Volume 10 Number 3 1982 Nucleic Acids Research

The sequence of the nucleoprotein gene of human influenza A virus, strain A/NT/60/68

J.A.Huddleston and G.G.Brownlee

Sir William Dunn School of Pathology, University of Oxford, South Parks Road, Oxford OX1 3RE, UK

Received 8 December 1981; Accepted 15 January 1982

ABSTRACT

The nucleotide sequence of the nucleoprotein gene of influenza A/NT/60/68 was established after using improved cloning methods to obtain full length cDNA clones in pBr322. The gene is 1565 residues long and codes for a basic protein of 498 amino acids. There are only 30 amino acid differences between it and the homologous sequence in A/PR/8/34, all occurring as point mutations. Assuming a common lineage, the evolutionary rate of divergence of the two strains is 0.18% amino acid per year. This confirms there is a slow but significant rate of evolution.

INTRODUCTION

Influenza A virus is a negative stranded segmented RNA virus with 8 essential segments coding for at least 10 genes (1). Molecular analyses of genes 4 and 6, which code for the haemagglutinin and neuraminidase, respectively, have been pursued in order to understand the molecular evolution of these surface proteins as they adapt to evade neutralization by the immune system. For example, the haemagglutinin is a highly variable protein with a rate of evolution of its HA1 subunit within one subtype close to 1% amino acid change per year (2). In addition occasional reassortment of genes occurs which effectively introduces entirely new viruses possessing a haemagglutinin with as much as 65% of its amino acids altered (3).

By contrast, some of the other genes have been less well studied. We chose here to study the gene 5 coding for the nucleoprotein - the type-specific antigen, to establish the rate of evolution of a protein whose evolution is apparently not influenced by selection imposed by the host immune system.

The nucleoprotein appears to be a multifunctional molecule. Firstly it interacts with the individual viral RNA segments to form discreet ribo-nucleoprotein complexes corresponding to each RNA segment (4). Further these structures specifically bind matrix protein (4) suggesting the

0305-1048/82/1003-1029$2.00/0

structures are precursors in the assembly of the intact virus. Secondly, the analyses of temperature sensitive mutants of the nucleoprotein (5,6) suggest it is also involved in viral replication.

The nucleoprotein was originally thought to be stable in all influenza A strains but is now known from recent antigenic studies with polyclonal (7) and monoclonal antibodies (8) to undergo antigenic variation. Further it is suggested that both point mutations and genetic reassortment contribute to this variation (8,9). To provide fuller details of the evolution of this protein, we chose to sequence the nucleoprotein of a 1968 influenza strain, A/NT/60/68, using recombinant DNA methods. This strain was sufficiently distant from the 1934 strain A/PR/8/34 whose sequence is known (10,11), to allow us to estimate its rate of evolution reasonably accurately.

MATERIALS AND METHODS

Preparation of gene 5 full-length cDNA clones

a) Double strand DNA synthesis

Full-length $[^{32}P]$-cDNA was prepared from A/NT/60/68 virion RNA (kindly supplied by Dr B M Moss) by modification of our previous protocol (12) as follows. 1.5 μg of the synthetic primer d(A-G-C-A-A-A-A-G-C-A-G-G), complementary to the 3' end of all virion RNA segments (13) was phosphorylated at its 5' hydroxyl end using 10 μCi of $\gamma[^{32}P]$-ATP (3,000 Ci/mMole) and 5u of T4 phosphokinase (Boehringer) in a 10 μl reaction containing a final concentration of 0.1mM ATP, 50mM Tris-HCl, pH 7.5, 10mM $MgCl_2$ and 10mM dithiothreitol (DTT) for 1h at 37°C in a sealed capillary tube. After heat inactivation (90°C, 5 min) the reaction mix was added to a reverse transcriptase reaction using 20 μg virion RNA and 80u reverse transcriptase in a 100 μl reaction containing 20 μCi $\alpha[^{32}P]$-dATP (Amersham 3,000 Ci/mMole) and a final concentration of 50mM Tris-HCl pH 8.0, 5mM $MgCl_2$, 5mM DTT, 70mM KCl and 0.5mM of each of dATP, dCTP, dGTP and dTTP. Incubation was for 1h at 42° after which phenol extraction and ethanol (3 vols) precipitation of the complementary DNA (cDNA) was carried out. The cDNA sample was loaded as a 5 cm band and fractionated on 3% denaturing acrylamide gel (14). After radioautography the full length segment 5 cDNA was eluted as before (15) except that 2M ammonium acetate was used. Second strand DNA synthesis was carried out using the 13-long primer complementary to the 3' end of cDNA and *E.coli* DNA polymerase I (Klenow subfragment) (Boehringer). 2 μg of this primer d(A-G-T-A-G-A-A-A-C-A-A-G-G) was phosphorylated with T4 phosphokinase

as above (except for the omission of the $\gamma[^{32}P]$- ATP). After heat inactivation, 0.4 µg was added to a 100 µl reaction containing the band 5 cDNA, 5u Klenow subfragment of DNA polymerase I (E.coli) and a final concentration of 0.5mM of each of dATP, dCTP, dGTP and dTTP, 50mM Hepes (Na^+) pH 7.0, 25mM NaCl, 5mM $MgCl_2$ and 1.5mM DTT. After incubation for 2h at 25°, phenol extraction and ethanol precipitation was used to deproteinize and concentrate the double-stranded DNA (ds DNA) in 25 µl deionized water. Analysis for ds DNA on a 4% native acrylamide gel (15) showed that about 30% of the cDNA had been converted to a faster moving discreet ds DNA band (results not shown).

b) Blunt-end ligation and cloning in pBr322

5 µg of supercoiled pBr322 was incubated with PvuII in 6mM Tris-HCl pH 7.5, 6mM $MgCl_2$, 60mM NaCl and 6mM β-mercaptoethanol using 10u of enzyme for 2h at 37° in a volume of 50 µl. 5 µl calf intestinal phosphatase (Boehringer) (0.1 mg/ml in 10mM Tris-HCl pH 8.0) was added and the incubation continued for 10 min at 37°C. After heat treatment at 70° for 10 min followed by vortexing with phenol for 5 min, the aqueous phase was extracted with excess ether (3 times) and made up to 70 µl giving "phosphatased" vector at an estimated concentration of 35 µg/µl.

Blunt-end ligation was carried out for 16h at 20° in a 10 µl volume using 10 µg of vector, 1 µl of band 5 ds DNA (see above) in 0.4mM ATP, 50mM Tris-HCl pH 7.4, 10mM $MgCl_2$ and 10mM DTT. Half of the ligation reaction was used to transform competent E.coli X1974 cells (16) using a high efficiency protocol (17) and plated in ampicillin-containing agar plates. 196 colonies grew and 77 of these were screened by Grunstein-Hogness hybridization (18) with short copy $\alpha[^{32}P]$-cDNA derived by reverse transcription of total virion RNA (2 µg) with the 12-long primer (see above) under conditions where $\alpha[^{32}P]$-dATP [20 µCi at 3000 Ci/mMole in a 20 µl reaction] was the sole source of dATP. 6 influenza positive clones were obtained and 1ml cultures of these clones were grown up overnight and mini-plasmid preparations were prepared (19). Sizing of the plasmids by 1% agarose gel electrophoresis indicated that 5 clones had identical mobilities, moving slower than marker pBr322, and one had a pBr322 mobility. Digestion of the 5 long clones with PvuII gave rise to a high MW band on agarose gel electrophoresis. This suggested that as the original PvuII site in pBr322 had been destroyed in the cloning, all 5 clones were full length clones with an internal PvuII site derived from gene 5 of influenza. DNA from one clone

(labelled NT/60/5/4) was allowed to transform competent E.coli HB101 cells and a preparation of recombinant plasmid was prepared from a 1l culture (20) giving a yield of 0.8 mg.

Sequencing of A/NT/60/5.4

Sequencing was carried out using the Maxam-Gilbert method (21) except for the formic acid protocol (22) for the A + G reaction. Radioactive restriction enzyme fragments were derived from HinfI and Sau3A digests "filled in" using E.coli DNA polymerase (Klenow subfragment) and the appropriate $\alpha[^{32}P]$- dNTPs before fractionation on 4 or 6% native acrylamide gels. Strand-separation or recutting with second restriction enzymes was used to prepare fragments labelled at one end for sequencing (21). The sequence from residues 1-220 and 900 to 1565 was sequenced on both strands to ensure accuracy.

M13 cloning of restriction fragments and sequencing

Ds DNA (see above) was cut separately with various restriction enzymes. AluI, HaeIII and HinfI digests (after "filling in" the HinfI site) were blunt-end ligated (see above) to HindII - phosphatased M13mp7 (23). Sau3A cut ds DNA was cloned into BamHI cut M13mp2(Bam) (24), and TaqI cut material was ligated into AccI cut M13mp7. Transformation of competent E.coli JM101 cells (23) was carried out and recombinants isolated by plating using inducer and an indicator to select for recombinants causing insertional inactivation of the production of β-galactosidase (23). Standard procedures were used for preparing single stranded DNA and for sequencing the recombinants using a 'universal' 17-long primer (25). One large section of sequence from residues 212-920 (Fig 1) was established unambiguously by sequence analysis of clones completely covering both strands. Data was handled and searched using the Staden computer programs (29).

RESULTS

Sequence of A/NT/60/68 nucleoprotein gene

Fig 1 shows the sequence of the gene which is 1565 residues long. It is definitively identified as the nucleoprotein gene as it is homologous to the influenza A/PR/8/34 gene (10,11). It is also full length as it contains the common terminal sequences (13). Besides the short 5' non-coding region (45 residues) and the 3' non-coding region (26 residues including the U-A-A terminator), the gene codes for a protein of 498 amino acid residues

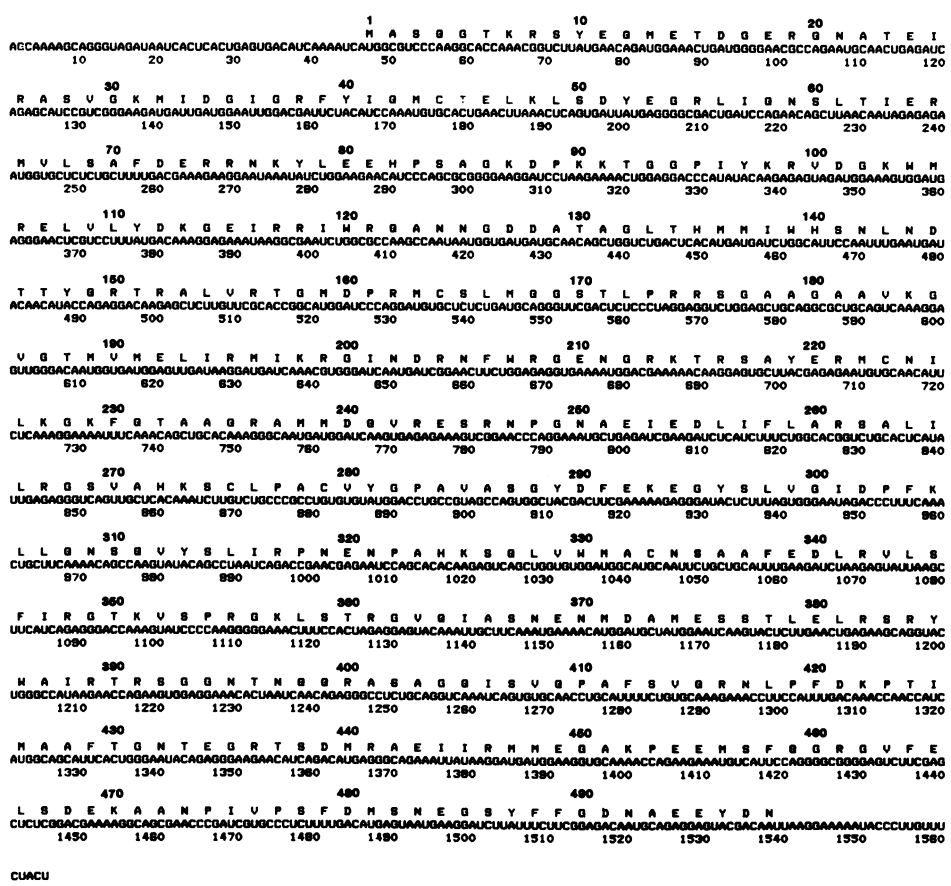

Fig. 1 The nucleotide and amino acid sequence of the nucleoprotein of A/NT/60/68 written in the mRNA sense. The amino acids are numbered above, and the nucleotides below their respective symbols. The first 12 and the last 13 nucleotide residues were derived from the universal terminal primers (see Methods) so that the actual sequence could differ slightly.

of 56,000 daltons. Its amino acid composition (Table 1) is remarkably arginine rich (this being the commonest amino acid) and is cysteine poor. The calculated net positive change at pH 7.0, assuming glutamic and aspartic acid are each -1, and arginine and lysine each +1, and histidine is +0.5 at this pH, is +13.5. This is in agreement with the high pI value of 8.7 (11). The codon usage (data not shown) is consistent with the usual low content of C-G in mammalian hosts and viruses (26); for example, the CGN codon for Arg is used 13 times whereas the AGN codon is used 33 times. Similarly the

Table 1 Amino acid composition of nucleoprotein*

Phe	(F)	17	Tyr	(Y)	15
Leu	(L)	32	His	(H)	5
Ile	(I)	27	Gln	(Q)	21
Met	(M)	26	Asn	(N)	27
Val	(V)	22	Lys	(K)	24
Ser	(S)	39	Asp	(D)	24
Pro	(P)	18	Glu	(E)	35
Thr	(T)	25	Cys	(C)	6
Ala	(A)	41	Trp	(W)	6
Arg	(R)	46	Gly	(G)	42

* Total number of residues = 498; MW 55,890

overall composition of the coding region of the cRNA (C 19.9%; A 32.7%; G 26.1% and U 21.3%) which is noticeably A-rich, is in general reflected in the fact that A is more commonly used in third positions than G.

DISCUSSION

a) Cloning of nucleoprotein gene

Our first sequence analysis by shotgun cloning of restriction fragments in M13 DNA (see Methods) failed to give a complete sequence (only residues 212-920 of Fig. 1 were obtained), so that full length cloning seemed desirable prior to further sequencing. The combination of specific priming, blunt-end ligation into pBr322 and Grunstein-Hogness hybridization with short influenza ^{32}P-cDNA probes, effectively allowed the selection of full length clones which were suitable for sequencing. This method would seem to have advantages over our previous M13 cloning method (12) where a significant number of clones shorter than full length were obtained.

b) Comparison with A/PR/8/34 nucleoprotein

Our A/NT/60/68 nucleoprotein sequence differs in length from one published A/PR/8/34 sequence (11) but is in agreement with a corrected version (Min Jou, personal communication) and a second independent sequence (10). The nucleotide sequence is also in very close agreement with our previous estimate prior to cloning of a length of 1560 residues (14).

Table 2 illustrates that the A/NT/60/68 nucleoprotein sequence has 30 amino acid changes compared to the Cambridge A/PR/8/34 version. 14 of these

Table 2 Amino acid changes between A/PR/8/34 (Cambridge) and A/NT/60/68

Nucleotide (Fig 1)	Mutation	Amino Acid	Change PR/8 → NT/60	Nucleotide (Fig 1)	Mutation	Amino Acid	Change PR/8 → NT/60
146	G → A	34	G → D	1102	G → T }	353	V → S
338	G → A	98	R → K	1103	T → C }		
346	A → G	101	N → D	1161	G → T	372	E → D
370	A → G	109	I → V	1162	A → G	373	T → A
386	A → G	114	E → G	1267	A → G	408	I → V
481	G → A	146	A → T	1276	A → G	411	T → A
625	G → A	194	V → I	1310	G → A	422	R → K
695	T → G	217	I → S	1312	A → C	423	T → P
752	A → G	236	K → R	1318	G → A	425	V → I
784	G → A	247	D → N	1369	A → G	442	T → A
802	T → A	253	F → I	1393	A → G	450	S → G
815	C → T	257	T → I	1400	G → A	452	R → K
923	G → A	293	R → K	1410	T → A	455	D → E
959	G → A	305	R → K	1411	G → A	456	V → M
1045	C → A	334	H → N	1463	G → A	473	S → N
1088	A → G	348	K → R				

changes are strictly conservative (I → V, R → K, D → E or vice versa, and F → I and V → M). All of the changes can be explained by single point mutations except that at amino acid 353 which requires two adjacent point mutations. 21 of the total of 31 point mutations causing amino acid changes involve A → G or G → A transitions. Some special mechanism must favour either the mutation or the selection of such transitions as compared with the other possibilities. We note that 94 silent mutations occur which is about 3 times the rate of non-silent mutations giving rise to the amino acid changes. The 5' non-coding region is identical in the two strains.

The distribution of amino acid changes between A/NT/60/68 and A/PR/8/34 is asymmetric, 10 occurring in the N-terminal half and 20 in the C-terminal half of the nucleoprotein. Also there are clusters of changes e.g. at amino acids 372 and 373; at 422, 423 and 425 and at 450, 452, 455 and 456. It is tempting to suggest that one or more of the clusters represents the antigenic determinants known to differ (8) between A/PR/8/34 and A/Hong Kong/68 (the latter being closely related to A/NT/60/68). Amino acid 423, within one of these clusters, is the best candidate for an antigenic site as there is a proline residue in A/NT/60/68 (Table 2), which is likely to cause a conformational change when compared with the A/PR/8/34 nucleoprotein

sequence.

c) Evolution of nucleoprotein compared with other influenza proteins

Table 3 shows an evolutionary comparison (expressed as % amino acid changes per year) for the nucleoprotein, the M1 and M2 matrix proteins and the NS1 and NS2 proteins. In all except the last example, two selected human strains were compared. For example, the number of amino acid changes between the nucleoprotein of A/PR/8/34 and A/NT/60/68 is 30/498 or 6.2%. Assuming mutation occurred at a linear rate from 1934 to 1968 and there is a common lineage, this is 0.18% per year.

In all cases sequences can be aligned without deletions or additions suggesting that the observed differences can be explained by point mutation, rather than reassortment. Assuming the % changes are significant, which would clearly require confirmation by the sequence analysis of intermediate strains, we note the very low value for matrix M1, the intermediate value of the nucleoprotein, the higher value for NS1 and M2 and the very high value for the haemagglutinin. These figures demonstrate the polymorphism attainable by mutation and natural selection in each of these molecules that is compatible with their functional role. But we must be aware that the overall figures (Table 3) do not take into account that some regions of the structures may be strictly conserved whereas others are free to diverge. Clearly in vivo the haemagglutinin is under strong selective pressure to change to evade neutralization by antibody. We must also assume that the other proteins including the nucleoprotein are under selective pressure, in

Table 3 Calculated % amino acid changes per year for nucleoprotein and other human influenza proteins

Protein	Data from	Strain compared	% amino acid per year
Nucleoprotein	ref 10	A/PR/8/34 v A/NT/60/68	0.18
Matrix M1	refs 1,30	" v A/UDORN/72	0.07
" M2		" v "	0.27
NS1	refs 12,31	" v "	0.31
NS2		" v "	0.15
Haemagglutinin H3 (HA1 subunit)	ref 2	A/NT/60/68 v A/Bangkok/79	1.0

that occasional mutations are either "neutral" or have conferred some selective advantage. The 0.18% figure for the nucleoprotein shows that mutations are being fixed by selection and that considerable polymorphism is possible in the protein. The M1 matrix protein is obviously less susceptible to change implying it has very stringent amino acid sequence requirements.

d) General points

The predicted primary amino acid sequence reported here gives us no information on the secondary or tertiary structure of the nucleoprotein. We do not know whether disulphide bonds interconnect the 6 cysteine residues, nor where the phosphorylated serine (27) is located. We would predict however that the lack of large clusters of basic amino acids and the fact that these are well dispersed over the entire length of the nucleoprotein suggests the protein-RNA interaction occurs over a high proportion of the length of the protein. Like others (10), we agree that the proteolytic processing (28) is likely to occur near the N-terminus rather than the C-terminus of the nucleoprotein.

ACKNOWLEDGEMENTS

We thank Mrs M Robertson for assistance at the beginning of this project, Dr J Beard for generously supplying reverse transcriptase, Dr S Fields for a gift of T4 ligase, and Dr A Caton and Professor D H L Bishop for critical reading of the manuscript. This work was supported by an MRC programme grant to GGB.

REFERENCES

1. Lamb, R.A., Lai, C-J. & Choppin, P.W. (1981) Proc.Nat.Acad.Sci.USA 78, 4170-4174
2. Sleigh, M.J. & Both, G.W. (1981) ICN-UCLA Symposium in "Genetic Variation among Influenza Viruses", in press
3. Winter, G., Fields, S. & Brownlee, G.G. (1981) Nature 292, 72-75
4. Rees, P.J. & Dimmock, N.J. (1981) J.Gen.Virol. 53, 125-132
5. Barry, R.D. & Mahy, B.W.J. (1979) Brit.Med.Bull. 35, 39-46
6. Scholtissek, C. (1978) Curr.Top.Microbiol.Immunol. 80, 139-169
7. Schild, G.C., Oxford, J.S. & Newman, R.W. (1979) Virology 93, 569-573
8. Van Wyke, K.L., Hinshaw, V.S., Bean, W.J. & Webster, R.G. (1980) J.Virol. 35, 24-30
9. Dimmock, N.J., Carver, A.S. & Webster, R.G. (1980) Virology 103, 350-356
10. Winter, G. & Fields, S. (1981) Virology 114, 423-428
11. Van Rompuy, L., Min Jou, W., Huylebroek, D., Devos, R. & Fiers, W. (1981) Eur.J.Biochem. 116, 347-353

12. Winter, G., Fields, S., Gait, M. & Brownlee, G.G. (1981) Nucleic Acids Res. 9, 237-245
13. Skehel, J.J. & Hay, A.J. (1978) Nucleic Acids Res. 5, 1207-1220
14. Sleigh, M.J., Both, G.W. & Brownlee, G.G. (1979) Nucleic Acids Res. 6, 1309-1321
15. Sleigh, M.J., Both, G.W. & Brownlee, G.G. (1979) Nucleic Acids Res. 7, 879-893
16. Seed, B., personal communication
17. Hanahan, D., personal communication
18. Grunstein, M. & Hogness, D.S. (1975) Proc.Nat.Acad.Sci.USA 72, 3961-3965
19. Holmes, D. & Quigley, M., Analyt.Biochem., in press
20. Birnboim, H.C. & Doly, J. (1979) Nucleic Acids Res. 7, 1513-1523
21. Maxam, A.M. & Gilbert, W. (1980) Methods Enzymol. 65, 499-560
22. Hill, D.F. & Petersen, G.B., personal communication
23. Messing, J., Crea, R. & Seeburg, P.H. (1981) Nucleic Acids Res. 9, 309-321
24. Rothstein, R.J., Lall, L.F., Bahl, C.P., Narang, S.A. & Wu, R. (1979) Methods in Enzymol. 68, 101-110
25. Duckworth, M.L., Gait, M.J., Goelet, P., Hong, G.F., Singh, M. & Titmas, R.C. (1981) Nucleic Acids Res. 7, 1691-1706
26. Subak-Sharpe, J.H. (1967) Br.Med.Bull. 23, 161-168
27. Privalsky, M.L. & Penhoet, E.E. (1977) J.Virol. 24, 401-405
28. Zhirnov, O.P. & Bukrinskaya, A.G. (1981) Virology 109, 174-179
29. Staden, R. (1980) Nucleic Acids Res. 8, 3673-3694
30. Winter, G. & Fields, S. (1980) Nucleic Acids Res. 8, 1965-1974
31. Lamb, R.A. & Lai, C.J. (1980) Cell 21, 475-485