Cell, Vol. 19, 683-696, March 1980, Copyright © 1980 by MIT

Complete Structure of the Hemagglutinin Gene from the Human Influenza A/Victoria/3/75 (H3N2) Strain As Determined from Cloned DNA

Willy Min Jou, Martine Verhoeyen, René Devos, Eric Saman, Rongxiang Fang, Danny Huylebroeck and Walter Fiers
Laboratory of Molecular Biology
State University of Ghent
Ledeganckstraat 35
B-9000 Ghent, Belgium
Geoffrey Threlfall, Christine Barber, Norman Carey and Spencer Emtage
Searle Research Laboratories
Lane End Road
High Wycombe
Bucks HP12 4HL, England

Summary

The complete sequence of a hemagglutinin (HA) gene of a recent human influenza A strain, A/Victoria/3/75, is 1768 nucleotides long and contains the information for 567 amino acids. It codes for a signal peptide of 16 amino acids, the HA1 chain of the mature hemagglutinin of 329 amino acids, a connecting region between HA1 and HA2 consisting of a single arginine residue and the HA2 portion of 221 amino acids. The sequence is compared with the hemagglutinin of two members of other subtypes, the human H2 strain A/Jap/305/57 and the avian Hav1 strain A/FPV/Rostock/34, and with one of the same H3 subtype, A/Memphis/3/72. To align the HA1 chain of different major subtypes several deletions/insertions of single amino acids must be invoked, but two more extensive differences are found at both ends, one leading to an extension of the amino terminal sequence of HA1 and the other (four residues) occurring in the region processed away between HA1 and HA2. Comparison of the HA1 of two H3 strains suggests that drift probably depends on single base mutations, some of which change antigenic determinants. The HA2 region, which apparently is not involved in the immune response, is highly conserved even between different subtypes, and single base substitutions account for all the observed diversity. A hydrophobic segment of 24 residues is present in the same position close to the carboxyl terminus of HA2 in both Victoria and FPV, and presumably functions in implantation into the lipid bilayer. The many conserved features not only in HA2 but also in HA1 suggest a rather rigid architecture for the whole hemagglutinin molecule.

Introduction

Influenza A virus infections of humans continue to cause worldwide epidemics because of the frequent and extensive antigenic variation of the virus. These changes may be drastic (antigenic shift) or relatively minor (antigenic drift). Whereas shifts only occur approximately every 10 years, drifts occur every 1-2 years. The influenza A virus has a divided genome consisting of eight negative-stranded RNA segments (Palese and Schulman, 1976b; McGeoch, Fellner and Newton, 1976; Scholtissek et al., 1976). The surface antigens, hemagglutinin and neuraminidase, are glycoproteins coded for by genes 4 and 5 or 6 (depending on the strain), respectively (Palese and Schulman, 1976b; Scholtissek et al., 1976). Although both surface antigens change independently, the hemagglutinin (HA) is the antigen against which neutralizing antibodies are directed (Laver and Kilbourne, 1966; Drzenick, Seto and Rott, 1966). The hemagglutinin is responsible for attachment to virus receptors (Hirst, 1942) and is involved in the initial stages of infection (Lazarowtiz and Choppin, 1975; Klenk et al., 1975). Thus the hemagglutinin is considered to be the most important antigen in determining the unique epidemiology of influenza.

To try to understand the antigenic variation of the hemagglutinin, two groups (Ward and Dopheide, 1979; Waterfield et al., 1979) have determined partial peptide sequences of the HA of a Hong Kong strain (H3N2) and an Asian strain (H2N2), respectively. Protein sequencing is relatively time consuming, however, and certain regions (of hydrophobic character) even pose difficult technical problems. The approach we have used was to determine the structure of a synthetic DNA copy of the HA gene cloned in a bacterial plasmid. We have previously worked out a method to convert a nonpolyadenylated viral RNA into an approximately full-size double-stranded (ds) DNA copy suitable for cloning (Devos et al., 1979b). Now we have applied this technique to the negative-stranded (and nonpolyadenylated) viral RNA corresponding to the hemagglutinin gene of a recent human influenza H3N2 strain, A/Victoria/3/75. The complete sequence is compared with the sequence of an animal strain (Fowl Plague Virus) determined by a similar procedure (Porter et al., 1979) and with the available amino acid sequence data of two human strains mentioned above: an H3N2 strain (Ward and Dopheide, 1979) and an H2N2 strain (Waterfield et al., 1979).

Results and Discussion

Characterization of the Hemagglutinin Gene Inserts

Plasmid DNA was purified on a small scale (Kahn et al., 1980) from the clones which hybridized specifically with the hemagglutinin RNA (see Experimental Procedures). Preliminary assessment of the length of the inserts by double digestion with Bgl I and Hind III and with Hinc II and Pvu II revealed insert lengths (including the tails) of approximately 720 bp (clone pVHA55; pVHA stands for plasmid-Victoria-Hemagglutinin) and 600 bp (pVHA57 and pVHA101). The other inserts in the plasmids from the first experiment were less than 350 bp long. Further characterization using the enzymes Eco RI, Hinf I and Ava I showed that pVHA55 was an extension of pVHA57 but that pVHA55 and pVHA101 were apparently not overlapping. Evidently this cloning experiment had not yielded a full-size gene 4 copy. Nevertheless clones 55 and 101 were retained for further characterization (see below) and for sequencing.

In the second cloning experiment, five plasmids showed positive hybridization with gene 4 RNA and all had an insert of at least 1700 bp. By further restriction mapping, four of the inserts were found to be nearly identical. The fifth clone, pVHA18, contained the same piece of DNA (the "near-complete" hemagglutinin information) but also an extra insert of about 760 bp. The latter was tentatively identified as the translocatable element IS1 on the basis of both size and the presence of one Pst I and three Hinf I restriction sites at appropriate positions relative to one another (Ohtsubo and Ohtsubo, 1978; Johnsrud, 1979). Two similar observations have been made previously in our laboratory using the same protocol for cloning (insertion with poly(dA)·poly(dT) tails into the Pst I site of pBR322): an IS1 sequence was found in MS2 clones (Devos et al., 1979a, 1979b) and also in satellite tobacco necrosis virus clones (J. van Emmelo et al., manuscript in preparation).

The five clones from experiment II contained all the information present in pVHA55 and in pVHA101; one of the clones without the IS1, pVHA14, was arbitrarily chosen for further study. In the experiment involving poly(dG)·poly(dC) tailing, one presumed full-length clone, pVHA10, was selected for further analysis. Restriction maps were made of pVHA55, pVHA101 and later of pHVA14 and pVHA10, and, as nucleotide sequence information became available, additional restriction sites were localized by a computer search using the summary list of restriction enzymes of Roberts (1978). The complete restriction map is shown in Figure 1.

DNA Sequencing

5′ and (occasionally) 3′ terminally labeled fragments were prepared and sequenced according to the

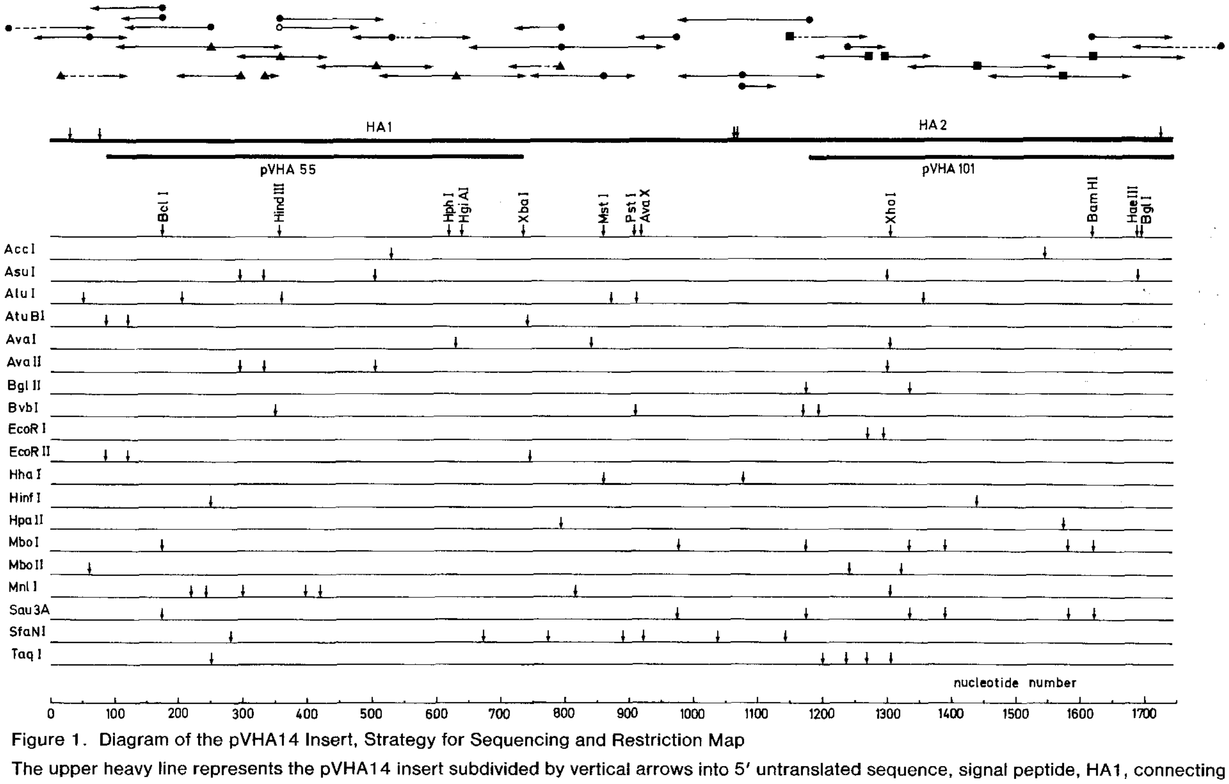

Figure 1. Diagram of the pVHA14 Insert, Strategy for Sequencing and Restriction Map

The upper heavy line represents the pVHA14 insert subdivided by vertical arrows into 5′ untranslated sequence, signal peptide, HA1, connecting region (one amino acid), HA2 and 3′ untranslated sequence. The region spanned by the clones pVHA55 and pVHA101 is represented underneath. Horizontal arrows in the top part indicate schematically the procedures used for sequencing: (●→) DNA from pVHA14, 5′ end-labeled; (○→) DNA from pVHA14, 3′ end-labeled; (▲→) DNA from pVHA55, 5′ end-labeled; (■→) DNA from pVHA101, 5′ end-labeled. Dashed segments refer to regions which were allowed to run off the gel; many more experiments than shown here have been carried out. Below is a restriction endonuclease cleavage map constructed by computer search on the basis of the complete nucleotide sequence using the list of enzymes compiled by Roberts (1978); enzymes which cut only once in the insert are indicated on the top line and enzymes cutting at multiple sites are indicated on the lower lines.

Table 1. Immunological Characterization of the Hemagglutinin Gene by the Hemagglutination Inhibition Test

	Post-infection Ferret Sera		
Virus Strain	A/Hong Kong/ 1/68	A/England/ 864/75	A/Victoria/ 3/75
A/Hong Kong/ 1/68	5120	<40	40
A/England/ 864/75	<40	5120	320
A/Victoria/ 3/75	<40	60	1280
X47	<40	5120	2560

Titers obtained in the homologous system are underlined. This experiment was carried out by Dr. J. Skehel.

method of Maxam and Gilbert (1977). Nearly all of the sequence was read several times, either on the same strand but from different restriction sites or on the opposite strand; approximately 70% of the sequence was determined on both strands (Figure 1). Thus ambiguities in one strand (such as an occasional difficult discrimination between T and C) could be easily resolved by reading the other strand. In a few cases a region close to the end (within approximately 15 nucleotides) of a restriction fragment did not yield a simple pattern (Hind III site in both directions, Ava I site to the left). However, either reading the same strand from a more distant site or reading the opposite strand provided the correct sequence. To rule out the possibility that a small restriction fragment between two identical restriction sites would have been overlooked, we read through all the sites used for sequencing (Figure 1). By this procedure, for example, we detected a region of 22 nucleotides bounded by two Eco RI sites.

Gaps in the ladder suggested the presence of methylcytosine at these positions in Eco RII sites (Ohmori, Tomizawa and Maxam, 1978). The nucleotide sequence was unambiguously established by information from both strands. Furthermore, the position of three Eco RII sites was verified by restriction mapping using the enzyme Bst NI, an isoschizomer of Eco RII that cleaves in spite of cytosine methylation.

Characterization of the Influenza Virus Strain

The influenza strain A/Victoria/3/75 (an H3N2 strain) is a natural field isolate. To obtain an attenuated high yielder strain for vaccination, it was subsequently crossed with A/PR/8/34 (an H0N1 strain), resulting in the recombinant X47. The latter has retained the P1 polymerase, the hemagglutinin, the neuraminidase and the nucleoprotein genes from A/Victoria/3/75 (P. Palese, personal communication). This X47 strain propagated under conditions as for vaccine preparation was the source of the A/Victoria/3/75 hemagglutinin gene in the present study. The identity of this gene was again checked by hemagglutination-inhibition assay (experiments carried out for us by J. Skehel). As shown in Table 1, there is indeed an immunologically detectable difference with the original field isolate. Possibly a mutant hemagglutinin gene has been fortuitously cloned during the construction of X47, or else a rare mutation in the population was selectively propagated under the conditions used to maintain and grow the X47 stock. An example of two allelic mutants in a recombinant stock and selective propagation of one type has been described by Kilbourne (1978). The exact difference between the hemagglutinin gene in the natural A/Victoria/3/75 and in X47 remains to be further characterized; for example, by probing with monoclonal antibodies.

General Features of the Nucleotide Sequence

The total length of the cloned hemagglutinin gene is 1745 bp; the sequence is shown in Figure 2. The presence of the sequence 5′-AGCAAAAGCAGG at the 5′ end of the pVHA14 insert proves that the complete information starting from the 3′ end of the viral RNA (negative strand) has been cloned, as the complementary sequence has been found at the 3′ end of all influenza A viral RNAs studied thus far (Skehel and Hay, 1978; Air, 1979; Robertson, 1979; Both and Air, 1979; Desselberger et al., 1980; McCauley et al., 1979; Porter et al., 1979). The only nucleotide which is variable in the first segment of 12 residues is nucleotide 4, which can be A (FPV; X31; HA, NP and NS from PR8; HA from A/RI/5/57) or G (the five other genes of PR8). Beyond position 12 the sequence of the different genes diverges. The conserved stretch (5′ end of the coding strand of the insert; that is, 3′ end of the virion RNA) is preceded in the clone by (dT)$_8$, originating from the primer used for reverse transcription [in fact a (dT)$_{10}$ primer had been used] and by a poly(A) tail approximately 70 nucleotides long (not shown). Although the complete coding information of the hemagglutinin gene is present in clones pVHA14 and pVHA10, some genetic information is missing at the 3′ end of the insert. Overlapping information determined by sequencing the 5′ end of the viral RNA by reverse transcription using a labeled restriction fragment as primer revealed that 23 nucleotides are missing from the pVHA14 insert and 22 from pVHA10. The first 22 nucleotides at the 5′ end of the virion RNA—except for residues 14, 15 and 16—are highly conserved in all genes of all A-type viruses examined thus far (Skehel and Hay, 1978; Robertson, 1979; Desselberger et al., 1980). The addition of 23 residues makes the exact length of the viral RNA 1768 nucleotides, which is very similar to the length of the FPV hemagglutinin RNA—that is, 1742 nucleotides (Porter et al., 1979). The loss of 22–23 nucleotides is similar to what has been found in comparable genetic engineering experiments (Efstratiadis, Kafatos and Maniatis, 1977; Devos et al.,

```
                           AGCAAAAGC AGGGGATAAT TCTATTAACC ATG AAG.ACT.ATC.ATT.GCT.TTG.AGC.TAC.ATT.TTC.TGT.CTG.GTT.TTC.GCC.CAA.GAC.CTT.CCA.GGA.AAT.GAC.
                           TCGTTTTCG TCCCCTATTA AGATAATTGG TAC TTC TGA TAG TAA CGA AAC TCG ATG TAA AAG ACA GAC CAA AAG CGG GTT CTG GAA GGT CCT TTA CTG
A/Victoria/3/75 (X47)                                     Met-Lys-Thr-Ile-Ile-Ala-Leu-Ser-Tyr-Ile-Phe-Cys-Leu-Val-Phe-Ala-Gln-Asp-Leu-Pro-Gly-Asn-Asp-
A/Memphis/102/72
A/Jap/305/57                                              Met-Ala-Ile-Ile-Tyr-Leu-Ile-Leu-Leu-Phe-Thr-Ala-Val-Arg-Gly ——————————————————
A/FPV/Rostock/34                                          Met-Asn-Thr-Gln-Ile-Leu-Val-Phe-Ala-Leu-Val-Ala-Val-Ile-Pro-Thr-Asn-Ala ——————————

99
AAC.AAC.AGC.ACA.GCA.ACG.CTG.TGC.CTG.GGA.CAT.CAT.GCG.GTG.CCA.AAC.GGA.ACG.CTA.GTG.AAA.ACA.ATC.ACG.AAT.GAT.CAG.ATT.GAA.GTG.ACT.AAT.GCT.ACT.GAG.
TTG TTG TCG TGT CGT TGC GAC ACG GAC CCT GTA GTA CGC CAC GGT TTG CCT TGC GAT CAC TTT TGT TAG TGC TTA CTA GTC TAA CTT CAC TGA TTA CGA TGA CTC
Asn-Asn-Ser-Thr-Ala-Thr-Leu-Cys-Leu-Gly-His-His-Ala-Val-Pro-Asn-Gly-Thr-Leu-Val-Lys-Thr-Ile-Thr-Asn-Asp-Gln-Ile-Glu-Val-Thr-Asn-Ala-Thr-Glu 42
——————————— Asp-Gln-Ile-Cys-Ile-Gly-Tyr-His-Ala-Asn-Asn-Ser-Thr-Glu-Lys-Val-Asp-Thr-Ile-Leu-Glu-Arg-Asn-Val-Thr-Val-Thr-His-Gly-Arg
——————————— Asp-Lys-Ile-Cys-Leu-Gly-His-His-Ala-Val-Ser-Asn-Gly-Thr-Lys-Val-Asn-Thr-Leu-Thr-Glu-Arg-Gly-Val-Glu-Val-Val-Asn-Ala-Thr-Glu

204
CTG.GTT.CAG.AGT.TCC.TCA.ACG.GGT.AAA.ATA.TGC.AAC.AAT.CCT.CAT.CGA.ATC.CTT.GAT.GGA.ATA.AAC.TGC.ACA.CTG.ATA.GAT.GCT.CTA.TTG.GGG.GAC.CCT.CAT.TGT.
GAC CAA GTC TCA AGG AGT TGC CCA TTT TAT ACG TTG TTA GGA GTA GCT TAG GAA CTA CCT TAT TTG ACG TGT GAC TAT CTA CGA GAT AAC CCC CTG GGA GTA ACA
Leu-Val-Gln-Ser-Ser-Ser-Thr-Gly-Lys-Ile-Cys-Asn-Asn-Pro-His-Arg-Ile-Leu-Asp-Gly-Ile-Asn-Cys-Thr-Leu-Ile-Asp-Ala-Leu-Leu-Gly-Asp-Pro-His-Cys 77
                                    Ile-Cys-Asn-Asn-Pro-His-Arg

Thr-Val-Glu-Arg-Thr-Asn-Ile-Pro-Lys-Ile-Cys-Ser-Lys-Gly-Lys-Arg-Thr-Thr-Asp-Leu-Gly-Gln-Cys-Gly-Leu-Leu-Gly-Thr-Ile-Thr-Gly-Pro-Pro-Gln-Cys-

309
GAT.GGA.TTT.CAA.AAT.GAG.AAA.TGG.GAC.CTT.TTC.GTT.GAA.CGC.AGC.AAA.GCT.TTC.AGC.AAC.TGT.TAC.CCT.TAT.GAT.GTG.CCA.GAT.TAT.GCC.TCC.CTT.AGG.TCA.CTA.
CTA CCT AAA GTT TTA CTC TTT ACC CTG GAA AAG CAA CTT GCG TCG TTT CGA AAG TCG TTG ACA ATG GGA ATA CTA CAC GGT CTA ATA CGG AGG GAA TCC AGT GAT
Asp-Gly-Phe-Gln-Asn-Glu-Lys-Trp-Asp-Leu-Phe-Val-Glu-Arg-Ser-Lys-Ala-Phe-Ser-Asn-Cys-Tyr-Pro-Tyr-Asp-Val-Pro-Asp-Tyr-Ala-Ser-Leu-Arg-Ser-Leu- 112

Asp-Gln-Phe-Leu-Glu-Phe-Ser-Ala-Asp-Leu-Ile-Ile-Glu-Arg-Arg-Glu-Gly-Asn-Asp-Val-Cys-Tyr-Pro-Gly-Lys-Phe-Val-Asn-Glu-Glu-Ala-Leu-Arg-Gln-Ile-

414
GTT.GCC.TCG.TCA.GGC.ACT.CTG.GAG.TTT.ATC.AAT.G.A.GGC.TTC.AAT.TGG.ACT.GGG.GTC.ACT.CAG.AAT.GGG.GGA.AGC.AGT.GCT.TGC.AAA.AGA.GGA.CCT.GAT.AGC.GGT.
CAA CGG AGC AGT CCG TGA GAC CTC AAA TAG TTA CIT CCG AAG TTA ACC TGA CCC CAG TGA GTC TTA CCC CCT TCG TCA CGA ACG TTT TCT CCT GGA CTA TCG CCA
Val-Ala-Ser-Ser-Gly-Thr-Leu-Glu-Phe-Ile-Asn-Glu-Gly-Phe-Asn-Trp-Thr-Gly-Val-Thr-Gln-Asn-Gly-Gly-Ser-Ser-Ala-Cys-Lys-Arg-Gly-Pro-Asp-Ser-Gly- 147
                                                                                                        Gly-Pro-Asp-Ser-Gly-

Leu-Arg-Gly-Ser-Gly-Gly-Ile-Asp-Lys-Glu-Thr-Met-Gly-Phe-Thr-Tyr-Ser-Gly-Ile-Arg-Thr-Asn-Gly-Thr-Thr-Ser-Ala-Cys-Arg-Arg-Ser-Gly-Ser-Ser-Phe-

519
TTT.TTC.AGT.AGA.CTG.AAC.TGG.TTG.TAC.AAA.TCA.GGA.AGC.ACA.TAT.CCA.GTG.CAA.AAC.GTG.ACC. dl ATG.CCA.AAC.AAT.GAC.AAT.TCT.GAC.AAA.CTA.TAC.ATT.TGG.
AAA AAG TCA TCT GAC TTG ACC AAC ATG TTT AGT CCT TCG TGT ATA GGT CAC GTT TTG CAC TGG    TAC GGT TTG TTA CTG TTA AGA CTG TTT GAT ATG TAA ACC
Phe-Phe-Ser-Arg-Leu-Asn-Trp-Leu-Tyr-Lys-Ser-Gly-Ser-Thr-Tyr-Pro-Val-Gln-Asn-Val-Thr ——  Met-Pro-Asn-Asn-Asp-Asn-Ser-Asp-Lys-Leu-Tyr-Ile-Trp- 181
Phe-Phe-Ser-Arg                                                                          Met-Pro-Asn-Asn-Asp-Asn-Phe-Asp-Lys-Leu-Tyr-Ile-Trp-
                                                                                   Gly-Ser-Tyr-Thr-Asn-Asn-Glx-Ser-Gly-Met-Leu-Ile-Ile-Trp-
Tyr-Ala-Glu-Met-Glu-Trp-Leu-Leu-Ser-Asn-Thr-Asp-Asn-Ala-Ser-Phe-Pro-Gln-Met-Thr-Lys-Ser-Tyr-Lys-Asn-Thr-Arg-Arg-Glu-Ser-Ala-Leu-Ile-Val-Trp-

621
GGG.GTT.CAC.CAC.CCG.AGC.ACG.GAC.AAA.GAA.CAA.ACC.AAC.CTA.TAT.GTT.CAA.GCA.TCA.GGG.AAA.GTC.ACA.GTC.TCC.ACC.AAG.AGA.AGC.CAG.CAA.ACT.ATA.ATC.CCG.
CCC GAA GTG GTG GGC TCG TGC CTG TTT CTT GTT TGG TTG GAT ATA CAA GTT CGT AGT CCC TTT CAG TGT CAG AGG TGG TTC TCT TCG GTC GTT TGA TAT TAG GGC
Gly-Val-His-His-Pro-Ser-Thr-Asp-Lys-Glu-Gln-Thr-Asn-Leu-Tyr-Val-Gln-Ala-Ser-Gly-Lys-Val-Thr-Val-Ser-Thr-Lys-Arg-Ser-Gln-Gln-Thr-Ile-Ile-Pro- 216
Gly-Val-His-His-Pro-Ser-Thr-Asp-Gln-Glu-Gln-Thr-Ser-Leu-Tyr-Val-Gln-Ala-Ser-Gly-Arg-Val-Thr-Val-Ser-Thr-Lys-Arg-Ser-Gln-Gln-Thr-Ile-Ile-Pro-
Gly-Val-His-His-Pro-Ile-Asp-Glu-Thr-Glu-Gln-Arg
Gly-Ile-His-His-Ser-Gly-Ser-Thr-Thr-Glu-Gln-Thr-Lys-Leu-Tyr-Gly-Ser-Gly-Asn-Lys-Leu-Ile-Thr-Val-Gly-Ser-Ser-Lys-Tyr-His-Gln-Ser-Phe-Val-Pro-

726
AAT.GTC.GGG.TCT.AGA.CCC.TGG.GTA.AGG.GGT.CTG.TCT.AGT.AGA.ATA.AGC.ATC.TAT.TGG.ACA.ATA.GTT.AAA.CCG.GGA.GAC.ATA.CTG.GTA.ATT.AAT.AGT.AAT.GGG.AAC.
TTA CAG CCC AGA TCT GGG ACC CAT TCC CCA GAC AGA TCA TCT TAT TCG TAG ATA ACC TGT TAT CAA TTT GGC CCT CTG TAT GAC CAT TAA TTA TCA TTA CCC TTG
Asn-Val-Gly-Ser-Arg-Pro-Trp-Val-Arg-Gly-Leu-Ser-Ser-Arg-Ile-Ser-Ile-Tyr-Trp-Thr-Ile-Val-Lys-Pro-Gly-Asp-Ile-Leu-Val-Ile-Asn-Ser-Asn-Gly-Asn- 251
Asn-Ile-Gly-Ser-Arg-Pro-Trp-Val-Arg-Gly-Leu-Ser-Ser-Arg-Ile-Ser-Ile-Tyr-Trp-Thr-Ile-Val-Lys-Pro-Gly-Asp-Ile-Leu-Val-Ile-Asn-Ser-Asx-Gly-Asx-
                                                        Met-Gln-Phe-Ser-Trp-Thr-Leu-Leu-Asp-Met-Trp-Asp-Thr-Ile-Asn-Phe-Glu-Ser-Thr-Gly-Asn-
Ser-Pro-Gly-Thr-Arg-Pro-Gln-Ile-Asn-Gly-Gln-Ser-Gly-Arg-Ile-Asp-Phe-His-Trp-Leu-Ile-Leu-Asp-Pro-Asn-Asp-Thr-Val-Thr-Phe-Ser-Phe-Asn-Gly-Ala-

831
CTA.ATT.GCT.CCT.CGG.GGT.TAC.TTC.AAA.ATG.CGC.ACT.GGG.AAA.AGC.TCA dl ATA.ATG.AGG.TCA.GAT.GCA.CCT.ATT.GGC.ACC.TGC.AGC.TCT.GAA.TGC.ATC.ACT.CCA.
GAT TAA CGA GGA GCC CCA ATG AAG TTT TAC GCG TGA CCC TTT TCG AGT    TAT TAC TCC AGT CTA CGT GGA TAA CCG TGG ACG TCG AGA CTT ACG TAG TGA GGT
Leu-Ile-Ala-Pro-Arg-Gly-Tyr-Phe-Lys-Met-Arg-Thr-Gly-Lys-Ser-Ser ——— Ile-Met-Arg-Ser-Asp-Ala-Pro-Ile-Gly-Thr-Cys-Ser-Ser-Glu-Cys-Ile-Thr-Pro- 286
Leu-Ile-Ala-Pro-Arg-Gly-Tyr-Phe-Lys-Met-Arg-Thr-Gly-Lys-Ser-Ser ——— Ile-Met-Arg-Ser-Asp-Ala-Pro-Ile-Gly-Thr-Cys-Ile-Ser-Glu-Cys-Ile-Thr-Pro-
Leu-Ile-Ala-Pro-Glu-Tyr-Gly-Phe-Lys-Ile-Ser-Lys-Arg-Gly-Ser-Ser-Gly-Ile-Met-Lys-Thr-Glu-Gly-Thr-Leu-Glu-Asn-Cys-Glu-Thr-Lys-Cys-Gln-Thr-Pro-
Phe-Ile-Ala-Pro-Asn-Arg-Ala-Ser-Phe-Leu-Arg ——— Gly-Lys-Ser-Met-Gly-Ile-Gln-Ser-Asp-Val-Gln-Val-Asp-Ala-Asn-Cys-Glu-Gly-Glu-Cys-Tyr-His-Ser-

933
AAT.GGA.AGC.ATT.CCC.AAT.GAC.AAG.CCC.TTT.CAA.AAC.GTA.AAC.AAG.ATC.ACA.TAT.GGG.GCA.TGT.CCC.AAG.TAT.GTT.AAG.CAA.AAC.ACT.CTG.AAG.TTG.GCA.ACA.GGG.
TTA CCT TCG TAA GGG TTA CTG TTC GGG AAA GTT TTG CAT TTG TTC TAG TGT ATA CCC CGT ACA GGG TTC ATA CAA TTC GTT TTG TGA GAC TTC AAC CGT TGT CCC
Asn-Gly-Ser-Ile-Pro-Asn-Asp-Lys-Pro-Phe-Gln-Asn-Val-Asn-Lys-Ile-Thr-Tyr-Gly-Ala-Cys-Pro-Lys-Tyr-Val-Lys-Gln-Asn-Thr-Leu-Lys-Leu-Ala-Thr-Gly- 320
Asn-Gly-Ser-Ile-Pro-Lys-Pro-Asp-Asp-Phe-Gln-Asn-Val-Asn-Lys-Ile-Thr-Tyr-Gly-Ala-Cys-Pro-Lys-Tyr-Val-Lys-Gln-Asn-Thr-Leu-Lys-Leu-Ala-Thr-Gly-
Leu-Gly-Ala-Ile-Asn-Thr-Thr-Leu-Pro-Phe-His-Asn-Val-His-Pro-Leu-Thr-Ile-Gly-Glu-Cys-Pro-Lys-Tyr-Val-Lys-Ser-Glu-Lys-Leu-Val-Leu-Ala-Thr-Gly-
Gly-Gly-Thr-Ile-Thr-Ser-Arg-Leu-Pro-Phe-Gln-Asn-Ile-Asn-Ser-Arg-Ala-Val-Gly-Lys-Cys-Pro-Arg-Tyr-Val-Lys-Gln-Glu-Ser-Leu-Leu-Leu-Ala-Thr-Gly-

1038
ATG.CGG.AAT.GTA.CCA.GAG.AAA.CAA.ACT.AGA.    dl    GGC.ATA.TTC.GGC.GCA.ATA.GCA.GGT.TTC.ATA.GAA.AAT.GGT.TGG.GAG.GGA.ATG.ATA.GAC.GGT.TGG.
TAC GCC TTA CAT GGT CTC TTT GTT TGA TCT           CCG TAT AAG CCG CGT TAT CGT CCA AAG TAT CTT TTA CCA ACC CTC CCT TAC TAT CTG CCA ACC
Met-Arg-Asn-Val-Pro-Glu-Lys-Gln-Thr-Arg ——————————— Gly-Ile-Phe-Gly-Ala-Ile-Ala-Gly-Phe-Ile-Glu-Asn-Gly-Trp-Glu-Gly-Met-Ile-Asp-Gly-Trp- 21
Met-Arg-Asn-Val-Pro-Glu-Lys-Gln-Thr                 Gly-Leu-Phe-Gly-Ala-Ile-Ala-Gly-Phe-Ile-Glu-Asn-Gly-Trp-Glu-Gly-Met-Ile-Asp-Gly-Trp-
Leu-Arg-Asx-Val-Pro-Glx-Ser-Glx-Ile                 Gly-Leu-Phe-Gly-Ala-Ile-Ala-Gly-Phe-Ile-Glu-Gly-Gly-Trp-Glu-Gly-Met-Val-Asp-Gly-Trp-
Met-Lys-Asn-Val-Pro-Glu-Pro-Ser-Lys-Lys-Arg-Glu-Lys-Arg-Gly-Leu-Phe-Gly-Ala-Ile-Ala-Gly-Phe-Ile-Glu-Asn-Gly-Trp-Glu-Gly-Leu-Val-Asp-Gly-Trp-

1131
TAC.GGT.TTC.AGG.CAT.CAA.AAT.TCC.GAG.GGC.ACA.GGA.CAA.GCA.GCA.GAT.CTT.AAA.AGC.ACT.CAA.GCA.GCC.ATC.GAC.CAA.ATC.AAT.GGG.AAA.CTG.AAT.AGG.GTA.ATC.
ATG CCA AAG TCC GTA GTT TTA AGG CTC CCG TGT CCT GTT CGT CGT CTA GAA TTT TCG TGA GTT CGT CGG TAG CTG GTT TAG TTA CCC TTT GAC TTA TCC CAT TAG
Tyr-Gly-Phe-Arg-His-Gln-Asn-Ser-Glu-Gly-Thr-Gly-Gln-Ala-Ala-Asp-Leu-Lys-Ser-Thr-Gln-Ala-Ala-Ile-Asp-Gln-Ile-Asn-Gly-Lys-Leu-Asn-Arg-Val-Ile- 56
Tyr-Gly-Phe-Arg-His-Gln-Asn-Ser-Glu-Gly-Thr-Gly-Gln-Ala-Ala-Asp-Leu-Lys-Ser-Thr-Gln-Ala-Ala-Ile-Asp-Gln-Ile-Asp-Gly-Lys-Leu-Asn-Arg-Val-Ile-
Tyr-Gly-Tyr
Tyr-Gly-Phe-Arg-His-Gln-Asn-Ala-Gln-Gly-Glu-Gly-Thr-Ala-Ala-Asp-Tyr-Lys-Ser-Thr-Gln-Ser-Ala-Ile-Asp-Gln-Ile-Thr-Gly-Lys-Leu-Asn-Arg-Leu-Ile-
```

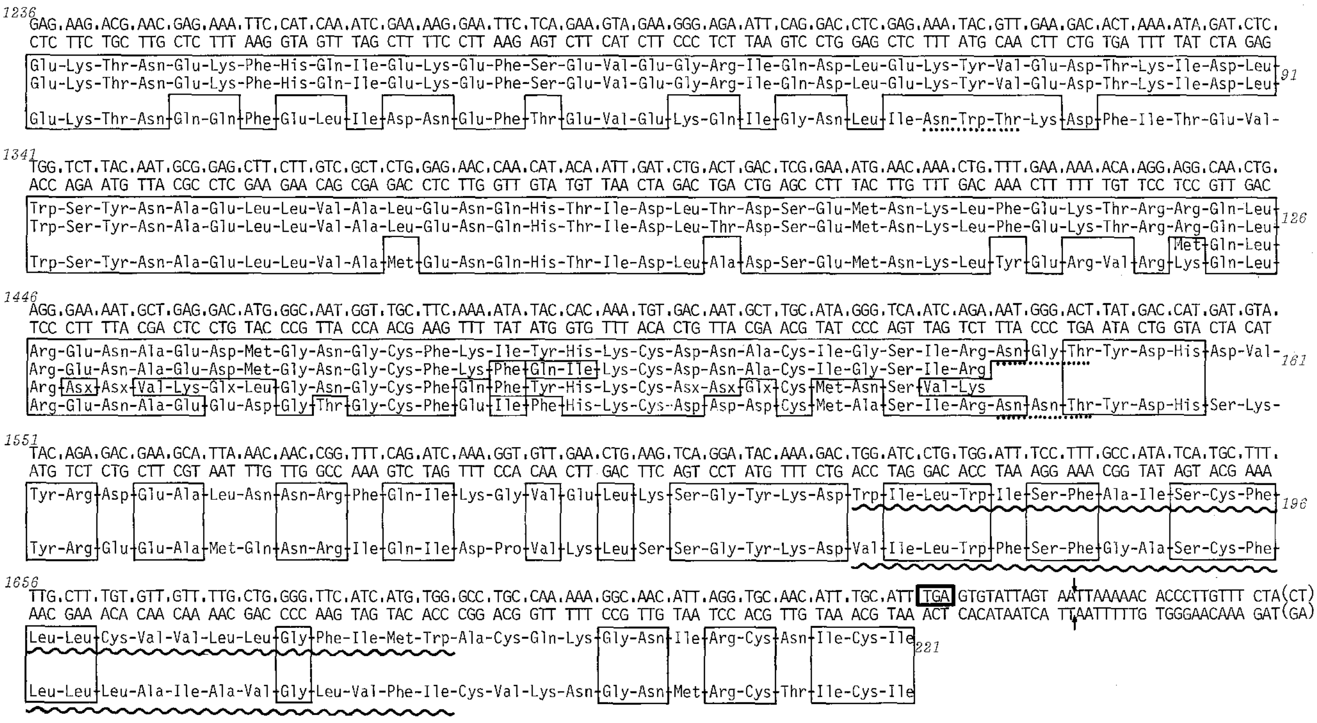
1236
GAG.AAG.ACG.AAC.GAG.AAA.TTC.CAT.CAA.ATC.GAA.AAG.GAA.TTC.TCA.GAA.GTA.GAA.GGG.AGA.ATT.CAG.GAC.CTC.GAG.AAA.TAC.GTT.GAA.GAC.ACT.AAA.ATA.GAT.CTC.
CTC TTC TGC TTG CTC TTT AAG GTA GTT TAG CTT TTC CTT AAG AGT CTT CAT CTT CCC TCT TAA GTC CTG GAG CTC TTT ATG CAA CTT CTG TGA TTT TAT CTA GAG
Glu-Lys-Thr-Asn-Glu-Lys-Phe-His-Gln-Ile-Glu-Lys-Glu-Phe-Ser-Glu-Val-Glu-Gly-Arg-Ile-Gln-Asp-Leu-Glu-Lys-Tyr-Val-Glu-Asp-Thr-Lys-Ile-Asp-Leu
Glu-Lys-Thr-Asn-Glu-Lys-Phe-His-Gln-Ile-Glu-Lys-Glu-Phe-Ser-Glu-Val-Glu-Gly-Arg-Ile-Gln-Asp-Leu-Glu-Lys-Tyr-Val-Glu-Asp-Thr-Lys-Ile-Asp-Leu
91
Glu-Lys-Thr-Asn-Gln-Gln-Phe-Glu-Leu-Ile-Asp-Asn-Glu-Phe-Thr-Glu-Val-Glu-Lys-Gln-Ile-Gly-Asn-Leu-Ile-Asn-Trp-Thr-Lys-Asp-Phe-Ile-Thr-Glu-Val-
1341
TGG.TCT.TAC.AAT.GCG.GAG.CTT.CTT.GTC.GCT.CTG.GAG.AAC.CAA.CAT.ACA.ATT.GAT.CTG.ACT.GAC.TCG.GAA.ATG.AAC.AAA.CTG.TTT.GAA.AAA.ACA.AGG.AGG.CAA.CTG.
ACC AGA ATG TTA CGC CTC GAA GAA CAG CGA GAC CTC TTG GTT GTA TGT TAA CTA GAC TGA CTG AGC CTT TAC TTG TTT GAC AAA CTT TTT TGT TCC TCC GTT GAC
Trp-Ser-Tyr-Asn-Ala-Glu-Leu-Leu-Val-Ala-Leu-Glu-Asn-Gln-His-Thr-Ile-Asp-Leu-Thr-Asp-Ser-Glu-Met-Asn-Lys-Leu-Phe-Glu-Lys-Thr-Arg-Arg-Gln-Leu
Trp-Ser-Tyr-Asn-Ala-Glu-Leu-Leu-Val-Ala-Leu-Glu-Asn-Gln-His-Thr-Ile-Asp-Leu-Thr-Asp-Ser-Glu-Met-Asn-Lys-Leu-Phe-Glu-Lys-Thr-Arg-Arg-Gln-Leu
126
Met-Gln-Leu
Trp-Ser-Tyr-Asn-Ala-Glu-Leu-Leu-Val-Ala-Met-Glu-Asn-Gln-His-Thr-Ile-Asp-Leu-Ala-Asp-Ser-Glu-Met-Asn-Lys-Leu-Tyr-Glu-Arg-Val-Arg-Lys-Gln-Leu
1446
AGG.GAA.AAT.GCT.GAG.GAC.ATG.GGC.AAT.GGT.TGC.TTC.AAA.ATA.TAC.CAC.AAA.TGT.GAC.AAT.GCT.TGC.ATA.GGG.TCA.ATC.AGA.AAT.GGG.ACT.TAT.GAC.CAT.GAT.GTA.
TCC CTT TTA CGA CTC CTG TAC CCG TTA CCA ACG AAG TTT TAT ATG GTG TTT ACA CTG TTA CGA ACG TAT CCC AGT TAG TCT TTA CCC TGA ATA CTG GTA CTA CAT
Arg-Glu-Asn-Ala-Glu-Asp-Met-Gly-Asn-Gly-Cys-Phe-Lys-Ile-Tyr-His-Lys-Cys-Asp-Asn-Ala-Cys-Ile-Gly-Ser-Ile-Arg-Asn-Gly-Thr-Tyr-Asp-His-Asp-Val-
Arg-Glu-Asn-Ala-Glu-Asp-Met-Gly-Asn-Gly-Cys-Phe-Lys-Phe-Gln-Ile-Lys-Cys-Asp-Asn-Ala-Cys-Ile-Gly-Ser-Ile-Arg
161
Arg-Asx-Asx-Val-Lys-Glx-Leu-Gly-Asn-Gly-Cys-Phe-Gln-Phe-Tyr-His-Lys-Cys-Asx-Asx-Glx-Cys-Met-Asn-Ser-Val-Lys
Arg-Glu-Asn-Ala-Glu-Glu-Asp-Gly-Thr-Gly-Cys-Phe-Glu-Ile-Phe-His-Lys-Cys-Asp-Asp-Asp-Cys-Met-Ala-Ser-Ile-Arg-Asn-Asn-Thr-Tyr-Asp-His-Ser-Lys-
1551
TAC.AGA.GAC.GAA.GCA.TTA.AAC.AAC.CGG.TTT.CAG.ATC.AAA.GGT.GTT.GAA.CTG.AAG.TCA.GGA.TAC.AAA.GAC.TGG.ATC.CTG.TGG.ATT.TCC.TTT.GCC.ATA.TCA.TGC.TTT.
ATG TCT CTG CTT CGT AAT TTG TTG GCC AAA GTC TAG TTT CCA CAA CTT GAC TTC AGT CCT ATG TTT CTG ACC TAG GAC ACC TAA AGG AAA CGG TAT AGT ACG AAA
Tyr-Arg-Asp-Glu-Ala-Leu-Asn-Asn-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-Cys-Phe
196
Tyr-Arg-Glu-Glu-Ala-Met-Gln-Asn-Arg-Ile-Gln-Ile-Asp-Pro-Val-Lys-Leu-Ser-Ser-Gly-Tyr-Lys-Asp-Val-Ile-Leu-Trp-Phe-Ser-Phe-Gly-Ala-Ser-Cys-Phe
1656
TTG.CTT.TGT.GTT.GTT.TTG.CTG.GGG.TTC.ATC.ATG.TGG.GCC.TGC.CAA.AAA.GGC.AAC.ATT.AGG.TGC.AAC.ATT.TGC.ATT TGA GTGTATTAGT AATTAAAAAC ACCCTTGTTT CTA(CT)
AAC GAA ACA CAA CAA AAC GAC CCC AAG TAG TAC ACC CGG ACG GTT TTT CCG TTG TAA TCC ACG TTG TAA ACG TAA ACT CACATAATCA TTAATTTTTG TGGGAACAAA GAT(GA)
Leu-Leu-Cys-Val-Val-Leu-Leu-Gly-Phe-Ile-Met-Trp-Ala-Cys-Gln-Lys-Gly-Asn-Ile-Arg-Cys-Asn-Ile-Cys-Ile
221
Leu-Leu-Leu-Ala-Ile-Ala-Val-Gly-Leu-Val-Phe-Ile-Cys-Val-Lys-Asn-Gly-Asn-Met-Arg-Cys-Thr-Ile-Cys-Ile

Figure 2. Nucleotide and Amino Acid Sequence of the A/Victoria/3/75 Hemagglutinin Gene

The nucleotide sequence is presented in the form of a double-stranded DNA copy, essentially as present in the plasmid pVHA14 (except for the 3′ terminal part beyond the arrow, which was sequenced directly by reverse transcription of the viral RNA). The initiation and termination codons are shown in a heavy box. The nucleotide numbering is shown in italics at the left side. The nucleotide sequence was translated into an amino acid sequence, which consists of a signal sequence (from the initiation site up to the first arrow), the HA1 chain of 329 amino acids, a small "connecting" region between HA1 and HA2 (indicated by arrows) and the HA2 chain of 221 amino acids. The amino acid numbering is shown in italics on the right side; separate numbering is used for HA1 and HA2 and the signal peptide is numbered backward. The results are compared with partial amino acid sequences from another H3 strain, A/Memphis/102/72 (Ward and Dopheide, 1979; Laver et al., 1979) and from an H2 strain (Air, 1979; McCauley et al., 1979; Waterfield et al., 1979; M. Waterfield, personal communication) and with the complete amino acid sequence deduced from another cloned influenza A hemagglutinin gene, that of Fowl Plague Virus (Porter et al., 1979). Homologous sequences are boxed and dotted lines indicate potential sites for carbohydrate attachment at asparagine residues in Asn-X-Ser or Asn-X-Thr sequences. The HA2 carboxyl-proximal hydrophobic portion presumed to be associated with the membrane is indicated by a wavy line. Deletions introduced in the nucleotide sequence for alignment with other strains are indicated by the boxed symbol dl; the amino acid sequence has a continuous heavy line at these positions. The arrow close to the 3′ terminus indicates the end of pVHA101, while pVHA14 is one nucleotide shorter, ending at position 1744.

1979b; Porter et al., 1979; J. van Emmelo, personal communication), where between 13 and 46 nucleotides were missing from the 3′ end of the insert.

The noncoding RNA segment in the A/Victoria hemagglutinin plus-strand is 29 nucleotides long, as the first initiating AUG codon occurs at position 30–32 and the next in-phase AUG is found almost 500 nucleotides downstream. The latter AUG would be incompatible with the required coding capacity (Ward and Dopheide, 1979; Waterfield et al., 1979). Both other reading frames are frequently blocked by stop codons. A noncoding region of 29 nucleotides is comparable to the 21 nucleotides of the FPV hemagglutinin gene (Robertson, 1979; Porter et al., 1979), whereas that of the hemagglutinin gene of two closely related H2 strains is 43 nucleotides long (Air, 1979; McCauley et al., 1979). It may be recalled that the plus strand information as shown in Figure 2 is not the same as the hemagglutinin mRNA, which is probably slightly different at both ends. In vitro transcription experiments have shown that a cap-containing oligonucleotide can be transferred from added globin mRNA to the viral messages, resulting in approximately 10–15 extra nucleotides (Plotch, Bouloy and Krug, 1979); in vivo the 5′ ends of the influenza mRNAs are likely to be donated by the host mRNA population (Krug, Broni and Bouloy, 1979). In addition, at the 3′ end the polyadenylated viral mRNAs do not contain the sequence complementary to the very 5′ termini of the virion RNAs (Skehel and Hay, 1978). The length of noncoding RNA at the 3′ end of the plus strand of the insert is 12 residues (not including the termination codon). In fact, by sequencing this region of the viral RNA by the reverse transcription approach, we know that this noncoding segment of the viral RNA is 35 nucleotides long, the last 21 of which are identical to those of FPV (Porter et al., 1979). The noncoding sequence in the FPV hemagglutinin gene is 29 nucleotides long. The Victoria hemagglutinin is terminated by a UGA nonsense codon followed six nucleotides further by UAG.UAA. This is somewhat reminiscent of the UAA.UAG termination signal of the coat protein gene in group I RNA phages (Min Jou and Fiers, 1976). Double stop signals separated by four codons have been described for chicken ovalbumin mRNA (McReynolds et al., 1978) and for an immunoglobulin mRNA (Hamlyn et al., 1978). The FPV hemagglutinin is terminated by a single UAA signal. Neither hemagglutinin gene contains the AAUAAA sequence found at a short distance preceding the poly(A) tail of all eucaryotic cellular and many viral messengers [and possibly implicated in the polyadenylation process (Proudfoot and Brownlee, 1976)]. The majority of picornaviruses (Porter et al., 1978) and one of the VSV mRNAs (McGeoch and Turnbull, 1978) also do not have this signal.

Porter et al. (1979) have noted that the 3′ terminal segment of the cloned insert contains several mutations compared to the bona fide virion RNA. This result can be explained on the basis of the terminal hairpin used for the self-primed synthesis of the second DNA strand, followed by a correction of base mismatches in the bacterial cell. Direct sequencing of the 5′ end of the virion RNA by a reverse transcription approach, however, revealed that the cloned Victoria gene does not contain mutations in this region.

As shown in Figure 1, clones pVHA55 and pVHA101 cover most of the HA1 and HA2 genetic information, respectively. These clones were sequenced first, and later the complete primary structure of pVHA14 was established. Not a single nucleotide substitution was noted. This confirms our previous conclusion that the procedure used to convert the viral RNA into DNA followed by cloning is itself not highly mutagenic (Devos et al., 1979b). In addition, the large-scale preparation of influenza X47 must have been fairly homogeneous.

The Protein Sequence

The hemagglutinin of the infectious virions consists of two disulphide-bonded polypeptide chains (Ward and Dopheide, 1979; Waterfield et al., 1979) referred to as HA1 and HA2. The hemagglutinin is synthesized as a precursor with an amino-terminal hydrophobic signal sequence, the order of biosynthesis being NH_2-signal-HA1-HA2-COOH (Elder et al., 1979). From the nucleotide sequence reported in Figure 2, one can deduce a continuous sequence of 567 amino acids as the primary translation product of the hemagglutinin gene.

It is now well established that the large majority of excretory proteins start with a hydrophobic signal peptide which is subsequently processed away (Blobel and Dobberstein, 1975; Blobel et al., 1979). The amino terminus of HA1 is blocked in H3 strains (Laver and Webster, 1977; Ward and Dopheide, 1979; Waterfield et al., 1979), and for this reason no data are yet available about amino-terminal sequences. It has been suggested, however, that the blocked amino terminus is a pyroglutamic acid residue (Ward and Dopheide, 1979). We therefore propose a signal peptide of 16 amino acids; indeed, this segment has the typical hydrophobic character of such a prepeptide. The HA1 chain would then start at residue 17 of the precursor polypeptide. Beyond this position, amino acids with polar side chains occur fairly frequently. The amino-terminal glutamine residue would then become blocked by a post-translational process. Several other proteins with an amino-terminal pyroglutamic acid are known, such as bacteriorhodopsin (Ovchinnikov et al., 1979) and a proline-rich phosphoprotein from human saliva (Wong, Hofmann and Bennick, 1979).

As both the carboxyl-terminal amino acid sequence of the HA1 chain and the amino-terminal amino acid sequence of the HA2 chain are known for another H3 strain, namely A/Memphis/102/72 (Ward and Dopheide, 1979), and in view of the fact that the conservation in amino acid sequence in these regions is almost 100%, the sequence of the Victoria strain and of the Memphis strain can be unambiguously aligned (Figure 2). The carboxyl terminus of the HA1 chain is -Val-Pro-Glu-Lys-Glu-Thr; then a single arginine residue is excised by a protease system of the host; the HA2 chain starts with Gly-Ile-Phe-Gly-Ala-Ile-Ala-. The excision of the arginine residue can be accomplished, for example, by a trypsin-like activity followed by a carboxypeptidase B activity. Processing of the precursor hemagglutinin polypeptide with subsequent formation of HA1 and HA2 is required for production of infectious virions, but it is known that this step can also be produced by trypsin (Klenk et al., 1975; Lazarowitz and Choppin, 1975; Klenk, Rott and Orlich, 1977; Bosch et al., 1979). One can easily envisage the possibility for cleavage by trypsin right after the arginine residue to yield "quasi-normal" HA1 and HA2 (and infectious virus) under conditions where the host cellular protease is absent or inactive. The HA1 chain starting at residue 17 of the precursor polypeptide and ending at the carboxyl-terminal threonine residue is 329 amino acids long.

The information discussed above also completely defines the HA2 chain, which consists of 221 amino acid residues, in reasonable agreement with earlier estimates (Ward and Dopheide, 1979; Waterfield et al., 1979). Other aspects of the amino acid sequence will be discussed from a comparative point of view.

Comparison with Other Strains

The hemagglutinin gene of a fowl strain (Fowl Plague Virus) has also been cloned and completely sequenced (Porter et al., 1979). Other data include the sequence of a 180 nucleotide segment from the 5′ end of the coding strand of the hemagglutinin gene of A/Japan/305/57 (H2N2) (McCauley et al., 1979), an identical sequence of A/RI/5/57 (another H2N2 strain) (Air, 1979), and rather extensive amino acid sequencing data from an H3 strain (A/Memphis/102/72) (Ward and Dopheide, 1979; Laver et al., 1979) and from an H2 strain (A/Japan/305/57) (Waterfield et al., 1979). This information is summarized in Figure 2.

The signal peptide shows almost no sequence conservation, but is similar in hydrophobic character and approximate length. The segment Leu-Val-Phe-Ala is present in Victoria and in FPV but at a different distance from the amino terminus or the processing site; also, this segment is not present in the H2 strain. As discussed above, the presumed signal peptide of the Victoria precursor is 16 residues long, that of the H2 strain is 15 residues long (Air, 1979; McCauley et al., 1979) and that of the FPV is 18 residues long (Porter et al., 1979).

The two primary translation products of Victoria and FPV can be aligned by introducing two insertions (of 9 amino acids and 1 amino acid, respectively) and three deletions (2 × 1 and 1 × 4 amino acids, respectively) into the Victoria sequence with respect to the PFV sequence. The longer insertion (9 residues) lies between the signal peptide and the body of the HA1 coding information (that is, leading to an extension of the amino-terminal sequence) and the longest deletion (4 residues) lies in the region processed away in the formation of HA1 and HA2. All other insertions and deletions, taking into account also the available data on the H2 strain, involve only an occasional deletion/insertion of one amino acid. This observation allows the conclusion that the overall architecture of the hemagglutinin molecule is not that flexible. Also, from the data available so far there is no reason to believe that the drift between different H3 strains involves anything more than base changes.

Of 320 corresponding positions in HA1, 116 amino acids (36.2%) have been conserved in FPV, with the conservation being higher towards both ends (amino acids 15–44 and 295–326 of Victoria). A similar situation holds for the Japanese H2 strain. The carboxyl-terminal 160 amino acids of the other H3 strain, A/Memphis/102/72, match perfectly with the Victoria information; in fact, the amino acid conservation is 93% in this region.

Whereas only one arginine residue has to be excised to give rise to HA1 and HA2 in the Victoria strain (based on comparison of the coding region spanning the carboxyl terminus of HA1 and the amino terminus of HA2 with the corresponding amino acid sequence of the Memphis strain), in FPV five residues are removed. Not only is this an extraordinary basic amino acid sequence, Lys$\overset{\downarrow}{-}$Lys-Arg-Glu-Lys-Arg$\overset{\downarrow}{-}$Gly, but the coding nucleotide sequence is also remarkable: 18 purines in a row, 14 being adenine residues (in the coding strand) (Porter et al., 1979). This purine-rich segment is less pronounced around the connecting region of the Victoria sequence.

The HA2 chain contains 221 amino acids both in Victoria and FPV, and no insertions or deletions have to be invoked in order to maximize the amino acid sequence homology, which amounts to 65.6% overall. This value is much higher than that in the HA1 part (36.2%). At the nucleotide level a similar percentage of homology (65.7%) is observed in the HA2 region. The small conserved part and the large variable part detected by RNA cross-hybridization between members of different subtypes of influenza A (Scholtissek, 1979) thus essentially correspond in molecular terms to the HA2 and the HA1 region, respectively. The homology is even higher within the first 60 residues (83.3%) and in the middle region of the HA2 molecule (amino acids 92–120) (89.7%). The conservation between the amino-terminal 153 amino acids of Victoria and the Memphis H3 strain is 96.7% (five differences). Amino-terminal sequence analysis of the HA2 polypeptides from different human type-A viruses, equine viruses and a single type-B influenza virus has revealed extensive sequence homology in the first 17 amino acids (Waterfield et al., 1979). Because of this homology, the relative hydrophobic character of this region (11 out of 24 residues in Victoria) and the fact that this site has to be released by proteolytic cleavage to yield infectious virus, it is tempting to speculate that this region would be involved in an interaction with the cellular membrane during the infection process. Homology (although more limited) with the amino-terminal part of the Sendai virus fusion protein (Gething, White and Waterfield, 1978), which also requires proteolytic processing (Homma and Ohuchi, 1973; Scheid and Choppin, 1974), further strengthens this hypothesis.

Another hydrophobic region can clearly be recognized at exactly the same position close to the carboxyl terminus of HA2 both in Victoria and FPV: 18 of the 24 residues from amino acids 185 to 208 are strongly hydrophobic in Victoria. An HA2 carboxyl-terminal portion has been shown to be associated with the membrane (Skehel and Waterfield, 1975; Goldman et al., 1979). Although there is a high degree of sequence conservation in the first part (10 out of 14 residues), only the hydrophobic character is maintained in the last 10 residues of that section, while the nucleotide sequence is almost totally different. Furthermore, the region contains two conserved serine residues (position 190 and 194), one of which (or both) could be the fatty acid attachment site(s) (Schmidt, Bracha and Schlesinger, 1979). A stretch of 24 amino acids which spans the lipid bilayer would be analogous to the peptide of 22 amino acids proposed as the intramembranous segment of human erythrocyte glycophorin A (Furthmayr et al., 1978) and to a stretch of 25 amino acids forming the portion of the HLA-A and HLA-B antigen heavy chain spanning the membrane (Springer and Strominger, 1976). The last 11 carboxyl-terminal amino acids of Victoria (and

FPV) hemagglutinin could, in analogy with both aforementioned membrane proteins (Tomita and Marchesi, 1975; Robb, Terhorst and Strominger, 1978), form an internal hydrophilic tail. In the other two proteins the hydrophilic tail is approximately 30 residues long. In at least one eveloped virus, Semliki Forest virus, there is evidence that the spike glycoprotein extends through the membrane and comes into close contact with the internal nucleoprotein (Garoff and Symons, 1974). So far there is no direct evidence for such an interaction in influenza (for example, with the matrix or possibly nucleoprotein).

All over the HA1 and HA2 chains cysteine residues are extremely well conserved. In all cases where comparative data are available, the nine positions in Victoria HA1 where cysteine is found are conserved in other strains, and no new cysteine residues appear (Figure 2). In Victoria the cysteine residues in HA1 occur at positions 15, 53, 65, 77, 98, 140, 278, 282 and 306. The total of nine cysteines is close to a previous estimate of ten (Waterfield et al., 1979). In HA2 there are eight cysteines, at positions 137, 144, 148, 195, 199, 210, 217 and 220. The cysteine at position 199 is not found in FPV, a second one occupies a slightly different position (209 instead of 210) and the others are conserved in FPV and in the sequence of other strains as far as information is available. This almost complete conservation of the cysteine residues, even in the otherwise more variable HA1 chain, indicates that the disulphide bonds in different hemagglutinins follow the same pattern, and suggests a similar general folding of the molecule in various strains.

Codon Usage

Codon usage (and amino acid frequency) in the HA1 (including the "connecting" arginine) and in the HA2 coding sequence is summarized in Table 2. The most striking difference in amino acid composition between both chains is the complete absence of proline in HA2, compared to 20 in HA1. Based on amino acid composition data of the subunits from different strains, it seems that a proline content of zero up to only two or three residues is a general rule for HA2 chains (Ward and Dopheide, 1979; Waterfield et al., 1979). The FPV HA2 contains a single proline residue (Porter et al., 1979) (Figure 2). Victoria hemagglutinin is rich in asparagine (45 residues) and aspartic acid (31 residues) and to a lesser extent in glutamine (23 residues) and glutamic acid (27 residues).

Vertebrate DNA is known to be very low in the dinucleotide C-G. C-G-containing codons (for serine, proline, threonine, alanine and arginine) are lower than expected (except for threonine), but the effect is not as drastic as that observed in SV40 DNA (Fiers et al., 1978). It has been argued in the past that C-G in vertebrate DNA is low because its methylated form is mutagenic (although the C-G sequences in SV40 DNA are not methylated). Because the messengers are deficient in C-G-containing codons, the tRNA population of the cell might have adapted to this situation, and the tRNAs recognizing C-G-containing codons might therefore be sparse or inefficient. Influenza would have adapted to this situation by also partially selecting away these C-G-containing codons. If this hypothesis were correct, then there would be no reason for selecting away bridging C-G dinucleotides; that is, those formed by a codon ending with C followed by a G-N-N codon. In fact, we observe only 18 C-G bridges, while 41 are expected. Also FPV hemagglutinin is low in C-G-containing codons (Porter et al., 1979), and also the number of N-N-C.G-N-N bridges is considerably less than expected (19 actual versus approximately 35 expected). It seems, therefore, that there are also restrictions (not necessarily absolute) on certain codon nearest neighbors—that is, that the third letter of synonymous codons may in some instances be influenced by the following codon. Such restrictions for a few particular codons have recently also been discussed for the SV40 coding information (Van Heuverswyn et al., 1980).

Most of the other synonymous codons are used rather indiscriminately, except for leucine, which is frequently coded for by C-U-G and very seldom by U-U-A or C-U-C (there are some other instances where the codon choice is apparently nonrandom; for example, for serine).

Site of Glycosylation

Since our amino acid sequence data are completely derived from the known DNA structure, carbohydrate attachment sites can only be inferred by comparison with actual glycoprotein chemical analysis. From the carbohydrate composition of different influenza strains and the sequence at two modified sites, it is believed that all substitutions are via N-glycosidic linkage to asparagine (Ward and Dopheide, 1979; Waterfield et al., 1979) in the sequence asparagine-X-serine or asparagine-X-threonine, the presence of this sequence being a necessary but not a sufficient condition for glycosylation (Neuberger et al., 1972). In the Memphis/72 (H3) strain, a glycosylated site has been described at position 286 (Ward and Dopheide, 1979) (Figure 3), and since the sequence in the corresponding region is completely conserved it is very probable that the Victoria strain also has the same modification. In the same region, but not exactly the same position (corresponding to residue 290 in Victoria), a similar site has been found for the Japanese/57 (H2) strain (Waterfield et al., 1979). From the data of Waterfield et al. (1979) a second site can be inferred at position 173. Further potential sites for glycosylation in Victoria are at positions 8, 9 (two overlapping sites; it must be regarded as improbable that both are modified), 23, 39, 64, 127 and 166 of HA1, and position 154 of HA2. Three of these se-

Table 2. Use of Codons in A/Victoria/3/75 Gene 4

	U	C	A	G	
U	Phe 4 + 4 = 8	Ser 4 + 1 = 5	Tyr 7 + 1 = 8	Cys 3 + 2 = 5	U
	5 + 7 = 12	3 + 2 = 5	4 + 6 = 10	6 + 6 = 12	C
	Leu 0 + 1 = 1	7 + 4 = 11	Ochre	Opal	A
	3 + 2 = 5	1 + 1 = 2	Amber	Trp 6 + 6 = 12	G
C	Leu 4 + 4 = 8	Pro 6 + 0 = 6	His 4 + 4 = 8	Arg 0 + 0 = 0	U
	0 + 2 = 2	4 + 0 = 4	2 + 1 = 3	2 + 0 = 2	C
	6 + 0 = 6	7 + 0 = 7	Gln 9 + 8 = 17	1 + 0 = 1	A
	9 + 8 = 17	3 + 0 = 3	4 + 2 = 6	2 + 1 = 3	G
A	Ile 6 + 6 = 12	Thr 10 + 4 = 14	Asn 16 + 9 = 25	Ser 5 + 0 = 5	U
	7 + 8 = 15	4 + 0 = 4	14 + 7 = 21	12 + 1 = 13	C
	8 + 8 = 16	8 + 3 = 11	Lys 13 + 12 = 25	Arg 6 + 3 = 9	A
	Met 4 + 4 = 8	5 + 1 = 6	6 + 3 = 9	3 + 6 = 9	G
G	Val 7 + 4 = 11	Ala 5 + 3 = 8	Asp 8 + 4 = 12	Gly 4 + 6 = 10	U
	4 + 1 = 5	2 + 3 = 5	9 + 10 = 19	3 + 5 = 8	C
	4 + 3 = 7	5 + 6 = 11	Glu 4 + 11 = 15	11 + 3 = 14	A
	6 + 0 = 6	1 + 1 = 2	4 + 8 = 12	10 + 5 = 15	G

The first number refers to the frequency with which the triplet is used in the HA1 chain (including the connecting arginine residue), the second number shows the frequency for the HA2 chain and the third number gives the total frequency.

quences are conserved in FPV, at positions 23 and 39 of HA1 and a single site at position 154 of HA2. The latter site could well correspond to the single glycosylated site in the Japanese strain (Waterfield et al., 1979), near the bromelain cleavage site in HA2.

Concluding Remarks: "Shift" and "Drift" of Influenza A Strains

Influenza is one of the few unconquered viral diseases of man. It is unique in its frequent and extensive antigenic variation, and in the present paper we have explored the molecular basis of this intriguing adaptability.

Since 1933, when the first influenza virus (H0N1) was isolated from man, a major antigenic variation of the virus, called a SHIFT, has occurred about every 10 years, causing a pandemic. Such a novel strain spreading rapidly through the immunologically naive world population is due to the appearance of a new subtype of hemagglutinin (H) on the virus, occasionally accompanied also by a new subtype of the neuraminidase (N). Such pandemics have started in 1947 (H1N1), in 1957 (Asian flu: H2N2) and in 1968 (Hong Kong: H3N2). In addition to these four "human" H subtypes, at least 12 others have been immunologically identified in viruses isolated from pigs, horses and different avian species (Laver and Webster, 1979). The molecular basis of a shift can be partly understood by comparison of the Victoria, the Fowl Plague and (as far as the data are available) the Japanese strain hemagglutinin (Figure 2).

One major conclusion that can be drawn from this comparison is that many features of the amino acid sequence are remarkably constant. The hemagglutinin molecule is a specific three-dimensional structure which fulfills various functions; its genetic variability is most remarkable but may nevertheless be limited. The molecule must make the proper interactions to form a trimer and must have the proper conformation to be cleaved into an HA1 and HA2 part. The amino-terminal and middle part of the latter is remarkably constant and in fact may play a role in the fusion-penetration of the virion into the cell membrane (Gething et al., 1978), a process which may be analogous to an enzymatic reaction. The function of the carboxyl-terminal part of HA2 is implantation in the lipid bilayer, and possibly making contact with internal proteins; this region is constant in length and in several special features, as discussed in a preceding section. The function of HA1 is to bind to sialyl group-containing receptors of the cell membrane. As far as is known, all virus-inactivating antibodies are also hemagglutination-inhibiting. This suggests that this "active site" should be the predominant domain where antigenic changes occur.

The data in Figure 2 show that a shift means not only an extensive set of amino acid changes, but also occasional deletion and/or insertion of one amino acid. There are only two loci where more extensive deletions/insertions occurred, namely a prolongation of the amino terminus and an additional segment between HA1 and HA2 as in FPV. Apparently the formation of deletions or insertions resulting in viable virus is an extremely rare event in nature. Rather, the different shifts in the recent viral history of man can be accounted for by a limited number of preexisting genetic subtypes of H. The total number of such genetic subtypes, that is, candidates for pandemics, may be as low as three (supposing that H0 and H1 are variants of the same subtype; Scholtissek et al., 1977), or as high as 16 or more if hemagglutinin genes of animal influenza strains can occasionally recombine in a human strain (compare discussion by Laver and Webster, 1979).

Once a new H subtype has spread through the human population, minor antigenic changes called DRIFT occur and the resulting variant strains cause epidemics every 1 or 2 years. As mentioned above, the domains responsible for antigenic variation are part of HA1 (Potter and Oxford, 1979), namely, at least predominantly, in the amino-terminal 170 amino acids segment (Jackson et al., 1979). Gerhard and Webster (1978) estimated that there may be as many as 15–60 individual immunogenic determinants, and each determinant consists of several amino acids. It is probable, however, that the change of one determinant by mutation of an amino acid is not sufficient to generate a new strain capable of overcoming the immunity of the population (Gerhard and Webster, 1979). Monoclonal antibodies are a powerful tool for probing the individual determinants, and, using such techniques, Laver et al. (1979) found that at least amino acid positions 55, 144 and 206 are part of such sites, but many more remain to be revealed.

By comparing the sequence of the Victoria hemagglutinin with the information available for the other H3 strain, A/Memphis/72, it seems probable that drift is nothing more than the accumulation of a series of single base changes. What is the molecular basis for the remarkable genetic plasticity of influenza, and more particularly of HA1? RNA-dependent RNA polymerases of RNA bacteriophages have a very high error frequency, perhaps because they lack a proofreading mechanism (Domingo et al., 1978; Fiers, 1979). Nevertheless, a unique primary structure of the viral RNA is quite well maintained in the population, and this means that most reversions to wild-type sequence must be selectively advantageous, not only the positions which change the amino acid sequence, but also the so-called silent mutations (not necessarily neutral). It is known that the terminal noncoding sequences are more stringently conserved than the coding regions (Min Jou and Fiers, 1976). In addition, however, third letters of synonymous codons may not be freely interchangeable, for reasons of optimal codon usage, including so far little understood effects of codon nearest-neighbor restrictions, of secondary and tertiary structure of the RNA, of protein-viral RNA interactions (for example, coat protein acts as a repressor; regions recognized by the replicase and by host factors) and occasionally even of double function in a second reading frame. But these restrictions on base changes are not strict; if RNA phage is grown in the presence of antiserum, resistant mutants readily appear (Van Assche et al., 1972).

Hence there is no reason to believe that there exists a special mechanism to generate base mutations in the influenza hemagglutinin gene or even that the influenza RNA polymerase(s) are especially error prone. The unique ability to generate new influenza strains can simply be explained by appropriate immunological selection pressures. The fact that other animal RNA viruses, like picorna viruses, do not show this remarkable genetic variability may be for several reasons, either at the biological level (mode of infection; interaction with the immune system of the host) or at the molecular biological level (a capsid protein molecule of a polyhedral virus has multiple interaction domains and therefore presumably less freely variable immunogenic determinants; secondary and tertiary structure of the viral RNA may be important, which is not the case in the myxoviruses, where an unfolded genome is wrapped up in a nucleoprotein coil; a nonsegmented genome may have additional regions involved in translational control). It is very probable that the combined approach of comparison of different hemagglutinin sequences of the same subtype, of three-dimensional structure analysis of the hemagglutinin trimer (Wiley and Skehel, 1977) and of probing the various strains with monoclonal antibodies will soon lead to a better understanding of the possibilities and limitations of genetic drift in influenza.

Experimental Procedures

Bacterial Strains, Plasmids and Enzymes

Escherichia coli K12 strain HB101 (hsm^-, hsr^-, $recA^-$, gal^-, pro^-, str^R) and HB101 carrying the plasmid pBR322 were made available by H. Boyer. Recombinant plasmid DNA from 5-10 ml cultures and from 2-1 cultures was isolated according to methods described by Kahn et al. (1980). Avian myeloblastosis virus RNA-dependent DNA polymerase was supplied by J. W. Beard (U.S. National Cancer Institute, Viral Oncology Program). The restriction enzymes used were from New England Biolabs (Beverly, Massachusetts), except for Pst I, which was from MRE (Porton, England). Nitrocellulose filters (HATF) were from Millipore (Malsheim, France). They were autoclaved for 10 min before use. Other enzymes and reagents were from the sources described by Devos et al. (1979b).

Influenza Virus RNA

A large scale preparation of the influenza virus X47 was obtained from Evans Medical Co. (Liverpool, England). X47 is a high yield recombinant between A/Victoria/3/75 and A/PR/8/34 (an H0N1 strain) suitable for human vaccination and was isolated by E. D. Kilbourne. The viral RNA was extracted according to the procedure of Palese and Schulman (1976a).

Gel Electrophoresis

Agarose gel electrophoresis was carried out in horizontal 1.5% slab gels containing 50 mM Tris-acetate, 20 mM sodium acetate, 2 mM EDTA (pH 7.8). DNA molecules were separated on 4% or 6% acrylamide gel slabs containing 50 mM Tris-borate, 1mM EDTA (pH 8.3) (Maniatis, Jeffrey and van de Sande, 1975). The mixture of influenza RNAs was fractionated on a 3% acrylamide gel in the same buffer system supplemented with 7 M urea and 0.1% sodium dodecylsulphate (urea was present only in the gel, not in the buffer reservoirs). Detection of both unlabeled and labeled material, photography, autoradiography and elution of DNA from gels were as described previously (Devos et al., 1979b).

Construction of Molecular Chimeras

For the cloning of a dsDNA copy of the influenza hemagglutinin gene, we essentially followed the methodology described previously for the Bacteriophage MS2 genetic information (Devos et al., 1979b). Considering the limited amount of starting material available, we did not attempt to purify the hemagglutinin gene before polyadenylation, nor did we try to separate the $poly(rA)^+$ material after polyadenylation. However, several control experiments were set up on an analytical scale to test the quality of the starting material: first, analysis of the mixture of influenza RNAs on a low percentage acrylamide gel containing urea revealed the normal pattern of eight RNAs as generally observed for influenza A strains (McGeoch et al., 1976; Palese and Schulman, 1976b; Scholtissek et al., 1976); second, to show a physical attachment of a poly(rA) tail to all eight genome segments, a ^{32}P-labeled polyadenylated sample (using α-^{32}P-ATP instead of ^{14}C-ATP for tailing) was analyzed as described above. The result, shown in Figure 3A, revealed a similar pattern, only slightly more diffuse, presumably due to heterogeneity of tail length and possibly to some degradation. It could be concluded inter alia that gene 4 RNA (the hemagglutinin gene) remained essentially intact and accepted a poly(rA) tail. The average length of the poly(rA) tail was 100 AMP residues (based on incorporated counts).

The mixture of eight X47 RNAs (containing the hemagglutinin and neuraminidase genes from A/Victoria/3/75) was then polyadenylated with ^{14}C-ATP. It was calculated that an average of 100 AMP residues had been added to the influenza RNAs. This polyadenylated RNA was converted into dsDNA by a stepwise procedure involving avian myeloblastosis virus reverse transcriptase, RNAase treatment,

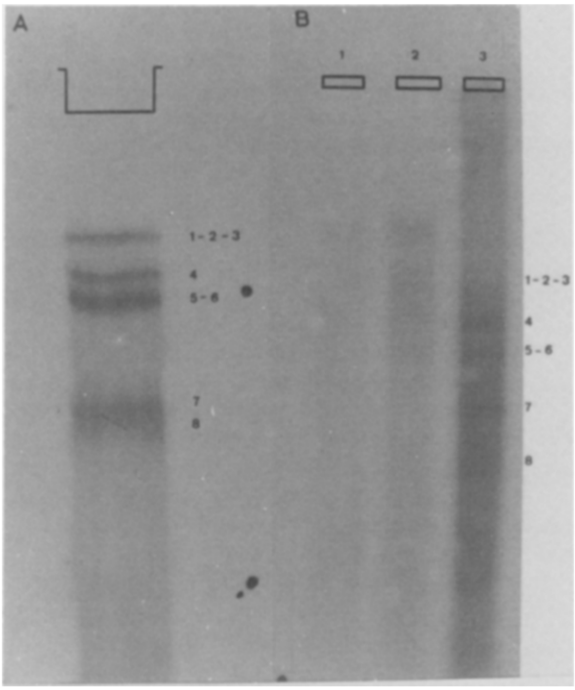

Figure 3. Electrophoretic Separation of "in Vitro" Polyadenylated Influenza RNA (X47) and of the Derived dsDNA

(A) 8 μg influenza RNA was polyadenylated with E. coli ATP:RNA adenyltransferase and α-^{32}P-ATP (1.4 Ci/mmole) for 3 min at 37°C. After phenol extraction, gel filtration and ethanol precipitation, the product was electrophoresed on a 3% polyacrylamide gel containing 7 M urea and 0.1% SDS.

(B) 80 μg total X47 influenza RNA polyadenylated in the presence of ^{14}C-ATP (15.6 Ci/mole) was used for synthesis of ^{32}P-labeled cDNA (α-^{32}P-dATP, spec. act. 0.35 Ci/mmole, was the labeled precursor). After treatment with ribonucleases A and T1, the cDNA was converted into the double-stranded form using E. coli DNA polymerase I (α-^{32}P-dATP, spec. act. 1.1 Ci/mmole, was the labeled dNTP). Following incubation with S1 nuclease, the influenza dsDNA was electrophoresed on a native 1.5% agarose gel (slot 3) together with untreated (slot 1) or S1-treated (slot 2) ^{32}P-labeled MS2 dsDNA. Gene 4 (hemagglutinin, 75 ng) was eluted, purified on hydroxylapatite and used for molecular cloning.

E. coli DNA polymerase I and S1 nuclease (for details see Experimental Procedures and Devos et al., 1979b). The dsDNA mixture was sized on a 1.5% agarose gel (Figure 3B), and presumed full-length gene 4 dsDNA was recovered from the gel as a source of material for cloning. At this step of the procedure we simply relied on the electrophoretic pattern for the identification of the genes, and we did not try to make precise measurements of the size of the product dsDNAs.

The virus-derived DNA copies were tailed by means of polynucleotide terminal transferase and TTP and cloned in the Pst I site of the plasmid pBR322 by the poly(dA)·poly(dT) tailing method. Alternatively, the insert was tailed with dCTP and cloned in the Hind III site of pBR322.

Identification of Bacterial Clones Carrying the Hemagglutinin Gene

118 tetracycline-resistant colonies were obtained with the hemagglutinin DNA preparation in experiment I, and 47 colonies were obtained in experiment II. As the insertion took place in the Pst I site, located approximately two thirds within the ampicillin gene, one would expect that plasmids which have received an insert will have lost their amp^R character. Indeed, 85 of the 118 colonies and 45 of the 47 colonies were found to be ampicillin-sensitive. All the colonies (including the ampicillin-resistant ones) were tested for the presence of a hemagglutinin gene insertion by using partially fragmented ^{32}P-labeled hemagglutinin RNA as a hybridization probe (Grunstein and Hogness, 1975). Ten of the colonies in experiment I hybridized with various efficiencies, whereas in experiment II five colonies showed positive hybridization (all about equally labeled). These colonies were all ampicillin-sensitive in both cases. We do not know the nature of the majority of the nonhybridizing clones. A control experiment in which tailed vector (that is, without addition of an insert) was used for transformation yielded only a few colonies. Cross-hybridization experiments (probe from gene 5 with the colonies obtained from gene 4 dsDNA and vice versa) also revealed only one or two possible positives. Probably at least part of the population consisted of fragments of the higher molecular weight genes.

Construction of Double-Stranded Influenza DNA Molecules

The mixture of eight influenza RNA molecules (80 μg) was polyadenylated with E. coli ATP:RNA adenyltransferase and ^{14}C-ATP (15 μCi/μmole) for 4 min at 37°C in a reaction volume of 0.4 ml (Devos et al., 1976). The reaction was stopped by the addition of 3 μl 20% SDS. A 10 μl aliquot was used to measure macromolecular incorporation by TCA precipitation. The aqueous phase was extracted with an equal volume of a mixture of phenol:chloroform:isoamylalcohol (25:24:1) equilibrated with 50 mM Tris-HCl (pH 8.0), 1% SDS. 50 μl 2 M potassium acetate and 10 μl 0.2 M EDTA were added to the aqueous phase and the RNA was precipitated with ethanol. The precipitate was directly used for reverse transcription under conditions favoring full-size product synthesis in a reaction volume of 0.1 ml (Kacian and Myers, 1976; Devos et al., 1979b). After 30 min at 41°C, the reaction was stopped with 10 μl 0.2 M EDTA and the incorporation was measured on an aliquot by TCA precipitation. The reaction volume was extracted with phenol as described above, passed through a G50 Sephadex column and precipitated with ethanol.

The RNA strand of the DNA·RNA hybrid was removed by treatment with RNAases A and T1 (Devos et al., 1979b). The second DNA strand was then synthesized on the mixture of influenza cDNAs by means of E. coli DNA polymerase I and the terminal loop was subsequently cleaved with S1 nuclease (Efstratiadis et al., 1976). The reaction was stopped with EDTA and the solution was extracted with phenol as described above. The product was precipitated with ethanol and dissolved in 40 μl 20 mM EDTA (pH 8.0), heated for 20 min at 55°C and loaded on a 1.5% agarose gel. Presumed full-length gene 4 dsDNA, gene 5 + 6 dsDNA, gene 7 dsDNA and gene 8 dsDNA were eluted, purified by adsorption to hydroxylapatite, eluted and precipitated with ethanol.

Cloning

The poly(dA)·poly(dT) joining procedure as originally proposed by Jackson, Symons and Berg (1972) was used to construct molecular chimeras consisting of the influenza hemagglutinin dsDNA and the bacterial plasmid pBR322 DNA. Pst I-cleaved pBR322 DNA was polydeoxyadenylated by means of terminal deoxynucleotidyl transferase in the presence of α-^{32}P-dATP (spec. act. 0.3 mCi/μmole) at a final concentration of 0.2 mM and Co^{++} at a final concentration of 1 mM; incubation was for 9 min at 37°C. Polyadenylated material was purified by two passages through an oligo(dT)-cellulose column. The poly(dA)-pBR322 DNA had an average tail length of 50 residues, as estimated on the basis of radioactive counts incorporated. A poly(dT) tail was added to the purified hemagglutinin dsDNA under the same conditions, except that dATP was replaced by ^{3}H-dTTP (spec. act. 10 mCi/μmole) at a concentration of 0.2 mM. Incubation was for 5 min at 37°C. The reaction was stopped and the solution was extracted with phenol, gel-filtrated and precipitated as indicated above. Incorporation measurements revealed a tail length of approximately 60 residues (experiment I) and of approximately 300 residues (experiment II).

Poly(dA)-pBR322 DNA (200 ng) and poly(dT)-hemagglutinin DNA (~70 ng) were annealed for 10 min at 65°C in 0.1 ml 10 mM Tris-HCl (pH 7.5), 100 mM NaCl, 1 mM EDTA and slowly cooled to room temperature over a period of 4 hr. E. coli HB101 cells were transformed as described by Lederberg and Cohen (1974). The transformation mixture was plated out on Luria agar plates containing 10 μg/ml tetracycline. In another experiment, hemagglutinin DNA was synthesized as described by Emtage, Catlin and Carey (1979) and was tailed with dCTP; pBR322 DNA opened at the Hind III site was tailed with dGTP. The average number of residues added was 20-30. Annealing and cloning were carried out as described above for the poly(dA)·poly(dT)-linked molecules.

Tetracycline-resistant colonies were transferred to master plates containing 100 μg/ml carbenicillin to test for ampicillin resistance and to nitrocellulose filters for colony hybridization. Colony hybridization was according to a simplified version of the method of Grunstein and Hogness (1975); the proteinase K treatment and the chloroform and 0.3 M NaCl washings before the actual hybridization were omitted. ^{32}P-hemagglutinin RNA probe was obtained as follows: the mixture of influenza RNAs (unlabeled) was separated on a denaturing polyacrylamide gel; the banding pattern was visualized by ethidium bromide staining; and the RNA band corresponding to the hemagglutinin gene (band 4) was eluted from the gel. The RNA was partially degraded to yield fragments of approximately 100 nucleotides using an adaptation of the method of Coffin and Billeter (1976). These RNA fragments were then end-labeled by means of γ-^{32}P-ATP and T4 polynucleotide kinase (Maxam and Gilbert, 1977). The ^{32}P-labeled polynucleotides were isolated by gel filtration on a G50 Sephadex column and precipitated with ethanol.

The experiments involving recombinant DNA were carried out in Category III physical containment facilities as specified in the U.K. "Williams Guidelines."

Restriction Endonuclease Digestions

Restriction digests were carried out as specified in the technical data sheet of the supplier, occasionally using 2 fold over digestion conditions. A physical restriction map was derived according to the method of Smith and Birnstiel (1976).

5′ and 3′ Terminal Labeling and DNA Sequencing

The 5′ ends of DNA fragments were dephosphorylated with bacterial alkaline phosphatase and kinase-labeled using γ-^{32}P-ATP and T4 polynucleotide kinase (Maxam and Gilbert, 1977). For 3′ terminal labeling we used T4 DNA polymerase and α-^{32}P-dNTP according to the procedure of Soeda, Kimura and Miura (1978). The labeled ends (5′ or 3′) were separated by cleaving the fragment with an appropriate restriction enzyme and separation of the fragments on an acrylamide gel. Sequencing was performed according to the method of Maxam and Gilbert (1977). 10% (short run) and 8% (long run) acrylamide gels 90 cm in length were routinely used (long run means that the first 30-50 nucleotides were allowed to run off the gel).

Acknowledgments

We are much indebted to Dr. P. Palese, Dr. C. Scholtissek, Dr. J. Skehel and Dr. M. Waterfield for valuable comments, information and materials, and to Dr. M. Ysebaert for computer analysis of the sequence. The research in Ghent was supported by grants from the Fonds voor Kollektief Fundamenteel Onderzoek and from the Gekoncerteerde Akties of the Belgian Ministry of Science. Research at Searle was carried out with facilities provided by Dr. A. J. Hale. D.H. and M.V. thank the IWONL for fellowships.

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

Received December 13, 1979

References

Air, G. M. (1979). Nucleotide sequence coding for the 'signal peptide' and N-terminus of the hemagglutinin from an Asian (H2N2) strain of influenza virus. Virology *97*, 468-472.

Blobel, G. and Dobberstein, B. (1975). Transfer of proteins across membranes. I. Presence of proteolytically processed and unprocessed nascent immunoglobulin light chains on membrane-bound ribosomes of murine myeloma. J. Cell Biol. *67*, 835-851.

Blobel, G., Walter, P., Chang, C. N., Goldman, B. M., Erickson, A. H. and Lingappa, R. (1979). Translocation of proteins across membranes: the signal hypothesis and beyond. In Society for Experimental Biology Symposium XXXIII, Secretory Mechanisms, C. R. Hopkins and C. J. Duncan, eds. (Cambridge, England: Cambridge University Press).

Bosch, F. X., Orlich, M., Klenk, H.-D. and Rott, R. (1979). The structure of the hemagglutinin, a determinant for the pathogenicity of influenza viruses. Virology *95*, 197-207.

Both, G. W. and Air, G. M. (1979). Nucleotide sequence coding for the N-terminal region of the matrix protein of influenza virus. Eur. J. Biochem. *96*, 363-372.

Coffin, J. M. and Billeter, M. A. (1976). A physical map of the Rous Sarcoma virus genome. J. Mol. Biol. *100*, 293-318.

Desselberger, U., Racaniello, V. R., Zazra, J. R. and Palese, P. (1980). The 3′ and 5′ terminal sequences of influenza A, B and C virus genes are highly conserved and show partial inverted complementarity. Gene, *8*, 315-328.

Devos, R., Gillis, E. and Fiers, W. (1976). The enzymic addition of poly(A) to the 3′-end of RNA using bacteriophage MS2 RNA as a model system. Eur. J. Biochem. *62*, 401-410.

Devos, R., Contreras, R., van Emmelo, J. and Fiers, W. (1979a). Identification of the translocatable element IS1 in a molecular chimera constructed with plasmid pBR322 DNA into which a bacteriophage MS2 DNA copy was inserted by the poly(dA)·poly(dT) linker method. J. Mol. Biol. *128*, 621-632.

Devos, R., van Emmelo, J., Contreras, R. and Fiers, W. (1979b). Construction and characterization of a plasmid containing a nearly full-size DNA copy of bacteriophage MS2 RNA. J. Mol. Biol. *128*, 595-619.

Domingo, E., Sabo, D., Taniguchi, T. and Weissmann, C. (1978). Nucleotide sequence heterogeneity of an RNA phage population. Cell *13*, 735-744.

Drzenick, R., Seto, J. T. and Rott, R. (1966). Characterization of neuraminidases from myxoviruses. Biochim. Biophys. Acta *128*, 547-558.

Efstratiadis, A., Kafatos, F. C. and Maniatis, T. (1977). The primary structure of rabbit β-globin mRNA as determined from cloned DNA. Cell *10*, 571-585.

Efstratiadis, A., Kafatos, F. C., Maxam, A. M. and Maniatis, T. (1976). Enzymatic in vitro synthesis of globin genes. Cell *7*, 279-288.

Elder, K. T., Bye, J. M., Skehel, J. J., Waterfield, M. D. and Smith, A. E. (1979). In vitro synthesis, glycosylation and membrane insertion of influenza virus haemagglutinin. Virology *95*, 343-350.

Emtage, J. S., Catlin, G. H. and Carey, N. H. (1979). Polyadenylation and reverse transcription of influenza viral RNA. Nucl. Acids Res. *6*, 1221-1239.

Fiers, W. (1979). Structure and function of RNA bacteriophages. In Comprehensive Virology, *13*, H. Fraenkel-Conrat and R. R. Wagner, eds. (New York: Plenum Press), pp. 69-204.

Fiers, W., Contreras, R., Haegeman, G., Rogiers, R., Van de Voorde, A., Van Heuverswyn, H., Van Herreweghe, J., Volckaert, G. and Ysebaert, M. (1978). Complete nucleotide sequence of SV40 DNA. Nature *273*, 113-120.

Furthmayr, H., Galardy, R. E., Tomita, M. and Marchesi, V. T. (1978). The intramembranous segment of human erythrocyte glycophorin A. Arch. Biochem. Biophys. *185*, 21-29.

Garoff, H. and Symons, K. (1974). Location of the spike glycoproteins in the Semliki Forest virus membrane. Proc. Nat. Acad. Sci. USA *71*, 3988-3992.

Gerhard, W. and Webster, R. G. (1978). Antigenic drift in influenza A viruses. I. Selection and characterization of antigenic variants of A/PR/8/34 (HON1) influenza virus with monoclonal antibodies. J. Exp. Med. *148*, 383-392.

Gething, M. J., White, J. M. and Waterfield, M. D. (1978). Purification of the fusion protein of Sendai virus: analysis of the NH_2-terminal sequence generated during precursor activation. Proc. Nat. Acad. Sci. USA *75*, 2737–2740.

Goldman, D. W., Pober, J. S., White, J. and Bayley, H. (1979). Selective labelling of the hybrophobic segments of intrinsic membrane proteins with a lipophilic photogenerated carbene. Nature *280*, 841–843.

Grunstein, M. and Hogness, D. S. (1975). Colony hybridization: a method for the isolation of cloned DNAs that contain a specific gene. Proc. Nat. Acad. Sci. USA *72*, 3961–3965.

Hamlyn, P. H., Brownlee, G. G., Cheng, C.-C., Gait, M. J. and Milstein, C. (1978). Complete sequence of constant and 3′ noncoding regions of an immunoglobulin mRNA using the dideoxynucleotide method of RNA sequencing. Cell *15*, 1067–1075.

Hirst, G. K. (1942). The quantitative determination of influenza virus and antibodies by means of red cell agglutination. J. Exp. Med. *75*, 49–64.

Homma, M. and Ohuchi, M. (1973). Trypsin action on the growth of Sendai virus in tissue culture cells. III. Structural difference of Sendai viruses grown in eggs and tissue culture cells. J. Virol. *12*, 1457–1465.

Jackson, D. C., Dopheide, T. A., Russell, R. J., White, D. O. and Ward, C. W. (1979). Antigenic determinants of influenza viruse hemagglutinin. II. Antigenic reactivity of the isolated N-terminal cyanogen bromide peptide of A/Memphis/72 hemagglutinin heavy chain. Virology *93*, 458–465.

Jackson, D., Symons, R. and Berg, P. (1972). Biochemical method for inserting new genetic information into DNA of simian virus 40: circular SV40 DNA molecules containing lambda phage genes and the galactose operon of Escherichia coli. Proc. Nat. Acad. Sci. USA *69*, 2904–2909.

Johnsrud, L. (1979). DNA sequence of the translocatable element IS1. Mol. Gen. Genet. *169*, 213–218.

Kacian, D. L. and Myers, J. C. (1976). Synthesis of extensive, possibly complete, DNA copies of poliovirus RNA in high yields and at high specific activities. Proc. Nat. Acad. Sci. USA *73*, 2191–2195.

Kahn, M., Kolter, R., Thomas, C., Figursky, D., Meyer, R., Remaut, E. and Helinsky, D. R. (1980). Plasmid cloning vehicles derived from plasmids colE1, F, R6K and RK2. In Methods of Enzymology, *68*, R. Wu, ed. (New York: Academic Press), in press.

Kilbourne, E. D. (1978). Genetic dimorphism in influenza viruses: characterization of stably associated hemagglutinin mutants differing in antigenicity and biological properties. Proc. Nat. Acad. Sci. USA *75*, 6258–6262.

Klenk, H.-D., Rott, R. and Orlich, M. (1977). Further studies on the activation of influenza A virus by proteolytic cleavage of the haemagglutinin. J. Gen. Virol. *36*, 151–161.

Klenk, H.-D., Rott, R., Orlich, M. and Blödorn, J. (1975). Activation of influenza viruses by trypsin treatment. Virology *68*, 426–439.

Krug, R. M., Broni, B. A. and Bouloy, M. (1979). Are the 5′ ends of influenza viral mRNAs synthesized in vivo donated by host mRNAs? Cell *18*, 329–334.

Laver, W. G. and Kilbourne, E. D. (1966). Identification in a recombinant influenza virus of structural proteins derived from both parents. Virology *30*, 493–501.

Laver, W. G. and Webster, R. G. (1977). Hemagglutinin molecules of Hong Kong, Equine 2 and Duck Ukraine influenza viruses lack N-terminal aspartic acid. Virology *81*, 482–485.

Laver, W. G. and Webster, R. G. (1979). Ecology of influenza viruses in lower mammals and birds. Brit. Med. Bull. *35*, 29–33.

Laver, W. G., Air, G. M., Webster, R. G., Gerhard, W., Ward, C. W. and Dopheide, T. A. A. (1979). Antigenic drift in type A influenza viruses: sequence differences in the hemagglutinin of Hong Kong (H3N2) variants selected with monoclonal hybridoma antibodies. Virology *98*, 226–237.

Lazarowitz, S. G. and Choppin, P. W. (1975). Enhancement of the infectivity of influenza A and B viruses by proteolytic cleavage of the hemagglutinin polypeptide. Virology *68*, 440–454.

Lederberg, E. M. and Cohen, S. N. (1974). Transformation of Salmonella typhimurium by plasmid deoxyribonucleic acid. J. Bacteriol. *199*, 1072–1074.

McCauley, J., Bye, J., Elder, K., Gething, M. J., Skehel, J. J., Smith, A. and Waterfield, M. D. (1979). Influenza virus hemagglutinin signal sequence. FEBS Letters, *108*, 422–426.

McGeoch, D. J. and Turnbull, N. T. (1978). Analysis of the 3′-terminal nucleotide sequence of vesicular stomatitis virus N protein mRNA. Nucl. Acids Res. *5*, 4007–4024.

McGeoch, D. J., Fellner, P. and Newton, C. (1976). Influenza virus genome consists of eight distinct RNA species. Proc. Nat. Acad. Sci. USA *73*, 3045–3049.

McReynolds, L., O'Malley, B. W., Nisbet, A. D., Fothergill, J. E., Girol, D., Fields, S., Robertson, M. and Brownlee, G. G. (1978). Sequence of chicken ovalbumin mRNA. Nature *273*, 723–728.

Maniatis, T., Jeffrey, A. and van de Sande, H. (1975). Chain length determination of small double- and single-stranded DNA molecules by polyacrylamide gel electrophoresis. Biochemistry *14*, 3787–3794.

Maxam, A. M. and Gilbert, W. (1977). A new method for sequencing DNA. Proc. Nat. Acad. Sci. USA *74*, 560–564.

Min Jou, W. and Fiers, W. (1976). Studies on the bacteriophage MS2. XXXIII. Comparison of the nucleotide sequences in related bacteriophage RNAs. J. Mol. Biol. *106*, 1047–1060.

Neuberger, A., Gottschalk, A., Marshall, R. D. and Spiro, R. G. (1972). Carbohydrate-peptide linkages in glycoproteins and methods for their elucidation. In The Glycoproteins: Their Composition, Structure and Function, A. Gottschalk, ed. (Amsterdam: Elsevier), pp. 450–490.

Ohmori, H., Tomizawa, J. I. and Maxam, A. M. (1978). Detection of 5-methylcytosine in DNA sequences. Nucl. Acids Res. *5*, 1479–1485.

Ohtsubo, H. and Ohtsubo, E. (1978). Nucleotide sequence of an insertion element IS1. Proc. Nat. Acad. Sci. USA *75*, 615–619.

Ovchinnikov, Y. A., Abdulaev, N. G., Feigina, M. Y., Kiselev, A. V. and Lobanov, N. A. (1979). The structural basis of the functioning of bacteriorhodopsin: an overview. FEBS Letters *100*, 219–224.

Palese, P. and Schulman, J. L. (1976a). Differences in RNA patterns of influenza A viruses. J. Virol. *17*, 876–884.

Palese, P. and Schulman, J. L. (1976b). Mapping of the influenza virus genome: identification of the hemagglutinin and the neuraminidase genes. Proc. Nat. Acad. Sci. USA *73*, 2142–2146.

Plotch, S. J., Bouloy, M. and Krug, R. M. (1979). Transfer of 5′-terminal cap of globin mRNA to influenza viral complementary RNA during transcription in vitro. Proc. Nat. Acad. Sci. USA *76*, 1618–1622.

Porter, A. G., Fellner, P., Black, D. N., Rowlands, D. J., Harris, T. J. R. and Brown, F. (1978). 3′-terminal nucleotide sequences in the genome RNA of picornaviruses. Nature *276*, 298–301.

Porter, A. G., Barber, C., Carey, N. H., Hallewell, R. A., Threlfall, G. and Emtage, J. S. (1979). Complete nucleotide sequence of an influenza virus haemagglutinin gene from cloned DNA. Nature *282*, 471–477.

Potter, C. W. and Oxford, J. S. (1979). Determination of immunity to influenza infection in man. Brit. Med. Bull. *35*, 69–75.

Proudfoot, N. J. and Brownlee, G. G. (1976). 3′ non-coding region sequences in eukaryotic messenger RNA. Nature *263*, 211–214.

Robb, R. J., Terhorst, C. and Strominger, J. L. (1978). Sequence of the COOH-terminal hydrophilic region of histocompatibility antigens HLA-A2 and HLA-B7. J. Biol. Chem. *253*, 5319–5324.

Roberts, R. J. (1978). Restriction and modification enzymes and their recognition sequences. Gene *4*, 183–193.

Robertson, J. S. (1979). 5′ and 3′ terminal nucleotide sequences of the RNA genome segments of influenza virus. Nucl. Acids Res. *6*, 3745–3757.

Scheid, A. and Choppin, P. W. (1974). Identification of biological activities of paramyxovirus glycoproteins. Activation of cell fusion, hemolysis, and infectivity by proteolytic cleavage of an inactive precursor protein of Sendai virus. Virology *57*, 475–490.

Schmidt, M. F. G., Bracha, M. and Schlesinger, M. J. (1979). Evidence for covalent attachment of fatty acids to Sindbis virus glycoproteins. Proc. Nat. Acad. Sci. USA *76*, 1687–1691.

Scholtissek, C. (1979). The genes coding for the surface glycoproteins of influenza A viruses contain a small conserved and a large variable region. Virology *93*, 594–597.

Scholtissek, C., Harms, E., Rohde, W., Orlich, M. and Rott, R. (1976). Correlation between RNA fragments of Fowl Plague virus and their corresponding gene functions. Virology *74*, 332–344.

Scholtissek, C., Rohde, W., Harms, E. and Rott, R. (1977). Correlation between base sequence homology of RNA segment 4 and antigenicity of the hemagglutinin of influenza viruses. Virology *79*, 330–336.

Skehel, J. J. and Waterfield, M. D. (1975). Studies on the primary structure of the influenza virus hemagglutinin. Proc. Nat. Acad. Sci. USA *72*, 93–97.

Skehel, J. J. and Hay, A. J. (1978). Nucleotide sequences at the 5′-termini of influenza virus RNAs and their transcripts. Nucl. Acids Res. *5*, 1207–1219.

Smith, H. O. and Birnstiel, M. L. (1976). A simple method for DNA restriction site mapping. Nucl. Acids Res. *3*, 2387–2398.

Soeda, E., Kimura, G. and Miura, K.-L. (1978). Similarity of nucleotide sequences around the origin of DNA replication in mouse polyoma virus and simian virus 40. Proc. Nat. Acad. Sci. USA *75*, 162–166.

Springer, T. A. and Strominger, J. L. (1976). Detergent-soluble HLA antigens contain a hydrophilic region at the COOH-terminus and a penultimate hydrophobic region. Proc. Nat. Acad. Sci. USA *73*, 2481–2482.

Tomita, M. and Marchesi, V. T. (1975). Amino-acid sequence and oligosaccharide attachment sites of human erythrocyte glycophorin. Proc. Nat. Acad. Sci. USA *72*, 2964–2968.

Van Assche, W., Vandekerckhove, J., Gielen, J. and Van Montagu, M. (1972). Anti-serum-resistant mutants of the RNA bacteriophage MS2. Arch. Intern. Physiol. Biochim. *80*, 410–411.

Van Heuverswyn, H., Van de Voorde, A., Van Herreweghe, J., Volckaert, G., DeWinne, P. and Fiers, W. (1980). Nucleotide sequence of simian virus 40 DNA: structure of the middle segment of the HindII + III restriction fragment B (sixth part of the T antigen gene) and codon usage. Eur. J. Biochem., in press.

Ward, C. W. and Dopheide, T. A. (1979). Primary structure of the Hong Kong (H3) haemagglutinin. Brit. Med. Bull. *35*, 51–56.

Waterfield, M. D., Espelie, K., Elder, K. and Skehel, J. J. (1979). Structure of the haemagglutinin of influenza virus. Brit. Med. Bull. *35*, 57–63.

Wiley, D. C. and Skehel, J. J. (1977). Crystallization and X-ray diffraction studies on the haemagglutinin glycoprotein from the membrane of influenza virus. J. Mol. Biol. *112*, 343–347.

Wong, R. S. C., Hofmann, T. and Bennick, A. (1979). The complete primary structure of a proline-rich phosphoprotein from human saliva. J. Biol. Chem. *254*, 4800–4808.