The complete sequence of RNA segment 2 of influenza A/NT/60/68 and its encoded P1 protein

D.H.L.Bishop, J.A.Huddleston and G.G.Brownlee

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

Received 5 January 1982; Accepted 28 January 1982

ABSTRACT

A DNA copy of influenza A/NT/60/68 viral RNA segment 2, corresponding to protein P1, has been cloned in the *E.coli* plasmid pBr322. The clone is 2341 nucleotides long and represents a full-length copy of the viral RNA. In the viral complementary (plus sense) strand there is an open reading frame that is 2271 nucleotides long. The predicted primary gene product is a basic 86,300 dalton protein with a net charge at neutral pH of +23. A 29 amino acid stretch of the protein (coded by nucleotide residues 583-669) is highly basic and contains 7 lysine and 8 arginine residues. Other smaller clusters of basic amino acids are also present in the protein.

INTRODUCTION

The influenza A virus genome consists of 8, negative sense, single-stranded RNA species that code for 10 distinct gene products (1). Partial or complete nucleotide sequences have been reported (for one or another influenza A virus strain) for the RNA segments coding for the nonstructural polypeptides (NS1, NS2, segment 8, references 2-5), the matrix polypeptides (M1, M2, segment 7, references 6-8), the neuraminidase (NA, segment 6, references 9,10), the nucleoprotein (NP, segment 5, references 11-13), the haemagglutinin (HA, segment 4, for examples see references 8, 10, 14-19) and two of the three P polypeptides (segments 1 and 3 of influenza A/PR/8/34, reference 20). We report here the 2341 nucleotide sequence of a full-length DNA copy of segment 2 of A/NT/60/68 cloned in pBr322. The sequence data indicate that an open reading frame in the viral-complementary (plus sense) RNA codes for a 757 amino acid (86,314 daltons) basic protein. With the information reported here for segment 2, complete sequences are now available for each of the 8 RNA species of influenza A (albeit different virus strains).

MATERIALS AND METHODS

Preparation of full-length DNA for influenza A/NT/60/68 segments 1, 2 and 3

Full-length double-stranded DNA derived from RNA segments 1-3 of influenza A/NT/60/68 virus was synthesized by procedures similar to those described previously (5,13). In brief, 20 μ g of virion RNA (a generous gift of Dr. B.M. Moss) was used as a template for cDNA synthesis by reverse transcriptase and primed with a 5' phosphorylated dodecanucleotide d(pA-G-C-A-A-A-A-G-C-A-G-G). This oligonucleotide is complementary to the 3' terminal sequence of the viral RNA (21,22). The product DNA was denatured and resolved by polyacrylamide gel electrophoresis. The bands that corresponded in size to segments 1-3 were eluted and back-copied with the "Klenow" large fragment of *E.coli* DNA polymerase. For the back-copying reaction a 5' phosphorylated tridecanucleotide primer d(pA-G-T-A-G-A-A-A-C-A-A-G-G) was used. This oligonucleotide is complementary to the 3' end of full-length cDNA (21,22).

Blunt-end ligation into pBr322 of putative full-length DNA representing influenza A/NT/60/68 segments 1, 2 and 3

The back-copied DNA was blunt-end ligated into *E.coli* plasmid pBr322 that had been cut at its unique *Pvu* II site and, in order to inhibit self-ligation, treated with alkaline phosphatase to remove the terminal phosphates (13). The ligation mixtures were used to transform competent *E.coli* X1974 cells (23). Twelve clones containing influenza DNA inserts were identified by Grunstein-Hogness screening (24) using terminal copy 32 P-cDNA prepared by reverse transcription of 2 μ g viral RNA under limiting α - 32 P-ATP concentration (0.35 μ molar), in order to select for DNA clones with completely back-copied cDNA (13). Plasmid DNA from one clone, A/NT/60/68/2/62, identified as containing a segment 2 insert (see Results), was used to transform competent *E.coli* HB101 cells. From a 500 ml culture of transformed cells derived from one colony, 0.9 mg of purified plasmid DNA was obtained.

Sequence analyses of A/NT/60/68/2/62

Restriction enzymes were purchased from New England Biolabs Inc. (Beverly, MA) and used according to their assay instructions. Specific fragments of plasmid DNA were obtained by treatment with *Hinf* I. The fragments were labelled at their 3' ends by "filling-in" using the large fragment of *E.coli* DNA polymerase and α - 32 P-dATP, followed by a dATP chase. The fragments were resolved on 4% native polyacrylamide gels, eluted, and heat denatured

for 2 min in 40% DMSO at 90°C. The individual DNA strands were resolved on 5% native polyacrylamide gels, eluted and subjected to standard Maxam and Gilbert sequencing procedures (25), except that A + G specific digestions involved incubation of the DNA sample in 64% formic acid for 5 min at 18°C (26). The data generated provided sequences for both strands of most of the Hinf I fragments containing influenza DNA (Fig. 1, residues 1-123, 129-313, 321-400, 437-505, 514-758, 764-893, 999-1041, 1059-1105, 1109-1167, 1805-1839 and 1844-2020). Short Hinf I fragments, and 2 very long Hinf I sequences (greater than 600 residues each, i.e. residues 1168 through 1800 and 2025 to the end of the insert, plus 307 nucleotides of pBr322) were not amenable to sequencing by this procedure. Analysis of each strand of a Dde I restriction fragment provided the sequence data for residues 1075-1305; analyses of a Sau3A I fragment gave the data for residues 2015-2236; and those of a Taq I fragment gave the sequences for residues 230-590. Sequence data of residues 2029-2341 plus a contiguous pBr322 sequence were deduced from analyses of the digestion products obtained by recutting labelled Hinf I fragments with Acc I. Complementary sequence information came from analyses of "filled-in" Acc I products recut with BstN I (pBr322 sequences plus influenza residues 2341-2058). Single-strand sequence data were obtained by recutting with Pvu II the labelled DNA produced after "filling-in" Bgl II restriction fragments (residues 83-523); similarly, data were obtained by recutting labelled Dde I restriction fragments (residues 757-1068) with Pvu II, and recutting the "filled-in" Ava I digestion products with BstN I (for residues 1698-1943), or Hinf I (for residues 1257-1584 and 1587-1800).

Sequence data derived from DNA clone 371

A section of the sequence representing residues 518-1693 was independently obtained from a clone derived by the poly G-C tailing method and inserted into the Pst I site of pBr322 (27). The resulting clone, labelled 371, was sequenced both by the Maxam and Gilbert method and by the Sanger dideoxynucleotide method (28) after subcloning derived restriction fragments into M13mp2 (29). The sequence information obtained corroborated the sequence data shown in Fig. 1 for A/NT/60/68/2/62 except for residue 1644 which was clearly a G in clone 371 and an A in A/NT/60/68/2/62. For either nucleotide, the encoded amino acid is a glycine residue. Taken together with the sequence information obtained as described above for A/NT/60/68/2/62, more than 99% of the sequence was obtained from analyses of both strands of cloned DNA.

RESULTS

Isolation of DNA clones representing segment 2 of A/NT/60/68

Unlike the 5 other RNA species of influenza A virus, segments 1-3 have such similar molecular weights that their individual isolation and recovery are often difficult to achieve (30). We have encountered similar problems for their cDNA copies prepared by reverse transcription. We therefore sought to obtain full-length clones of DNA from single-stranded cDNA representing all 3 of these RNA species, relying on characterization of the ends of the DNA inserts to deduce their origins. Previous studies have shown that the 5' and 3' ends of each viral RNA species of influenza A virus can be distinguished after the conserved terminal 13 nucleotides (21,22).

Using the protocols outlined in Materials and Methods, 12 influenza positive clones were initially identified. From sizing analyses in 1% agarose gels, by comparison with pBr322, all 12 were judged to contain DNA inserts of the order of 2 kilobase pairs. The Hinf I restriction digest patterns of 4 of the clones (including A/NT/60/68/2/62) were identical to each other. They contained (results not shown) 12 fragments greater than 50 base pairs in length in addition to all the normal pBr322 Hinf I fragments (except for the 344 nucleotide sequence containing the Pvu II site). The A/NT/60/68/2/62 clone was sequenced, and was deduced to represent the segment 2 information of this virus strain (vide infra).

Sequence of clone A/NT/60/68/2/62

Fig. 1 shows the sequence obtained for clone A/NT/60/68/2/62 expressed as the viral complementary, messenger sense (plus), strand. It represents a 2341 nucleotide long sequence with 5' and 3' ends complementary to the known ends of influenza A viral RNA and is definitively identified as segment 2 by comparison of its terminal sequence with that deduced for influenza A/FPV/Rostock RNA segment 2 (21,22). The viral RNA has a calculated base composition of 35.2% U, 22.4% A, 22.7% C and 19.6% G. The non-coding region at the 3' end of the viral RNA is 24 nucleotides in length; at the 5' end, after the two contiguous termination codons (UAG and UGA, positions 2296-2301 in Fig. 1), the non-coding sequence is 40 nucleotides in length. There is only one long open reading frame in the viral complementary RNA strand. It is preceded by a UGA termination codon and codes for a primary gene product that is 757 amino acids in length (approximately 86,300 daltons). In the alternate reading frames, there are no open reading frames initiated by a methionine codon longer than 95 amino acids in length (e.g. nucleotides

HC	HA	HT	TL	FL	VP	AS	HA	IS	TF	PT	YS	PP	YS	H
10	20	30	40	50	60	70	80	90	100	110	120			
GT	GT	GT	GT	GT	GT	GT	GT	GT	GT	GT	GT	GT	GT	GT
130	140	150	160	170	180	190	200	210	220	230	240			
LP	ES	NE	PE	PS	ST	ST	ST	ST	ST	ST	ST	ST	ST	ST
250	260	270	280	290	300	310	320	330	340	350	360			
VU	ST	TR	UD	RL	TS	SR	ST	YS	MT	LM	RS	PA	AT	AL
370	380	390	400	410	420	430	440	450	460	470	480			
HL	TA	NE	SR	LI	DF	LE	DU	NE	SH	KE	EE	EE	IT	TH
490	500	510	520	530	540	550	560	570	580	590	600			
SH	NT	KE	NT	RT	IS	KE	SR	VE	SR	VE	SR	VE	SR	VE
610	620	630	640	650	660	670	680	690	700	710	720			
RE	KL	RR	AI	AT	PS	HS	IR	BF	UY	FU	ET	LA	RI	CE
730	740	750	760	770	780	790	800	810	820	830	840			
VB	NE	KE	KA	KL	AN	VU	RE	NT	HS	BT	EL	ST	IT	IS
850	860	870	880	890	900	910	920	930	940	950	960			
SN	PR	HL	FA	NI	TY	IT	NS	PE	NR	HL	SI	AP	IF	SH
970	980	990	1000	1010	1020	1030	1040	1050	1060	1070	1080			
KE	YH	FE	SK	HL	RT	SI	PA	EN	LA	SI	DL	KY	PH	NE
1090	1100	1110	1120	1130	1140	1150	1160	1170	1180	1190	1200			
RP	LI	DS	TA	LS	PS	PH	NS	PH	NS	PH	NS	PH	NS	PH
1210	1220	1230	1240	1250	1260	1270	1280	1290	1300	1310	1320			
ET	TY	MD	BL	SS	SD	FA	LI	VA	PH	NE	SI	AS	VD	RF
1330	1340	1350	1360	1370	1380	1390	1400	1410	1420	1430	1440			
VI	IN	SK	KE	SY	IN	RT	ST	FE	FT	ST	FE	FT	ST	FE
1450	1460	1470	1480	1490	1500	1510	1520	1530	1540	1550	1560			
SV	SI	NE	SA	DS	IS	VT	VI	KN	HN	HN	HN	HN	HN	HN
1570	1580	1590	1600	1610	1620	1630	1640	1650	1660	1670	1680			
RY	YT	TY	TR	CH	RS	DT	IS	TR	RS	FE	LE	KL	ME	ST
1690	1700	1710	1720	1730	1740	1750	1760	1770	1780	1790	1800			
DS	PH	LY	NI	HL	NI	PE	UC	LE	NE	LE	NE	LE	NE	LE
1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920			
SH	KE	IE	SV	NA	VU	HP	AS	PA	KS	NE	YD	AV	AT	TH
1930	1940	1950	1960	1970	1980	1990	2000	2010	2020	2030	2040			
SI	LT	YS	OR	SI	LE	DE	SH	YK	CN	LF	ER	FP	SS	YS
2050	2060	2070	2080	2090	2100	2110	2120	2130	2140	2150	2160			
SH	VE	AN	VS	RR	AR	IS	AR	IS	AR	IS	AR	IS	AR	IS
2170	2180	2190	2200	2210	2220	2230	2240	2250	2260	2270	2280			
LR	RR	RR	RR	RR	RR	RR	RR	RR	RR	RR	RR			
2290	2300	2310	2320	2330	2340									

Fig. 1 The nucleotide and amino acid sequence of the influenza DNA insert in clone A/NT/60/68/2/62 written in the viral complementary (plus strand) sense. The nucleotides are numbered below their respective residues. Amino acids corresponding to the encoded P1 polypeptide are centred over the corresponding nucleotide triplets. The first 12 and last 13 nucleotide residues were derived from the primers (see Methods) so that the RNA sequence could differ slightly.

1970-2254). In the viral RNA strand (negative sense RNA), the longest open reading frame initiated by an AUG triplet codes for a stretch of 121 amino acids (the complementary sequence to nucleotides 1373-1011, Fig. 1).

The codon usage of the segment 2 gene product (data not shown) is very similar to that of the nucleoprotein gene (13). Notable is the low usage of 4 of the arginine codons (CGN, used 11 times), compared with the high usage of the arginine AGG and AGA codons (used 40 times). Likewise the threonine codon ACG is used only 4 times whereas the other 3 threonine codons (ACU, ACC and ACA) are used a total of 56 times.

The amino acid composition of the primary gene product of the segment 2 RNA (Table 1) indicates that the 757 amino acid protein is rich in arginine and lysine residues. The calculated net positive charge at pH 7.0 is +23, assuming glutamic and aspartic acid are each -1, and arginine and lysine each +1, with histidine assumed to have a +0.5 charge at this pH.

Examination of the distribution of basic amino acids in the primary sequence of the gene 2 polypeptide reveals that one particular stretch of 29 amino acids (coded by the nucleotide sequence 583-669, Fig. 1) contains 8 arginines and 7 lysines with only 1 intermediate negatively charged residue (aspartic acid). Clusters of other basic amino acids also exist in the molecule (e.g. 2 lysine and 3 arginine residues in the 7 amino acid sequence coded by nucleotides 721-741; 3 lysines and 2 arginines in the 8 amino acid sequence coded by nucleotides 1180-1203; 4 lysines and 1 arginine in the 11 amino acid sequence coded by nucleotides 856-888). The significance of these basic amino acid clusters is not known. To what extent the 10 cysteine

Table 1 Amino acid composition of the segment 2 gene product

Ala (A)	42	Leu (L)	57
Arg (R)	51	Lys (K)	50
Asn (N)	48	Met (M)	37
Asp (D)	32	Phe (F)	33
Cys (C)	10	Pro (P)	32
Gln (Q)	31	Ser (S)	51
Glu (E)	51	Thr (T)	60
Gly (G)	46	Trp (W)	9
His (H)	10	Tyr (Y)	25
Ile (I)	45	Val (V)	37

Total number of residues = 757; MW 86,314

residues are involved in forming intramolecular disulphide bonds is also not known, nor the effect of the asymmetric distribution of proline residues (11 out of the total of 32 occurring in the first 100 amino acids).

DISCUSSION

The 8 nucleocapsid structures of influenza A virus consist of complexes of minus-stranded RNA and 4 species of protein (31). The principal protein is the nucleoprotein, NP. Three minor components of the nucleocapsids are the large P proteins (P1, P2, P3).

It has been deduced that the segment 2 RNA of A/FPV/Rostock codes for the P1 polypeptide of that virus (22,32-34). We conclude that the gene product of the A/NT/60/68/2/62 clone is also a P1 type protein since the 5' and 3' terminal sequences of the clone are comparable to those of segment 2 of A/FPV/Rostock (22). Thus, the 5' terminal sequence of the negative sense strand of the A/NT/60/68/2/62 clone is identical to the 66 terminal nucleotides at the 5' end of segment 2 of A/FPV/Rostock viral RNA except for a C (instead of a U at residue 2292 in A/NT/60/68/2/62; note the U is the complement of the A in the positive sense sequence shown in Fig. 1). Similarly, the 67 nucleotide sequence reported for the 3' end of segment 2 of A/FPV/Rostock viral RNA (22) is identical to that of the A/NT/60/68/2/62 clone, except for a C at position 4 (which is a U in the A/NT/60/68/2/62 negative sense sequence), and a G at residue 36 (corresponding to an A in the negative sense sequence of the clone). However, it should be noted that the position 4 residue of the cloned DNA is derived from the primer oligonucleotide (see Fig. 1).

The P polypeptides of A/NT/60/68 have not been analyzed by gel electrophoresis, however for all influenza A virus strains so far analyzed, the P1 protein is basic and migrates significantly slower in SDS discontinuous gels than the P2 and P3 polypeptides (35). The basic nature of the P1 protein is consistent with the amino acid data shown in Table 1. Although the P2 and P3 polypeptides have different relative mobilities in SDS gels for different influenza A viruses, it has been demonstrated that one is acidic and the other basic (35). Since the P1, P2 and P3 nomenclature is based on their relative mobilities in SDS discontinuous gels, the P2 is the acidic polypeptide for some influenza A virus strains, in others it is the P3 polypeptide (35).

Recent analyses of the 2233 nucleotide sequence of RNA segment 3 of A/NT/60/68 (D.H.L. Bishop, K. Jones, J.A. Huddleston & G.G. Brownlee, unpubl-

ished data) and A/PR/8/34 (20) have established that the encoded gene product is an acidic polypeptide (net charge at neutral pH of -13.5) that is composed of 716 amino acids (82,500 daltons).

For influenza A/PR/8/34 the 2341 nucleotide sequence of RNA segment 1 (20) codes for a basic polypeptide (net charge at neutral pH of +28) involving 759 amino acids (86,000 daltons), i.e. larger than that of the P1 protein. Although the number of amino acids predicted for the primary gene product of segment 2 of A/NT/60/68 is less than that of the segment 1 gene product of A/PR/8/34 (757 by comparison with 759), its calculated molecular weight is slightly greater (A/NT/60/68 segment 2 gene product: 86,344; A/PR/8/34 segment 1 gene product: 86,044). It appears doubtful that such a small difference is sufficient to account for the slower mobility of the P1 polypeptide by comparison with P2 and P3. The aberrant mobilities in SDS discontinuous gels are possibly due to differences in their intrinsic charge, or relative binding of SDS; alternatively, processing or other post-translational modifications of the polypeptides may account for the anomaly.

Taken together with the recent sequence data for segments 1 and 3, it is evident that the segments 1-3 of influenza A each code for a single large protein. In view of the variable phenotypes of the P2 and P3 polypeptides in SDS discontinuous gels we would favour the adoption of a nomenclature system for the P polypeptides that is based on their gene assignments. Thus, following the precedence of the coding assignments made for influenza A/FPV/Rostock (22), the basic P2 polypeptide is coded by viral RNA segment 1, the basic P1 polypeptide is coded by RNA segment 2 and the acidic P3 polypeptide is coded by segment 3.

ACKNOWLEDGEMENTS

We thank Dr J. Beard for generously supplying reverse transcriptase, Drs G. Winter and S. Fields for a preprint of their paper (20) reporting sequence analyses of segments 1 and 3 of influenza A/PR/8/34 and Dr A. Caton for critical reading of the manuscript. This work was supported by an MRC programme grant to GGB and by NIH grant AI 14183 to DHLB.

REFERENCES

1. Lamb, R.A., Lai, C.-J. and Choppin, P.A. (1981) *Proc.Natl.Acad.Sci.USA* **78**, 4170-4174.
2. Lamb, R.A. and Lai, C.-J. (1980) *Cell* **21**, 475-485.
3. Porter, A.G., Smith, J.C. and Emtage, J.S. (1980) *Proc.Natl.Acad.Sci. USA* **77**, 5074-5078.
4. Baez, M., Taussig, R., Zazra, J.J., Young, J.F., Palese, P., Reiseld, A.

- and Skalka, A.M. (1980) *Nucleic Acids Res.* 8, 5845-5858.
5. Winter, G., Fields, S., Gait, M.J. and Brownlee, G.G. (1981) *Nucleic Acids Res.* 9, 237-245.
 6. Winter, G. and Fields, S., (1980) *Nucleic Acids Res.* 8, 1965-1974.
 7. Allen, H., McCauley, J., Waterfield, M. and Gething, M.-J. (1980) *Virology* 107, 548-551.
 8. Gething, M.-J. and Allen, H. (1981) in "The Replication of Negative Strand Viruses" (D.H.L. Bishop and R.W. Compans, eds.), pp. 241-249, Elsevier/North Holland, New York.
 9. Fields, S., Winter, G. and Brownlee, G.G. (1981) *Nature* 290, 213-217.
 10. Air, G.M., Blok, J. and Hall, R.M. (1981) in "The Replication of Negative Strand Viruses" (D.H.L. Bishop and R.W. Compans, eds.), pp. 225-239, Elsevier/North Holland, New York.
 11. Winter, G. and Fields, S. (1981) *Virology* 114, 423-428.
 12. Van Rompuy, L., Min Jou, W., Huylebroeck, D., Devas, R. and Fiers, W. (1981) *Eur.J.Biochem.* 116, 347-353.
 13. Huddleston, J.A. and Brownlee, G.G. (1982) *Nucleic Acids Res.* in press.
 14. Winter, G., Fields, S. and Brownlee, G.G. (1981) *Nature* 292, 72-75.
 15. Sleight, M.J., Both, G.W., Brownlee, G.G., Bender, V.J. and Moss, B.A. (1980) in "Structure and Variation in Influenza Virus" (G. Laver and G. Air, eds.), pp. 69-78, Elsevier/North Holland, New York.
 16. Both, G.W. and Sleight, M.J. (1980) *Nucleic Acids Res.* 8, 2561-2575.
 17. Min Jou, W., Verhoeven, M., Devos, R., Saman, E., Fang, R., Huylebroeck, D., Fiers, W., Threlfall, G., Barber, C., Carey, N. and Emtage, S. (1980) *Cell* 19, 683-696.
 18. Porter, A.G., Barber, C., Carey, N.H., Hallewell, R.A., Threlfall, G. and Emtage, J.S. (1979) *Nature* 282, 471-477.
 19. Hiti, A.L., Davis, A.R. and Nayak, D.P. (1981) in "The Replication of Negative Strand Viruses" (D.H.L. Bishop and R.W. Compans, eds.), pp. 217-223, Elsevier/North Holland, New York.
 20. Winter, G. and Fields, S. (1982) *Cell*, in press.
 21. Skehel, J.J. and Hay, A.J. (1978) *Nucleic Acids Res.* 5, 1207-1219.
 22. Robertson, J.S. (1979) *Nucleic Acids Res.* 6, 3745-3757.
 23. Seed, B., personal communication.
 24. Grunstein, M. and Hogness, D.S. (1975) *Proc.Natl.Acad.Sci.USA* 72, 3961-3965.
 25. Maxam, A.M. and Gilbert, W. (1980) *Methods Enzymol.* 65, 499-560.
 26. Hill, D.F. and Petersen, G.B., personal communication.
 27. Sleight, M.J., Both, G.W. and Brownlee, G.G. (1979) *Nucleic Acids Res.* 7, 879-893.
 28. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) *Proc.Natl.Acad.Sci.USA* 74, 5463-5467.
 29. Messing, J., Crea, R. and Seeburg, P.H. (1981) *Nucleic Acids Res.* 9, 309-321.
 30. Sleight, M.J., Both, G.W. and Brownlee, G.G. (1979) *Nucleic Acids Res.* 6, 1309-1321.
 31. Inglis, S.C., Carroll, A.R., Lamb, R.A. and Mahy, B.W.J. (1976) *Virology* 74, 489-503.
 32. Inglis, S.C., McGeoch, D.J. and Mahy, B.W.J. (1977) *Virology* 78, 522-536.
 33. Inglis, S.C., Barrett, T., Brown, C.M. and Almond, J.W. (1979) *Proc. Natl.Acad.Sci.USA* 76, 3790-3794.
 34. Almond, J.W. and Barry, R.D. (1979) *Virology* 92, 407-415.
 35. Horisberger, M.A. (1980) *Virology* 107, 302-305.