hae* III primer. Sequence data was obtained by the "dideoxy" method (16). The apparent missing residue in the cloned DNA copy of the HA gene (see (a)) is indicated (●).



DISCUSSION

Apparent deletion of a base from the HA gene copy in plasmid C89. A comparison of the nucleotide sequences determined for HA genes (bases 95-125) from the (+) strand of the cloned gene copies in C89 and C55 with the sequence obtained directly from the genome RNA indicates that at position 107, a residue present in the gene is missing in the C89 gene copy. This region of the HA gene can be drawn in a hairpin configuration (Fig. 4) with a stability of -4 Kcal (27). The presence of multiple bands on the sequencing gel (Fig. 3 b) between positions 103 and 111 may indicate that the hairpin structure is sufficiently stable to present reverse transcriptase with some difficulty in negotiating the 3' proximal side of the base-paired region. We speculate, therefore, that the presence of this hairpin may result in incorrect copying of the RNA by reverse transcriptase. Both Porter *et al.* (1), in cloning the FPV HA gene and Richards *et al.* (28), in studying copies of chicken β -globin mRNA found evidence for altered and missing bases in cloned DNA. However, they attributed this to repair or incorrect copying of mismatched regions associated with the terminal loop priming second strand DNA synthesis.

While it is possible that the HA gene copy in C89 represents a variant gene present in the viral population, such a deletion mutant should be extremely rare, since the deletion would result in the premature termination of synthesis of the HA protein, and this would be lethal in the next generation. Because the reverse transcriptase lacks a 3' exonuclease which could edit mistakes, it is possible that errors may occur with low frequency during

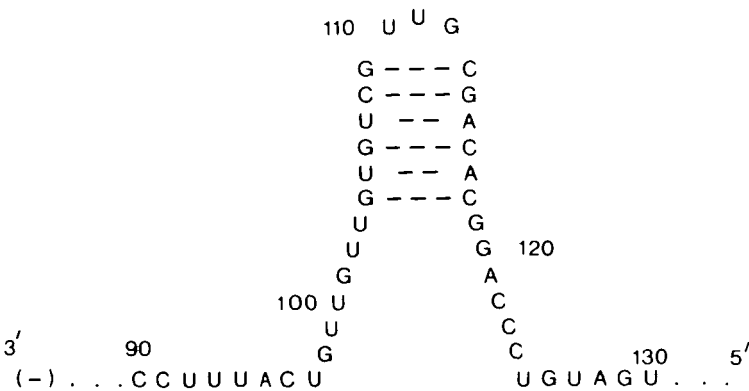


Figure 4. Structure of a hairpin loop which could form in the region of bases 100-120 of the gene.

the multi-step cloning procedure. Therefore, to guard against such errors when studying genes for which no protein sequence data are available, it may be necessary to derive nucleotide sequences from more than one cloned gene copy.

Structure of the HA gene from influenza of the Hong Kong subtype. Analyses by restriction enzyme mapping (7), nucleotide sequencing of the cloned HA gene copy and determination of the terminal sequence of the gene itself, revealed a length of 1765 nucleotides for the HA gene from the Hong Kong influenza strain 29C. This agrees with our previous estimate (1760 nucleotides) based on electrophoretic mobility (11) and compares with a length of 1742 nucleotides for the HA gene from the avian influenza strain FPV (Rostock) (1).

The arrangement of the HA genes from 29C and FPV are compared in Fig. 5. At the 3' end of the negative (genome) strand is a non coding sequence which appears to be completely transcribed into cRNA in vitro (23) and in vivo probably forms the 5' non-translated region of the mRNA. This section of mRNA may be subsequently modified in vivo if host-derived sequences and m⁷G caps are attached (29).

Of the potential initiation codons in the (+) strand, only the one following the first 29 bases is in the correct phase to provide a continuous reading frame, which is also the frame prescribed by the known amino acid sequences for HA from the Hong Kong strain A/Mem/102/72 (14,26). The next AUG in this phase occurs 578 bases into the gene. Commencement of protein synthesis at bases 30-32 would produce a very hydrophobic peptide of 16 amino acids preced-

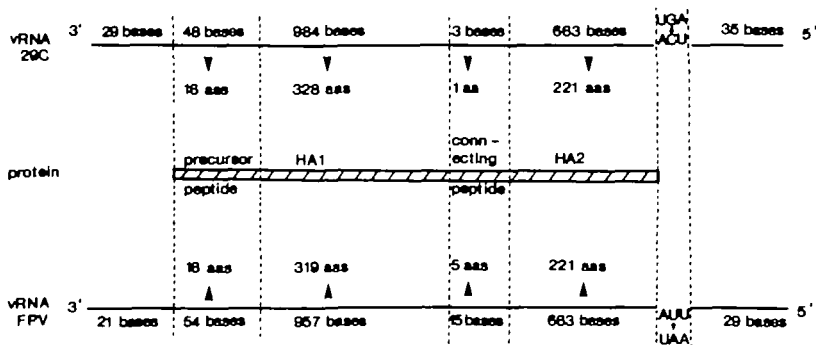


Figure 5. Comparison of the HA gene structures for Hong Kong and Fowl Plague viruses.

ing the glutamine residue (bases 79-81) found to be the N-terminal amino acid of the mature HA protein from A/Mem/102/72 (25).

The major and minor subunits (HA1 and HA2 respectively) of the mature HA protein appear to be generated by proteolytic cleavage of the primary translation product, with the loss of some amino acids connecting the two sections (30). Aligning the amino acid sequence found at the end of the HA1 and the beginning of HA 2 for influenza A/Mem/102/72 (15) with the amino acid sequence predicted by the HA gene from 29C, suggests that the connecting peptide consists of a single arginine residue. The HA subunits of A/Vic/3/75 are also linked by one arginine residue (31). In this respect, the HA of these strains resembles the H2-type HA from the Asian influenza strain A/Jap/305/57 (32) but differs from the FPV protein, where the HA subunits are connected in the immature protein by a basic pentapeptide (1).

The first in-phase termination codon (Fig. 5) is followed by only a short non coding sequence. How much of this sequence is transcribed into mRNA is not known, but it has been suggested that the U-rich sequence in the gene in this region may signal the end of transcription (1), providing a site for addition of poly A to the mRNA. Thus the 3' non-translated region of the mRNA following the termination codon could be as short as 14 bases in Hong Kong HA and 6 bases in FPV HA.

The amino acid sequence predicted from nucleotide sequence data for the HA gene of influenza A/Vic/3/75 (31) contained an additional asparagine residue following HA1 residue No. 8 (Fig. 2a). However, this additional residue may be unique to the particular isolate studied, since it is absent from H3-type HA1's in a total of six influenza strains isolated between 1968 and 1977. (Both and Sleight, unpublished results).

Comparison of nucleic acid sequences of Hong Kong and FPV HA genes. The genes from the two subtypes have similar base compositions: for 29C A24%, G 20.5%, C23.5%, U 32% and for FPV, A24%, G 18.4%, C 23.8%, U 33.8%. Codon utilisation in the Hong Kong HA gene is similar to that for FPV, with some exceptions which may reflect the availability of isoacceptor tRNAs in the host, e.g. CUG is preferred for leu in the Hong Kong gene while FPV uses AAA for lys in preference to AAG (Table 1). The incidence of CpG dinucleotides is low for both genes, as noted previously for FPV (1).

Comparison of amino acid sequences predicted by the two genes. The amino acid sequence predicted from the nucleotide sequence for the 29C HA gene (Fig. 2) is largely identical to that found for the HA protein from A/Mem/102/72 (14,26). As for HA molecules from other influenza strains, HA1 has a high

Table 1: Codon utilization in HA genes from Hong Kong and Fowl Plague Influenza Viruses. FPV data is in brackets below the corresponding figure for 29C.

	U	C	A	G		U	C	A	G
	9	6	9	6		9	6	8	0
U	(14)	(4)	(7)	(7)	U	(9)	(3)	(7)	(1)
	13	4	9	12		3	4	3	2
C	(12)	(4)	(8)	(9)	C	(4)	(4)	(4)	(0)
U	1	10	0	1	C	7	7	16	1
A	(6)	(13)	(1)	(0)	A	(4)	(6)	(14)	(4)
	7	2	0	12		16	3	8	3
G	(9)	(1)	(0)	(8)	G	(8)	(3)	(11)	(3)
	14	16	22	3		11	10	13	10
U	(15)	(13)	(23)	(11)	U	(9)	(12)	(17)	(5)
	18	6	21	14		4	5	20	7
C	(9)	(12)	(14)	(7)	C	(6)	(3)	(9)	(10)
A	14	13	19	10	G	9	11	16	13
A	(16)	(15)	(24)	(14)	A	(5)	(16)	(28)	(20)
	9	4	11	11		6	2	13	14
G	(11)	(1)	(7)	(8)	G	(10)	(1)	(10)	(15)

proline content relative to HA2. Also remarkable is the similarity with other strains in the number and distribution of cysteine residues in the 29C protein (9 in HA1, 8 in HA2) (1,14,30). Only one near the end of HA2 has no counterpart in the FPV molecule. If the FPV and Hong Kong HA amino acid sequences are aligned for maximum homology using the cysteine residues, seven of the ten proline residues in the C-terminal half of the HA1 are also conserved between the subtypes. This suggests that the shape of this part of the molecule is not permitted to vary extensively.

Potential sites for carbohydrate attachment (Fig.2), occurring (by analogy with HA from the Asian subtype) at sequences of the type Asn-X-Thr (30), are not conserved between subtypes. With the cysteine residues aligned, the sites at positions 22 and 38 in 29C are equivalent to those at 12 and 28 in FPV (1).

With the cysteine residues aligned, there is approximately 38% amino acid conservation in HA1 between FPV and 29C. In HA2 there is 65% homology, but in more than half of the 145 cases where the amino acid is conserved a different codon is used; 69 differ by one base, 5 differ by two

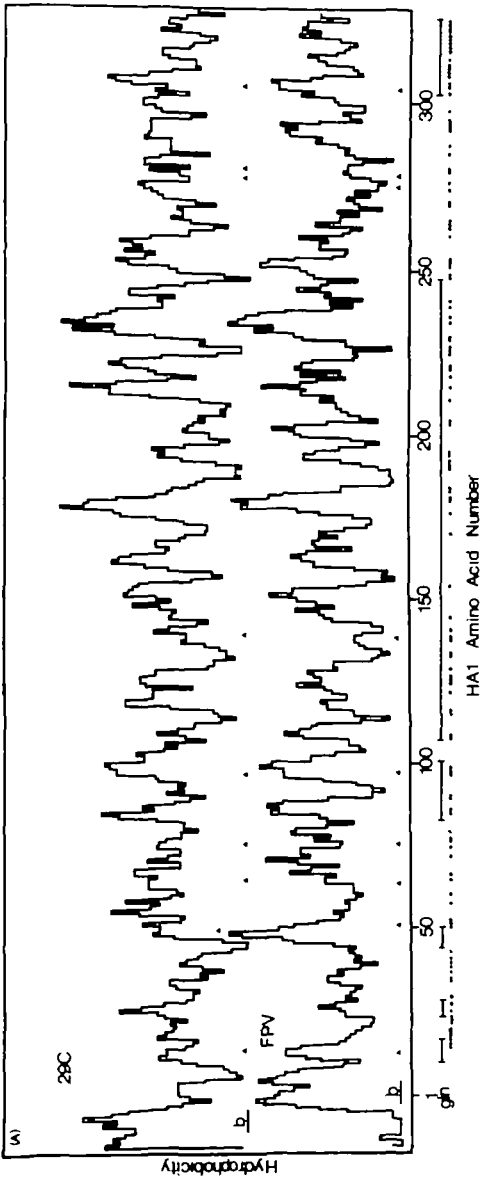


FIGURE 6. Relative hydrophobicities of the HA amino acid chains from FPV and strain 29C (a) HA1, (b) HA2. Computer-generated plots show hydrophobicities (20,21) as a moving average over five amino acids. In HA1 the two sequences were aligned to give maximum amino acid homology by introducing a single gap after residues 168 and 276 in 29C, and after residue 252 in FPV. Regions of similar hydrophobicity profile are indicated by solid lines below the figures, with homologous amino acids indicated by a dot. Cys residues are shown (▲) and a base line for each curve is indicated b.

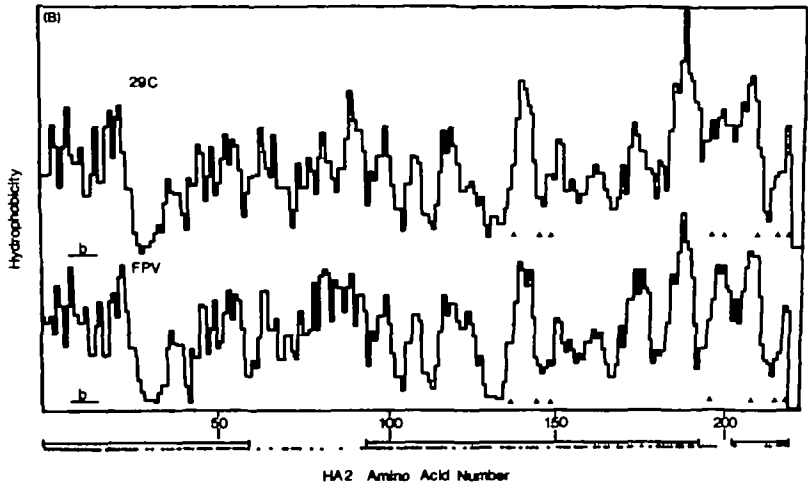


Fig. 6b

and in one case a serine uses an AGC instead of a UCA codon. Some areas of HA2 show a particularly high degree of amino acid conservation, e.g. the N-terminal region. In addition, in some areas of HA2 where the amino acid sequence is different, the character of the protein tends to be preserved. Figure 6 shows an analysis of the degree of hydrophobicity of different areas of the HA protein from 29C and FPV. In the C-terminal region of HA2, thought to be involved in anchoring the HA to the viral lipoprotein membrane (30), both proteins are highly hydrophobic in character, even though between residues 199 and 212, only one out of 13 amino acids is conserved. This effect extends to other regions of the HA as well. For example, the precursor peptides, cleaved from HA during maturation, differ in length and sequence among FPV, 29C and viruses from the H2 subtype (1, 32, 33); but are all hydrophobic in character. Also notable is the area between HA1 residues 85 and 240 of 29C for which the hydrophobicity profile is broadly similar to the equivalent area in HA1 of FPV, although the amino acid sequences show only 32% homology. This type of analysis suggests that amino acid divergence between HAs from different subtypes may be strictly limited in some areas to those changes which do not significantly disturb the local environment, while in other areas (e.g. residues 1-100 of HA1) little constraint is apparent. As sequence information on HA molecules from further influenza subtypes becomes available, it should be possible to identify regions of the protein which are essential to maintain HA structure and function. In addition, comparison in this way of closely related proteins from viruses

of the same subtype may help to identify the amino acid changes which are important in altering viral antigenicity.

ACKNOWLEDGEMENTS

We wish to thank Dr. Bernie Moss and Dr. Vera Bender for growing and purifying the virus, Caroline Bucholtz and Dr. Alex Reisner for constructing and adapting computer programmes for sequence storage and comparison and Elizabeth Hamilton for competent technical assistance. We are grateful to Dr. C. Hannoun of the Pasteur Institute for supplying the strain A/NT60/68/29C. We thank Dr. A. Reisner and Dr. G. Grigg for reading the manuscript.

REFERENCES

- 1 Porter, A.G., Barber, C., Carey, N.H., Hallewell, R.A., Threlfall, G., and Emtage, J.S. (1979). *Nature* 282, 471-477.
- 2 Palese, P. and Schulman, J.L. (1976). *Proc. Natl. Acad. Sci. USA.* 73, 2141-2146.
- 3 Ritchey, M.B., Palese, P. and Kilbourne, E.D. (1976). *J. Virol.* 18 738-744.
- 4 Scholtissek, C., Harms, E., Rhode, W., Orlich, M. and Rott, R. (1976). *Virology* 74, 332-344.
- 5 Scholtissek, C. (1978). *Curr. Top. Microbiol. Immunol.* 80, 139-169.
- 6 Inglis, S.C., Carroll, A.R., Lamb, R.A. and Mahy, B.W.J. (1976). *Virology* 74, 489-503.
- 7 Sleigh, M.J., Both, G.W. and Brownlee, G.G. (1979). *Nucl. Acids Res.* 7, 879-893.
- 8 Stuart-Harris, C.H. and Schild, G.C. (1976). *Influenza. The Viruses and the disease.* pp 57-68, Edward Arnold, London.
- 9 Laver, W.G., Air, G.M., Webster, R.G., Gerhard, W., Ward, C.W. and Dopheide, T.A. (1979). *Virology* 98, 226-237.
- 10 Moss, B.A., Underwood, P.A., Bender, V.J. & Whittaker, R.G. (1980). In *Structure and Variation in Influenza Virus.* (W.G. Laver and G. Air, eds.) pp. 329-338, Elsevier, New York.
- 11 Sleigh, M.J., Both, G.W. & Brownlee, G.G. (1979). *Nucl. Acids Res.* 6, 1309-1321.
- 12 Fazekas de St. Groth, S. (1967). *Cold Spring Harbor Symp. Quant. Biol.* 32, 525-536.
- 13 Fazekas de St. Groth, S. & Hannoun, C. (1973). *C.R. Acad. Sci. Paris. Ser D.* 276, 1917-1920.
- 14 Ward, C.W. & Dopheide, T.A. (1979). *Brit. Med. Bull.* 35, 51-56.
- 15 Maxam, A.M. & Gilbert, W. (1977). *Proc. Natl. Acad. Sci. USA.* 74, 560-564.
- 16 Sanger, F., Nicklen, S. & Coulson, A.R. (1977). *Proc. Natl. Acad. Sci. U.S.A.* 74, 5463-5467.
- 17 Both, G.W., Sleigh, M.J., Bender, V.J., Moss, B.A. (1980). In *Structure and Variation in Influenza Virus.* (W.G. Laver and G. Air, Eds.) pp. 81-90 Elsevier, New York.
- 18 Staden, R. (1977). *Nucl. Acids Res.* 4, 4037-4051.
- 19 Staden, R. (1979). *Nucl. Acids Res.* 6, 2601-2610.
- 20 Tanford, C. (1962). *J. Am. Chem. Soc.* 84, 4240-4247.
- 21 Nozaki, Y. & Tanford, C. (1970). *J. Biol. Chem.* 246, 2211-2217.

-
- 22 Bigelow, C.C. (1967). *J. Theoret. Biol.* 16, 187-211.
 - 23 Skehel, J.J. & Hay, A.J. (1978). *Nucl. Acids Res.* 5, 1207-1219.
 - 24 Robertson, J.S. (1979). *Nucl. Acids Res.* 6, 3745-3757.
 - 25 Sleigh, M.J., Both, G.W., Brownlee, G.G., Bender, V.J. & Moss, B.A. (1980). In *Structure and Variation in Influenza Virus*. (W.G. Laver and G. Air, eds). pp. 69-80, Elsevier, New York.
 - 26 Ward, C.W. & Dopheide, T.A. (1980). *Virology*. In Press.
 - 27 Tinoco, I., jun., Borer, P.N., Dengler, B., Levine, M.D., Uhlenbeck, O.C., Crothers, D.M. & Gralla, J. (1973). *Nature New Biol.* 246, 40-41.
 - 28 Richards, R.I., Shine, J., Ullrich, A., Wells, J.R. & Goodman, H.M. (1979). *Nucl. Acids Res.* 7, 1137-1146.
 - 29 Krug, R.M., Broni, B.A. & Bouloy, M. (1979). *Cell* 18, 329-334.
 - 30 Waterfield, M.D., Espelie, K., Elder, K. and Skehel, J.J. (1979). *Brit. Med. Bull.* 35, 57-63.
 - 31 Min-Jou, W., Verhoeven, M., Devos, R., Saman, E., Huylebroeck, D., van Rompuy, L., Fang, R.X. & Fiers, W. (1980). In *Structure and Variation in Influenza Virus* (W.G. Laver and G. Air, eds) pp. 63-68, Elsevier, New York.
 - 32 Gething, M.J., Bye, J., Skehel, J.J. and Waterfield, M.D. (1980). In *Structure and Variation in Influenza Virus* (W.G. Laver and G. Air, eds) pp. 1-10, Elsevier, New York.
 - 33 Air, G.M. (1979). *Virology* 97, 468-472.

