

Molecular characterisation of the 3'-end of the astrovirus genome

C. Monceyron, B. Grinde, and T. Ø. Jonassen

Department of Virology, National Institute of Public Health, Oslo, Norway

Accepted October 21, 1996

Summary. We have sequenced the 3'-end of the RNA genomes of 14 serotyped and 12 untyped isolates of human astrovirus. The sequences, which include all 8 serotypes, were used to predict secondary structures, postulate possible functional domains, reveal conserved regions suitable for nucleic acid amplification and perform phylogenetic analysis. The final nucleotides of the capsid protein precursor gene and the adjacent 3'-noncoding region were highly conserved and, except for 35 nucleotides with homology to a sequence in the 3'-end of a coronavirus RNA genome, unique to astrovirus family. This confirms that the 3'-end is a suitable target for universal and specific detection of astrovirus RNA. For the deduced 72 C-terminal amino acids of the capsid protein precursor, distances between the serotypes were found to vary from 0.1 substitution per site between serotypes 3 and 7 to more than one substitution per site between serotype 4 and the other serotypes. Different isolates of the same serotype were closely related, which indicates that the presently used type-specific antibodies differentiate between phylogenetically distinct groups. RNA secondary structures with minimal free energy were predicted using computer programs. Comparative sequence analysis verified the significance of certain of the predicted structural elements.

Introduction

Astroviruses are often observed in faeces in connection with juvenile diarrhoea [9, 11]. The 6.8 kb plus-strand RNA genome is organised in a 5'-non-coding region (NCR), three open reading frames, eighty nucleotides (nt) 3'-NCR, and a poly(A) tail of approximately 30 nt [4]. The capsid protein precursor gene (ORF 2) is located next to the 3'-NCR. The capsid protein precursor is translated from a subgenomic mRNA [15].

Complete nucleotide sequences are available for HAsV-1 [18] and for HAsV-2 [4]. Comparison of these two sequences shows that the 3'-NCR are 94% conserved, and that the 114 amino acids deduced from the 3' end of ORF

2 are 62% identical [14]. Limited nucleotide sequence information for the six other serotypes are available in the nucleotide databases [10, 16].

The 3'-end of RNA virus genomes are involved in regulating the initiation of synthesis of the minus-strand RNA. This region may also participate in the regulation of plus-strand synthesis, RNA stability, translation initiation and intracellular localisation. The aims of the present study were to predict secondary structures, postulate possible functional domains, reveal conserved regions suitable for nucleic acid amplification, and perform phylogenetic analysis by comparing the 3'-ends of the different human astrovirus serotypes.

Materials and methods

Astrovirus RNA was made available by dialysing and heating faecal extracts or CaCo-2 cell culture supernatants [5]. The templates for sequencing were made by reverse transcription of astrovirus RNA with an oligo d(T) primer, and subsequent PCR on the cDNA with the oligo d(T) primer and a 5' biotinylated upstream primer (5'-AAC CGT GTA ACC CTC (C/T) T, position 6 476–6 492). The nucleotide numbering used throughout this paper is adopted from the HAsV-2 sequence [4]. The PCR-product was captured on magnetic beads coated with streptavidin (Dynal, Norway), and the minus-strand eluted with NaOH according to the manufacturers' recommendations. After neutralisation the minus-strand was sequenced with a primer (5'-AC CGT GTA ACC CTC CTC TC) extending three nt downstream of the upstream PCR primer. The above PCR was not successful for HAsV-4. Another upstream PCR-primer (5'-GAG AGT A(A/C)(A/C/T) GAG G(A/G)(C/T) GAA, position 6 389–6 406) was used for HAsV-4 and untyped samples. A primer (5'-T GAT GAT GAA GACTCT AAT AC) extending eleven nt downstream of the 6 389–6 406 primer was used for sequencing the minus-strand of the samples suspected to be HAsV-4. The sequencing of the plus-strand attached to the beads was initiated with the primer 5'-TTTTTTTTTTTTTTTTTGC (position 6 815–6 796). Sequencing was done using the Sequenase 2.0 kit (United States Biochemical, USA), with [³⁵S]dATP incorporation, according to the manufacturer's recommendations.

The following list indicates the number of novel and previously published (GenBank) sequences within each serotype used in the present analyses: HAsV-1: 3 novel samples, 2 published; HAsV-2: 3 novel samples, 1 published; HAsV-3: 3 novel samples; HAsV-4: 2 novel samples, 1 published; HAsV-5: 1 novel sample; HAsV-6: 1 novel sample, 1 published; HAsV-7: 1 novel sample; HAsV-8: 1 in GenBank. The novel samples were from independent outbreaks of diarrhoea, collected in Norway and the UK. In addition cDNA from 12 untyped astroviruses from all over the world were sequenced. The Wisconsin sequence analysis package 8.0 (Genetics Computer Group, Wisconsin, USA) was used for sequence analyses. In addition the PHYLIP 3.5 software package (J. Felsenstein, Department of Genetics, University of Washington, Seattle, USA) was used for phylogenetic analyses.

Results and discussion

Alignment of one cDNA sequence of each serotype is shown in Fig. 1. The 3'-NCRs and the terminal 19 nt of the ORF2, were highly conserved between all the serotypes. In the rest of the sequenced part of the ORF2, however, there were substantial nucleotide and amino acid differences. This suggests a selective pressure on the level of RNA in the extreme 3'end of ORF2. In the rest of the

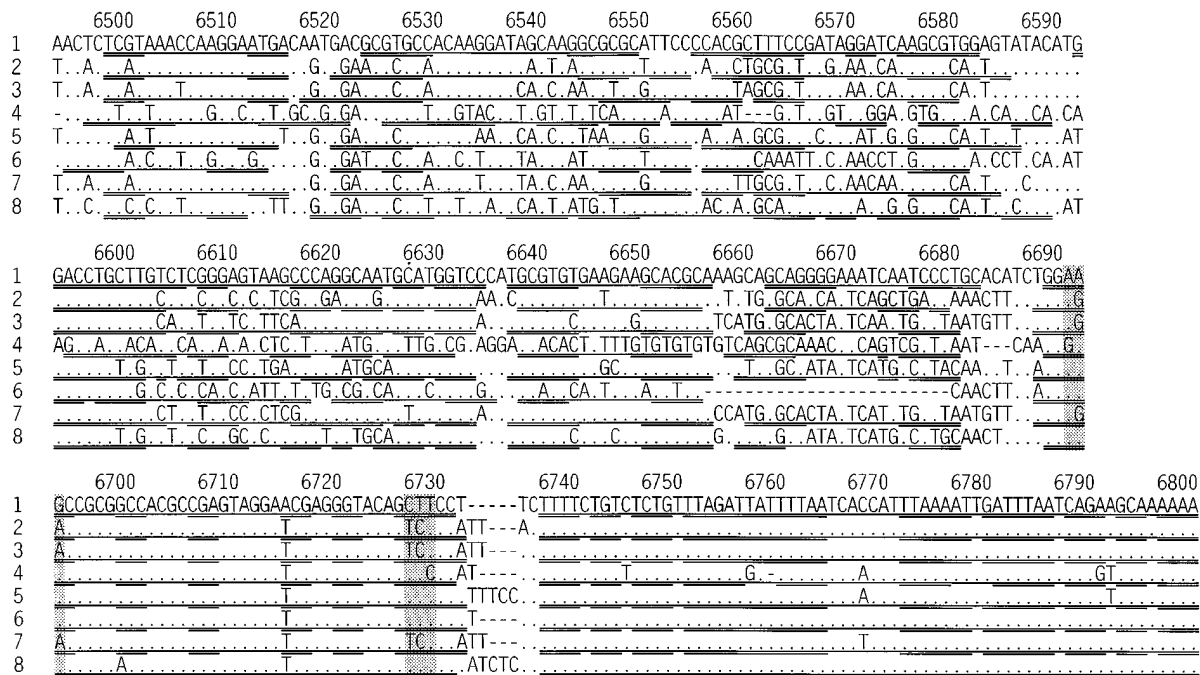


Fig. 1. Alignment of the 3'-ends of astrovirus genomes. The numbers above the sequences indicate the nucleotide numbers of HAstV-2. The numbers to the left of the sequences indicate serotype. The predicted loops and bulges are shown underlined, and the predicted stems are indicated by a double line under the base paired nucleotides (including G-U base pairs). The predicted covariations are shaded. The ORF 2 stop codon is in positions 6713–6715

sequenced region of the capsid protein precursor, most of the amino acids were not conserved. However, Kimura corrected ratios of synonymous to nonsynonymous nucleotide substitutions were above two, which suggests some conservation at the amino acid level.

The distances matrices in Table 1 indicate that HAstV-3 and 7 are closely related, while the C-terminal end of the capsid protein of HAstV-4 is very different from the other serotypes. All the untyped astroviruses had sequences related to one of the known serotypes. Eight of these isolates clustered with HAstV-1, two with HAstV-2 and two, one from Ghaza and one from Uganda, with HAstV-8. These untyped isolates were included in the calculation of the intratype differences shown in Table 1. The low intra-type RNA sequence variability and the distinct differences between the serotypes indicate that the serotypes based on reactivity to antibodies are phylogenetically distinct groups, as previously suggested [16]. However, the difference between HAstV-3 and 7 is only slightly higher than the differences within HAstV-1 which is the most prevalent serotype [11].

Table 1. Matrix of distances within and between astrovirus serotypes^a

Serotype	1	2	3	4	5	6	7	8
1	5	29	28	57	28	31	29	28
2	49	3	14	61	30	29	14	31
3	36	21	0	58	26	30	7	27
4	123	118	123	3	58	54	58	64
5	49	45	40	118	2	29	25	15
6	66	55	55	110	57	nd	30	30
7	42	29	10	123	42	55	nd	30
8	49	51	42	118	23	60	47	4

^a Correction method: Jukes-Cantor without gap penalty. Above diagonal: Mean number of substitutions per 100 nt in the 3' 0.3 kb of the HAstV RNA (not including the poly (A) tail). On the diagonal: Number of substitutions per 100 nt in the 3' 0.3 kb between the most divergent isolates within each HAstV serotype, including the untyped isolates in the same cluster. It was not determined for types 6 and 7, because more than one isolate has not been available. Below diagonal: Number of substitutions per 100 amino acids in the C-terminal 72 amino acids of the capsid protein precursor

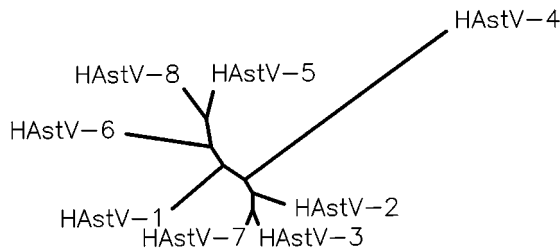


Fig. 2. Dendrogram of the maximum likelihood analysis of the 3' 0.3 kb of the astrovirus genomic RNA. The transversion:transition ratio was set to 2

Phylogenetic analyses, using neighbour joining, maximum parsimony and maximum likelihood, indicated a clade consisting of HAstV-2, 3 and 7 (Fig. 2). This clade was supported (90 %) by bootstrapping tests.

Major et al. [12] could not find any significant homology between the 3'-end of the HAstV-1 genome and other virus, while Willcocks and Carter [19] found a 17% identity between the 3'-NCR of HAstV-1 and certain enteroviruses. Searching data bases of virus nucleic acids, we found a sequence in the avian infectious bronchitis virus (IBV) RNA [1] with 90% identity to the first 35 nt of the conserved part of the 3'-end of the HAstV genome (position 6 696–6 730). The 35 nt IBV sequence is in a conserved part of the nucleocapsid gene about 100 nt from the poly(A) tail. The two sequences are not in the same reading frame in IBV and HAstV, and the sequence is not found in other coronaviruses. This makes it unlikely that the similarities are caused by convergent evolution. Statistically, a 90% similarity to a 35 nt sequence is expected to be found once in a sequence of

10^{17} nt. The virus sequence database we used contained 3×10^7 nt, and the sequence searched for was 300 nt, implying a 10^{-7} probability for the observed similarity to be incidental.

The 35 nt are included in the stem-loop II (Fig. 3), and in a similar structure in IBV. It has been suggested that transcription pauses caused by strong secondary structures can induce recombination [8]. Together with the high frequency of recombination reported for coronaviruses [7, 13], the data suggest the homology to be due to a recombination.

Although the regulation of initiation of plus- and minus-strand RNA replication probably differs, some common initiation signals may be located at the 3'-ends of the opposite RNA strands [17]. When we compared the 3'- and 5'-NCRs, a sequence complementary to an eight nucleotide motif (position 6 766–6 773) was found in the conserved part of the 5'-NCR (position 13–20). The 3'-NCR thus shares this motif with the 3'-end of the minus-strand RNA. The octamer sequence in the 3'-NCR is conserved, except for position 6 771 which differ from the consensus in HAstV-4, 5 and 7.

Our MFOLD predictions of secondary structures with minimal free energy gave eight stem-loops for most of the serotypes (Figs. 1 and 3). As shown for HAstV-2 in Fig. 3, we have extended the Roman numerals numbering of the stem-loops starting from the 3'-end as previously suggested [14].

In addition to the free energy of the predicted RNA folding, nucleation centres, direction of synthesis, long-range interactions, and extragenomic factors influence the secondary structure. These elements are difficult to analyse, but further evidence for the significance of a secondary structure may be obtained by comparing the predicted folding of closely related RNA-genomes [3].

Stem-loops I and II are formed in the highly conserved part of the 3'-end. Two distinctly different RNA secondary structures, based on RNA sequences of the 3'-ends of the HAstV-1 [19] and HAstV-1 and -2 [14] genomes, have previously been predicted. The stem-loop I predicted by MFOLD, corresponded to the one suggested by Monroe et al. [14]. It included six nt of the poly(A) tail which is not included in the folding suggested by Willcocks and Carter [19]. In the stem there are four mismatched base pairs or bulges (Fig. 3). There is some nucleotide variation between the serotypes in these positions (Fig. 1), but the bulges are maintained. This suggests a biological function of the bulges. Similar bulges are, for example, shown to be recognition sites for the binding of Rev and Tat to HIV RNA [2].

The octamer, which has a counterpart in the 3'-end of the minus-strand, is situated in the loop of stem-loop I. The MFOLD program did not predict any similar secondary structure in the 3'-end of the minus-strand, however, the minus-strand octamer appeared in a single stranded region, suggesting that the putative signal is in the primary sequence.

It has been reported that in the hinge between stem-loop I and II, the HAstV-1 lacks two residues found in HAstV-2 [14]. We found that HAstV-3 and -7 are similar to HAstV-2, while HAstV-4 and -6 are one residue shorter, and HAstV-5 and HAstV-8 are three residues longer than HAstV-2. These differences

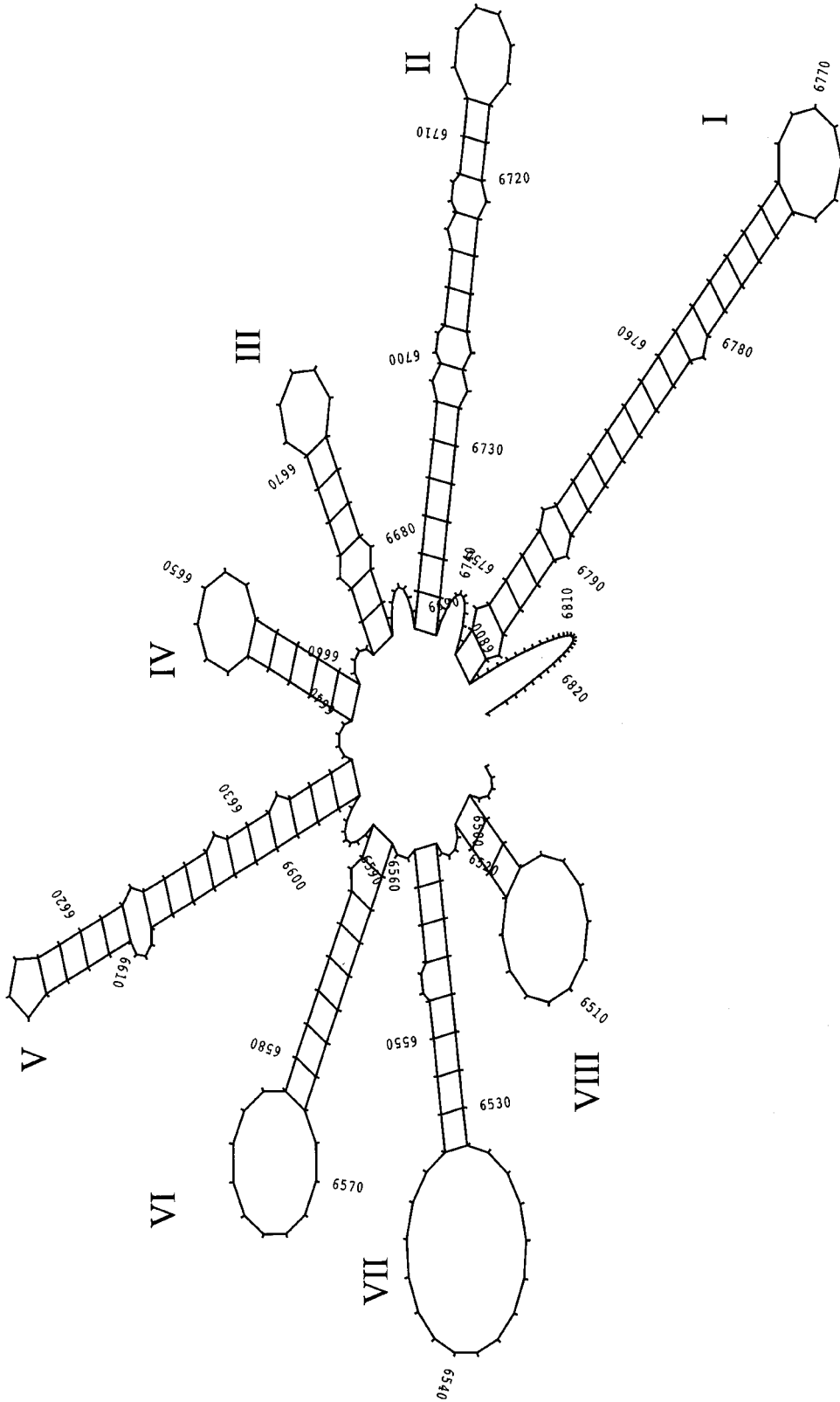


Fig. 3. Minimal free energy folding of the 3' 0.3 kb of the HAstV-2 RNA as predicted by the MFOLD program

are probably phylogenetically significant, and support the clade of serotypes 2, 3, and 7. The heterogeneity of this region is consistent with the suggested hinge structure.

The reported compensating nucleotide differences in the stem-loop II of HAstV-1 and 2 [14], are also found in the other serotypes (Fig. 1). HAstV-2, 3 and 7 have AGA in position 6 694–6 696 and a complementary TCT in position 6 730–6 732, while HAstV-4 has GAG and CTC, and the other serotypes AAG and CTT, giving an equally stable duplex. The compensating substitutions suggest that the stem is biologically important.

Nucleotide substitutions which require compensating substitutions to maintain a vital function, are less likely to have occurred several times during the evolution. The covariations therefore further support the phylogenetic relationship we have proposed between the HAstV-2, -3 and -7, as opposed to the phylogeny proposed by Noel et al. [16].

In stem-loops III—VIII the stems and loops of the various serotypes differ in length.

The sequences obtained in this work support our previous experimental observations that the 3'-NCR is a suitable target for universal and specific detection of astrovirus RNA [6]. The universal primers suggested in the former paper have positions 6 815–6 796 and 6 739–6 761. Our analyses also support the taxonomic classification of astroviruses as a separate family [14].

Acknowledgements

We would like to thank Dr. J. B. Kurtz (Public Health Laboratory, Oxford), Dr. C. A. Hart (Department of Medical Microbiology, University of Liverpool) and Prof. H. Ushijima (Institute of Public Health, Tokyo) for providing viral samples, and Dr. E. Kjeldsberg for serotyping samples collected in Norway.

References

1. Bourns MEG, Binns MM, Foulds IJ, Brown TDK (1985) Sequences of the nucleocapsid genes from two strains of avian infectious bronchitis virus. *J Gen Virol* 66: 573–580
2. Gait MJ, Karn J (1993) RNA recognition by the human immunodeficiency virus Tat and Rev proteins. *Trends Biochem Sci* 18: 255–259
3. James BD, Olsen GJ, Pace NR (1989) Phylogenetic comparative analysis of RNA secondary structure. *Methods Enzymol* 180: 227–239
4. Jiang B, Monroe SS, Koonin EV, Stine SE, Glass RI (1993) RNA sequence of astroviruses: distinctive genomic organization and a putative retrovirus-like ribosomal frameshifting signal that directs the viral replicase synthesis. *Proc Natl Acad Sci USA* 90: 10 539–10 543
5. Jonassen TØ, Kjeldsberg E, Grinde B (1993) Detection of human astrovirus serotype 1 by the polymerase chain reaction. *J Virol Methods* 44: 83–88
6. Jonassen TØ, Monceyron C, Lee TW, Kurtz JB, Grinde B (1995) Detection of all serotypes of human astrovirus by the polymerase chain reaction. *J Virol Methods* 52: 327–334
7. Kottier SA, Cavanagh D, Britton P (1995) Experimental evidence of recombination in coronavirus infectious bronchitis virus. *Virology* 213: 569–580

8. Lai MMC (1992) RNA recombination in animal and plant viruses. *Microbiol Rev* 56: 61–79
9. Lee TW, Kurtz JB (1994) Prevalence of human astrovirus serotypes in the Oxford region 1976–92: with evidence for two new serotypes. *Epidemiol Infect* 112: 187–193
10. Lewis TL, Greenberg HB, Herrmann JE, Smith LS, Matsui SM (1994) Analysis of astrovirus serotype 1 RNA, identification of the viral RNA-dependent RNA polymerase motif, and expression of a viral structural protein. *J Virol* 68: 77–83
11. Madeley CR, Cosgrove BP (1975) 28 nm particles in faeces in infantile gastroenteritis. *Lancet* ii: 451–452
12. Major ME, Eglin RP, Easton AJ (1992) 3' terminal nucleotide sequence of human astrovirus type 1 and routine detection of astrovirus nucleic acid and antigens. *J Virol Methods* 39: 217–225
13. Makino S, Keck JG, Stohlman SA, Lai MM (1986) High-frequency RNA recombination of murine coronaviruses. *J Virol* 57: 729–737
14. Monroe SS, Jiang B, Stine SE, Koopmans M, Glass RI (1993) Subgenomic RNA sequence of human astrovirus supports classification of astroviridae as a new family of RNA viruses. *J Virol* 67: 3 611–3 614
15. Monroe SS, Stine SE, Gorelkin L, Herrmann JE, Blacklow NR, Glass RI (1991) Temporal synthesis of proteins and RNAs during human astrovirus infection of cultured cells. *J Virol* 65: 641–648
16. Noel JS, Lee TW, Kurtz JB, Glass RI, Monroe SS (1995) Typing of human astroviruses from clinical isolates by enzyme immunoassay and nucleotide sequencing. *J Clin Microbiol* 33: 797–801
17. Pilipenko EV, Maslova SA, Sinyakov AN, Agol VI (1992) Towards identification of cis-acting elements involved in the replication of enterovirus and rhinovirus RNAs: a proposal for the existence of tRNA-like terminal structures. *Nucleic Acids Res* 20: 1 739–1 745
18. Willcocks MM, Brown TDK, Madeley CR, Carter MJ (1994) The complete sequence of a human astrovirus. *J Gen Virol* 75: 1 785–1 788
19. Willcocks MM, Carter MJ (1992) The 3' terminal sequence of a human astrovirus. *Arch Virol* 124: 279–289

Authors' address: Dr. T. Ø. Jonassen, Department of Virology, National Institute of Public Health, P. O. Box 4404, Torshov, N-0403 Oslo, Norway.

Received March 15, 1996