Proc. Natl. Acad. Sci. USA
Vol. 82, pp. 2627–2631, May 1985
Biochemistry

Primary structure and gene organization of human hepatitis A virus

(human picornavirus/structural analysis/nucleotide sequence/sequence homology/polyprotein)

RICHARD NAJARIAN*, DANIEL CAPUT*, WENDY GEE*, STEVEN J. POTTER*, ANDRE RENARD†, JAMES MERRYWEATHER*, GARY VAN NEST*, AND DINO DINA*

*Chiron Research Laboratories, Chiron Corporation, 4560 Horton Street, Emeryville, CA 94608; and †Laboratoire de Genie Genetique, Universite de Liege, 4000 Liege, Belgium

Communicated by Frank Lilly, December 26, 1984

ABSTRACT **The RNA genome of human hepatitis A virus (HAV) was molecularly cloned. Recombinant DNA clones representing the entire HAV RNA were used to determine the primary structure of the viral genome. The length of the viral genome is 7478 nucleotides. An open reading frame starting at nucleotide 734 and terminating at nucleotide 7415 encodes a polyprotein of M_r 251,940. Comparison of the HAV nucleotide sequence with that of other picornaviruses has failed to reveal detectable areas of homology. However, a computer analysis of the putative amino acid sequence of HAV and poliovirus demonstrated the existence of short areas of homology in virion protein 3 (VP3) and throughout the carboxyl-terminal portion of the polyproteins. In addition, extensive protein structural homologies with poliovirus were detected.**

Hepatitis A virus is a picornavirus belonging to the enterovirus genus. The viral RNA is 7.5 kilobases long, is of (+)-strand polarity, contains a poly(A) stretch at the 3′ end, and may be covalently linked to a protein at the 5′ terminus designated VPg (1–3). Evidence for virion proteins of M_rs 33,000, 29,000, and 27,000, designated VP1, VP2, and VP3, respectively, and additional peptides of M_rs 22,000 and 10,000 has been presented (4–6).

Two recent reports have described the molecular cloning of HAV-specific sequences (3, 7). We report here the molecular cloning and the nucleotide sequence of the entire HAV genome. The deduced amino acid sequence has been analyzed and compared to that of poliovirus. The sequence and structural homologies between these two human picornaviruses are discussed.

MATERIALS AND METHODS

Virus Isolation and Propagation. Several groups have reported the successful isolation and growth in tissue culture of HAV derived from both infected marmoset livers and human fecal samples (8–12). We have established our own isolate from the stools of a hepatitis A patient collected during an epidemic outbreak in Los Angeles (kindly provided by J. Rakela) according to published procedures (12). Viral RNA was extracted from virions by published procedures that included treatment with protease and extraction with phenol/chloroform followed by precipitation with ethanol (3).

Molecular Cloning. Details of the procedures and specific approaches used to obtain the entire HAV genome will be reported elsewhere. Briefly, 3′-proximal clones were obtained by screening recombinant cDNA libraries with the probes containing the following sequences (3):

HAV1 3′ A-C-A-A-A-T-A-A-A-G-A-A-A-A-T-A-G-T-C-A-T-T-T-A 5′

HAV2 3′ A-T-G-T-C-T-G-A-A-T-T-T-A-G-A-A-T-A-C-T-A-A-C-C-A-C-C 5′.

Further portions of the genome were cloned by "walking" techniques and by using the 5′-terminal *Pst* I–*Pst* I fragment derived from clone pHAV1307 (kindly provided to us by J. Ticehurst, National Institutes of Health).

RESULTS AND DISCUSSION

Analysis of cDNA Clones. A set of four recombinant cDNA clones representing the whole HAV genome was selected for nucleotide sequencing studies. The four clones, termed pHAV16, pHAV1, pHAV8, and pHAV47, were mapped with restriction endonucleases, and the overlapping regions were confirmed by cross-hybridization studies (Fig. 1). The total length of the nonoverlapping portions of these clones added up to approximately 7.5 kilobases, the reported length of the HAV genome (3).

The nucleotide sequence of the entire HAV genome derived from the four cDNA clones is shown in Fig. 2. According to this sequence, the HAV genome is 7478 nucleotides long. The heteropolymeric sequence is followed by a poly(A) tract of undetermined length. A single open reading frame starts with an AUG triplet at nucleotide 734 and terminates with an UGA codon at nucleotide 7415. The polyprotein thus encoded is 2227 amino acids long, has a M_r of 251,940, and is divided into P1, P2, and P3 regions (13).

The sequence at the 3′ end of the genome was determined from independent clones carrying the poly(A) 3′-terminal sequence. The sequence at the 5′ end of the viral RNA was derived from five independent clones that carried a terminal *Pst* I–*Pst* I fragment of approximately 180 nucleotides. Of these five, three were found to be identical and contained the 5′-terminal sequence shown in Fig. 2.

HAV Map. The genome of the independent HAV isolate cloned in our laboratory has a restriction map substantially different from that of the isolate cloned by Ticehurst *et al.* (3). The observed differences may be due to single nucleotide changes that do not substantially alter the primary amino acid sequence and antigenic properties of the viral proteins. This notion is supported by a comparison of the 3′-terminal 500 nucleotides of our sequence with the homologous sequence reported by Ticehurst *et al.* (3). Only a single amino acid difference at position 2159 (methionine for isoleucine) is detected in this portion of the genome. In the same region there are 24 nucleotide substitutions that do not alter the amino acid sequence. Additional changes are present in the 3′ noncoding region of the viral RNA. We conclude that

Abbreviation: HAV, hepatitis A virus.

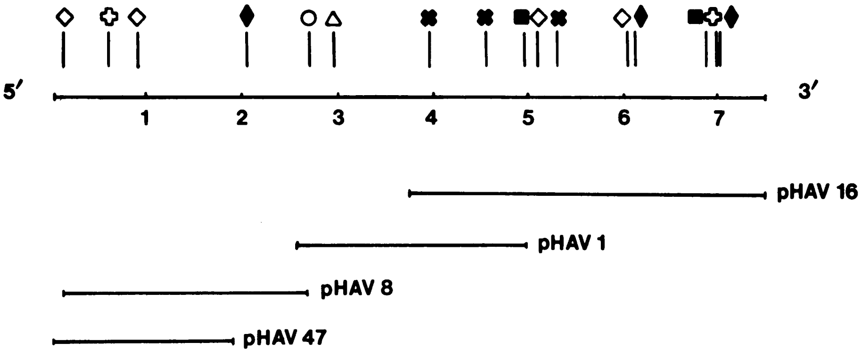

FIG. 1. Restriction map of HAV clones. (*Upper*) A composite map of the HAV cDNA clones. Numbers represent kilobases from the 5′ end of the genome. (*Lower*) Individual clones used to determine this map and their positions. The end positions of each clone and the extent of overlap with neighboring clones were determined by nucleotide sequencing. ◇, *Pst* I; ✥, *Bam*HI; ◆, *Hin*dIII; △, *Sac* I; ✖, *Xba* I; ■, *Eco*RI; ○, *Sal* I.

nucleotide sequence polymorphisms exist among independent isolates, most of which are not reflected in changes of the primary amino acid sequence.

The 5′ End. The sequence found at the 5′ terminus of the HAV genome has several of the properties described for other picornaviruses and in particular for poliovirus (14, 15). The 5′-terminal dinucleotide is U-U, as required for the presumed linkage to a tyrosine residue on VPg (16, 17). The first 75 nucleotides of the sequence can be arranged in a stable secondary structure comprising stems and loops. In addition, the 5′-terminal portion of the sequence displays long pyrimidine tracts (U+C repeats). Both features have been observed

```
TTCAAGAGGGGTCTCCGGAGGTTTCCGGAGCCCCTCTTGGAAGTCCATGGTGAGGGGACTTGATACCTCACCGCCGTTTGCCTAGGCTATAGGCTAAATTTCCCTTTCCCTGTCCCTCCCTTATTTCCCTTTGTTTTGCTTGTAAAT (147)
ATTAATTCCTGCAGGTTCAGGGTTCTTTAATCTGTTTCTCTATAAGAACACTCAATTTTCACGCTTTCTGTCTTCTTTCTTCCAGGGCTCTCCCCTTGCCCTAGGCTCTGGCCGTTGCGCCCGGCGGGGTCAACTCCATGATTAGCA (294)
TGGAGCTGTAGGAGTCTAAATTGGGGACGCAGATGTTTGGGACGTCACCTTGCAGTGTTAACTTGGCTCTCATGAACCTCTTTGATCTTCCACAAGGGGTAGGCTACGGGTGAAACCTCTTAGGCTAATACTTCTATGAAGAGATGC (441)
TTTGGATAGGGTAACAGCGGCGGATATTGGTGAGTTGTTAAGACAAAAACCATTCAACGCCGGAGGACTGGCTCTCATCCAGTGGATGCATTGAGTGGATTGATTGTCAGGGCTGTCTCTAGGTTTAATCTCAGACCTCTCTGTGCT (588)
TAGGGCAAACACCATTTGGCCTTAAATGGGATCCTGTGAGAGGGGGTCCCTCCATTGACAGCTGGACTGTTCTTTGGGGCCTTATGTGGTGTTTGCCTCTGAGGTACTCAGGGGCATTTAGGTTTTTCCTCATTCTTAAACAATA   (733)
P1,1A
1                                   10                                      20                                      30
Met Asn Met Ser Lys Gln Gly Ile Phe Gln Thr Val Gly Ser Gly Leu Asp His Ile Leu Ser Leu Ala Asp Ile Glu Glu Glu Gln Met Ile Gln Ser Val Asp Arg Thr
ATG AAT ATG TCC AAA CAA GGA ATT TTC CAG ACT GTT GGG AGT GGC CTT GAC CAC ATC CTG TCT TTG GCA GAT ATT GAG GAA GAG CAA ATG ATT CAG TCC GTT GAT AGG ACT (844)
        40                                      50                                      60                                      70
Ala Val Thr Gly Ala Ser Tyr Phe Thr Ser Val Asp Gln Ser Ser Val His Thr Ala Glu Val Gly Ser His Gln Ile Glu Pro Leu Lys Thr Ser Val Asp Lys Pro Gly
GCA GTG ACT GGA GCT TCT TAC TTC ACT TCT GTG GAC CAA TCT TCA GTT CAT ACT GCT GAG GTT GGC TCA CAT CAA ATT GAA CCT TTG AAA ACC TCT GTT GAT AAA CCT GGT (955)
                    80                                      90                                      100                                     110
Ser Lys Lys Thr Gln Gly Glu Lys Phe Phe Leu Ile His Ser Ala Asp Trp Leu Thr Thr His Ala Leu Phe His Glu Val Ala Lys Leu Asp Val Val Lys Leu Leu Tyr
TCT AAG AAA ACT CAG GGG GAA AAG TTT TTC CTG ATT CAT TCT GCT GAT TGG CTC ACT ACA CAT GCT CTC TTT CAT GAA GTT GCA AAA TTG GAT GTG GTG AAA CTA CTG TAT (1066)
                                120                                     130                                     140
Asn Glu Gln Phe Ala Val Gln Gly Leu Leu Arg Tyr His Thr Tyr Ala Arg Phe Gly Ile Glu Ile Gln Val Gln Ile Asn Pro Thr Pro Phe Gln Gln Gly Gly Leu Ile
AAT GAG CAG TTT GCC GTC CAA GGT TTG TTG AGA TAC CAT ACA TAT GCA AGA TTT GGC ATT GAG ATT CAA GTT CAG ATA AAT CCC ACA CCC TTT CAG CAA GGA GGA CTA ATT (1177)
    150                                     160                                     170                                     180
Cys Ala Met Val Pro Gly Asp Gln Ser Tyr Gly Ser Ile Ala Ser Leu Thr Val Tyr Pro His Gly Leu Leu Asn Cys Asn Ile Asn Asn Val Val Arg Ile Lys Val Pro
TGT GCC ATG GTT CCT GGT GAC CAA AGT TAT GGT TCA ATA GCA TCC TTG ACT GTT TAT CCT CAT GGT CTG TTA AAT TGC AAT ATC AAC AAT GTA GTT AGA ATA AAG GTT CCA (1288)
                190                                     200                                     210                                     220
Phe Ile Tyr Thr Arg Gly Ala Tyr His Phe Lys Asp Pro Gln Tyr Pro Val Trp Glu Leu Thr Ile Arg Val Trp Ser Glu Leu Asn Ile Gly Thr Gly Thr Ser Ala Tyr
TTT ATT TAT ACT AGA GGT GCT TAT CAT TTT AAA GAT CCA CAG TAC CCA GTT TGG GAA TTG ACA ATC AGA GTT TGG TCA GAG TTG AAT ATT GGA ACA GGA ACT TCA GCT TAC (1399)
                                                                                            1B
                            230                                     240                                     250
Thr Ser Leu Asn Val Leu Ala Arg Phe Thr Asp Leu Glu Leu His Gly Leu Thr Pro Leu Ser Thr Gln Met Met Arg Asn Glu Phe Arg Val Ser Thr Thr Glu Asn Val
ACT TCA CTC AAT GTT TTA GCT AGG TTT ACA GAT TTG GAG TTG CAT GGA TTA ACT CCT CTT TCT ACA CAG ATG ATG AGA AAT GAA TTT AGG GTC AGT ACT ACT GAA AAT GTT (1510)
260                                     270                                     280                                     290
Val Asn Leu Ser Asn Tyr Glu Asp Ala Arg Ala Lys Met Ser Phe Ala Leu Asp Gln Glu Asp Trp Lys Ser Asp Pro Ser Gln Gly Gly Gly Ile Lys Ile Thr His Phe
GTA AAT TTG TCA AAT TAT GAA GAT GCA AGG GCA AAA ATG TCT TTT GCT TTG GAT CAG GAA GAT TGG AAG TCT GAT CCT TCC CAA GGT GGT GGA ATT AAA ATT ACT CAT TTT (1621)
            300                                     310                                     320                                     330
Thr Thr Trp Thr Ser Ile Pro Thr Leu Ala Ala Gln Phe Pro Phe Asn Ala Ser Asp Ser Val Gly Gln Gln Ile Lys Val Ile Pro Val Asp Pro Tyr Phe Phe Gln Met
ACT ACC TGG ACA TCC ATT CCA ACC TTA GCT GCT CAG TTT CCA TTT AAT GCT TCA GAT TCA GTT GGA CAA CAA ATT AAA GTT ATT CCA GTG GAC CCA TAC TTT TTC CAA ATG (1732)
                        340                                     350                                     360                                     370
Thr Asn Thr Asn Pro Asp Gln Lys Cys Ile Thr Ala Leu Ala Ser Ile Cys Gln Met Phe Cys Phe Trp Arg Gly Asp Leu Val Phe Asp Phe Gln Val Phe Pro Thr Lys
ACA AAC ACT AAT CCT GAT CAA AAA TGT ATA ACT GCC TTG GCC TCT ATT TGT CAG ATG TTC TGC TTT TGG AGG GGA GAT CTT GTT TTT GAT TTT CAG GTT TTT CCA ACC AAA (1843)
                                    380                                     390                                     400
Tyr His Ser Gly Arg Leu Leu Phe Cys Phe Val Pro Gly Asn Glu Leu Ile Asp Val Thr Gly Ile Thr Leu Lys Gln Ala Thr Thr Ala Pro Cys Ala Val Met Asp Ile
TAT CAT TCA GGT AGA CTG TTG TTT TGT TTT GTT CCT GGG AAT GAG TTA ATA GAT GTT ACT GGA ATT ACA TTA AAA CAG GCA ACT ACT GCT CCT TGT GCA GTG ATG GAC ATT (1954)
        410                                     420                                     430                                     440
Thr Gly Val Gln Ser Thr Leu Arg Phe Arg Val Pro Trp Ile Ser Asp Thr Pro Tyr Arg Val Asn Arg Tyr Thr Lys Ser Ala His Gln Lys Gly Glu Tyr Thr Ala Ile
ACA GGA GTG CAG TCA ACC TTG AGA TTT CGT GTT CCT TGG ATT TCT GAT ACA CCT TAT CGA GTG AAT AGG TAC ACG AAG TCA GCA CAT CAA AAA GGT GAG TAC ACT GCC ATT (2065)
                    450                                     460                                     470                                     480
Gly Lys Leu Ile Val Tyr Cys Tyr Asn Arg Leu Thr Ser Pro Ser Asn Val Ala Ser His Val Arg Val Asn Val Tyr Leu Ser Ala Ile Asn Leu Glu Cys Phe Ala Pro
GGG AAG CTT ATT GTG TAT TGT TAT AAC AGA CTG ACT TCT CCT TCT AAT GTT GCC TCT CAT GTT AGA GTT AAT GTT TAT CTT TCA GCA ATT AAT TTG GAA TGT TTT GCT CCT (2176)
                                        1C
                                490                                     500                                     510
Leu Tyr His Ala Met Asp Val Thr Thr Gln Val Gly Asp Asp Ser Gly Gly Phe Ser Thr Thr Val Ser Thr Glu Gln Asn Val Pro Asp Pro Gln Val Gly Ile Thr Thr
CTT TAC CAT GCT ATG GAT GTT ACT ACA CAG GTT GGA GAT GAT TCA GGA GGT TTC TCA ACA ACA GTT TCT ACA GAG CAG AAT GTT CCT GAT CCC CAA GTT GGG ATA ACA ACC (2287)
    520                                     530                                     540                                     550
Met Arg Asp Leu Lys Gly Lys Ala Asn Arg Gly Lys Met Asp Val Ser Gly Val Gln Ala Pro Arg Gly Ser Tyr Gln Gln Gln Leu Asn Asp Pro Val Leu Ala Lys Lys
ATG AGG GAT TTA AAA GGA AAA GCC AAT AGG GGA AAG ATG GAT GTT TCA GGA GTG CAA GCA CCT CGT GGG AGC TAT CAG CAA CAA TTG AAC GAT CCA GTT TTA GCA AAG AAA (2398)
                560                                     570                                     580                                     590
Val Pro Glu Thr Phe Pro Glu Leu Lys Pro Gly Glu Ser Arg His Thr Ser Asp His Met Ser Ile Tyr Lys Phe Met Gly Arg Ser His Phe Leu Cys Thr Phe Thr Phe
GTA CCT GAG ACA TTT CCT GAA TTG AAG CCT GGA GAG TCC AGA CAT ACA TCA GAT CAC ATG TCT ATT TAT AAA TTC ATG GGA AGG TCT CAT TTT TTG TGC ACT TTT ACT TTC (2509)
                            600                                     610                                     620
Asn Ser Asn Asn Lys Glu Tyr Thr Phe Pro Ile Thr Leu Ser Ser Thr Ser Asn Pro Pro His Gly Leu Pro Ser Thr Leu Arg Trp Phe Phe Asn Leu Phe Gln Leu Tyr
AAT TCA AAT AAT AAA GAG TAC ACA TTT CCA ATA ACC CTG TCT TCG ACT TCT AAT CCT CCT CAT GGT TTA CCA TCA ACA TTA AGG TGG TTC TTC AAT TTG TTT CAG TTG TAT (2620)
630                                     640                                     650                                     660
Arg Gly Pro Leu Asp Leu Thr Ile Ile Ile Thr Gly Ala Thr Asp Val Asp Gly Met Ala Trp Phe Thr Pro Val Gly Leu Ala Val Asp Pro Trp Val Glu Lys Glu Ser
AGA GGA CCA TTG GAT TTA ACA ATT ATA ATC ACA GGA GCC ACT GAT GTG GAT GGT ATG GCC TGG TTT ACT CCA GTG GGC CTT GCT GTC GAC CCT TGG GTG GAA AAG GAG TCA (2731)
            670                                     680                                     690                                     700
Ala Leu Ser Ile Asp Tyr Lys Thr Ala Leu Gly Ala Val Arg Phe Asn Thr Arg Arg Thr Gly Asn Ile Gln Ile Arg Leu Pro Trp Tyr Ser Tyr Leu Tyr Ala Val Ser
GCT TTG TCT ATT GAT TAT AAA ACT GCC CTT GGA GCT GTT AGA TTT AAT ACA AGA AGA ACA GGA AAC ATT CAA ATT AGA TTG CCG TGG TAT TCT TAT TTG TAT GCC GTG TCT (2842)
                        710                                     720                                     730                                     740
Gly Ala Leu Asp Gly Leu Gly Asp Lys Thr Asp Ser Thr Phe Gly Leu Phe Leu Phe Glu Ile Ala Asn Tyr Asn His Ser Asp Glu Tyr Leu Ser Phe Ser Cys Tyr Leu
GGA GCA CTG GAT GGC TTG GGG GAT AAG ACA GAT TCT ACA TTT GGA TTG TTT CTA TTC GAG ATT GCA AAT TAC AAT CAT TCT GAT GAA TAT TTG TCC TTC AGT TGT TAT TTG (2953)
                                    750                                     760                                     770
Ser Val Thr Glu Gln Ser Glu Phe Tyr Phe Pro Arg Ala Pro Leu Asn Ser Asn Ala Met Leu Ser Thr Glu Ser Met Met Ser Arg Ile Ala Ala Gly Asp Leu Glu Ser
TCT GTC ACA GAG CAA TCA GAG TTC TAT TTT CCT AGA GCT CCA TTA AAT TCA AAT GCT ATG TTG TCC ACT GAA TCC ATG ATG AGT AGA ATT GCA GCT GGA GAC TTG GAG TCA (3064)
        780                                     790                                     800                                     810
Ser Val Asp Asp Pro Arg Ser Glu Glu Asp Arg Arg Phe Glu Ser His Ile Glu Cys Arg Lys Pro Tyr Lys Glu Leu Arg Leu Glu Val Gly Lys Gln Arg Leu Lys Tyr
TCA GTG GAT GAT CCC AGA TCA GAG GAG GAT AGA AGA TTT GAG AGT CAT ATA GAA TGT AGG AAA CCA TAC AAA GAA TTG AGA CTG GAG GTT GGG AAA CAA AGA CTC AAA TAT (3175)
                                                                                        P2,2A
                    820                                     830                                     840                                     850
Ala Gln Glu Glu Leu Ser Asn Glu Val Leu Pro Pro Pro Arg Lys Met Lys Gly Leu Phe Ser Gln Ala Lys Ile Ser Leu Phe Tyr Thr Glu Glu His Glu Ile Met Lys
GCT CAG GAA GAG TTA TCA AAT GAA GTG CTT CCA CCT CCT AGG AAA ATG AAG GGG TTA TTT TCA CAA GCT AAA ATT TCT CTT TTT TAT ACT GAG GAG CAT GAA ATA ATG AAG (3286)
                                860                                     870                                     880
Phe Ser Trp Arg Gly Val Thr Ala Asp Thr Arg Ala Leu Arg Arg Phe Gly Phe Ser Leu Ala Ala Gly Arg Ser Val Trp Thr Leu Glu Met Asp Ala Gly Val Leu Thr
TTT TCT TGG AGA GGA GTG ACT GCT GAT ACT AGG GCT TTG AGA AGA TTT GGA TTC TCT CTG GCT GCT GGT AGA AGT GTG TGG ACT CTT GAA ATG GAT GCT GGA GTT CTT ACT (3397)
    890                                     900                                     910                                     920
Gly Arg Leu Ile Arg Leu Asn Asp Glu Lys Trp Thr Glu Met Lys Asp Asp Lys Ile Val Ser Leu Ile Glu Lys Phe Thr Ser Asn Lys Tyr Trp Ser Lys Val Asn Phe
GGA AGA TTG ATC AGA TTG AAT GAT GAG AAA TGG ACA GAA ATG AAG GAT GAT AAG ATT GTT TCA TTA ATT GAA AAG TTC ACA AGC AAT AAA TAT TGG TCT AAA GTG AAT TTT (3508)
                930                                     940                                     950                                     960
Pro His Gly Met Leu Asp Leu Glu Glu Ile Ala Ala Asn Ser Lys Asp Phe Pro Asn Met Ser Glu Thr Asp Leu Cys Phe Leu Leu His Trp Leu Asn Pro Lys Lys Ile
CCA CAT GGA ATG TTG GAT CTT GAA GAA ATT GCT GCC AAT TCT AAG GAT TTT CCA AAT ATG TCT GAG ACA GAT TTG TGT TTC CTG TTA CAT TGG CTA AAT CCA AAG AAA ATC (3619)
                                                                        2B
                            970                                     980                                     990
Asn Leu Ala Asp Arg Met Leu Gly Leu Ser Gly Val Gln Glu Ile Lys Glu Gln Gly Val Gly Leu Ile Ala Glu Cys Arg Thr Phe Leu Asp Ser Ile Ala Gly Thr Leu
AAT TTA GCA GAT AGA ATG CTT GGA TTG TCT GGA GTG CAG GAA ATT AAG GAA CAG GGT GTT GGA CTG ATA GCA GAG TGT AGA ACT TTC TTG GAT TCT ATT GCT GGG ACT TTG (3730)
```

(*Fig. 2 continues on the next page.*)

```
1000                                    1010                                    1020                                    1030
Lys Ser Met Met Phe Gly Phe His His Ser Val Thr Val Glu Ile Ile Asn Thr Val Leu Cys Phe Val Lys Ser Gly Ile Leu Leu Tyr Val Ile Gln Gln Leu Asn Gln
AAA TCT ATG ATG TTT GGG TTT CAT CAT TCT GTG ACT GTT GAA ATT ATA AAT ACT GTG CTT TGT TTT GTT AAG AGT GGA ATC CTG CTT TAT GTC ATA CAA CAA TTG AAC CAA (3841)

            1040                                    1050                                    1060                                    1070
Asp Glu His Ser His Ile Ile Gly Leu Leu Arg Val Met Asn Tyr Ala Asp Ile Gly Cys Ser Val Ile Ser Cys Gly Lys Val Phe Ser Lys Met Leu Glu Thr Val Phe
GAT GAA CAC TCT CAC ATA ATT GGT TTG TTG AGA GTT ATG AAT TAT GCA GAT ATT GGC TGT TCA GTT ATT TCA TGT GGT AAA GTT TTT TCC AAA ATG TTA GAA ACA GTT TTT (3952)
            2C
                        1080                                    1090                                    1100                                    1110
Asn Trp Gln Met Asp Ser Arg Met Met Glu Leu Arg Thr Gln Ser Phe Ser Asn Trp Leu Arg Asp Ile Cys Ser Gly Ile Thr Ile Phe Lys Ser Phe Lys Asp Ala Ile
AAT TGG CAA ATG GAT TCT AGA ATG ATG GAG CTG AGG ACT CAG AGC TTC TCT AAT TGG TTA AGA GAT ATT TGT TCA GGA ATT ACT ATT TTT AAA AGT TTT AAG GAT GCC ATA (4063)

                                    1120                                    1130                                    1140
Tyr Trp Leu Tyr Thr Lys Leu Lys Asp Phe Tyr Glu Val Asn Tyr Gly Lys Lys Lys Asp Ile Leu Asn Ile Leu Lys Asp Asn Gln Gln Lys Ile Glu Lys Ala Ile Glu
TAT TGG TTA TAT ACA AAA TTG AAG GAT TTT TAT GAA GTA AAT TAT GGC AAG AAA AAG GAT ATT CTT AAT ATT CTC AAA GAT AAT CAG CAA AAA ATA GAA AAA GCC ATT GAA (4174)

        1150                                    1160                                    1170                                    1180
Glu Ala Asp Asn Phe Cys Ile Leu Gln Ile Gln Asp Val Glu Lys Phe Asp Gln Tyr Gln Lys Gly Val Asp Leu Ile Gln Lys Leu Arg Thr Val His Ser Met Ala Gln
GAA GCA GAC AAT TTT TGC ATT TTG CAA ATT CAA GAT GTA GAG AAA TTT GAT CAG TAT CAG AAA GGG GTT GAT TTA ATA CAA AAG CTG AGA ACT GTC CAT TCA ATG GCG CAA (4285)

                    1190                                    1200                                    1210                                    1220
Val Asp Pro Asn Leu Gly Val His Leu Ser Pro Leu Arg Asp Cys Ile Ala Arg Val His Gln Lys Leu Lys Asn Leu Gly Ser Ile Asn Gln Ala Met Val Thr Arg Cys
GTT GAC CCC AAT TTG GGG GTT CAT TTG TCA CCT CTC AGA GAT TGC ATA GCA AGA GTC CAC CAA AAG CTC AAG AAT CTT GGA TCT ATA AAT CAG GCC ATG GTA ACA AGA TGT (4396)

                                1230                                    1240                                    1250
Glu Pro Val Val Cys Tyr Leu Tyr Gly Lys Arg Gly Gly Gly Lys Ser Leu Thr Ser Ile Ala Leu Ala Thr Lys Ile Cys Lys His Tyr Gly Val Glu Pro Glu Lys Asn
GAG CCA GTT GTT TGC TAT TTG TAT GGC AAA AGA GGG GGA GGG AAA AGC TTG ACT TCA ATT GCA TTG GCA ACC AAA ATT TGT AAA CAC TAT GGT GTT GAA CCT GAG AAA AAT (4507)

    1260                                    1270                                    1280                                    1290
Ile Tyr Thr Lys Pro Val Ala Ser Asp Tyr Trp Asp Gly Tyr Ser Gly Gln Leu Val Cys Ile Ile Asp Asp Ile Gly Gln Asn Thr Thr Asp Glu Asp Trp Ser Asp Phe
ATT TAC ACC AAA CCT GTG GCC TCA GAT TAT TGG GAT GGA TAT AGT GGA CAA TTA GTT TGC ATT ATT GAT GAT ATT GGC CAA AAC ACA ACA GAT GAA GAT TGG TCA GAT TTT (4618)

                1300                                    1310                                    1320                                    1330
Cys Gln Leu Val Ser Gly Cys Pro Met Arg Leu Asn Met Ala Ser Leu Glu Glu Lys Gly Arg His Phe Ser Ser Pro Phe Ile Ile Ala Thr Ser Asn Trp Ser Asn Pro
TGT CAA TTA GTG TCA GGA TGC CCA ATG AGA TTG AAT ATG GCT TCT CTA GAG GAG AAG GGC AGA CAT TTT TCC TCT CCT TTT ATA ATA GCA ACT TCA AAT TGG TCA AAT CCA (4729)

                            1340                                    1350                                    1360
Ser Pro Lys Thr Val Tyr Val Lys Glu Ala Ile Asp Arg Arg Leu His Phe Lys Val Glu Val Lys Pro Ala Ser Phe Phe Lys Asn Pro His Asn Asp Met Leu Asn Val
AGT CCA AAA ACA GTT TAT GTT AAG GAA GCA ATT GAT CGT AGG CTT CAT TTT AAG GTT GAA GTT AAA CCT GCT TCA TTT TTT AAA AAT CCT CAC AAT GAT ATG TTG AAT GTT (4840)

1370                                    1380                                    1390                                    1400
Asn Leu Ala Lys Thr Asn Asp Ala Ile Lys Asp Met Ser Cys Val Asp Leu Ile Met Asp Gly His Asn Ile Ser Leu Met Asp Leu Leu Ser Ser Leu Val Met Thr Val
AAT TTG GCC AAA ACA AAT GAT GCA ATT AAG GAC ATG TCT TGT GTT GAT TTA ATA ATG GAT GGA CAC AAT ATT TCA TTG ATG GAT TTA CTT AGT TCC TTA GTG ATG ACA GTT (4951)
                                                                P3,3A
            1410                                    1420                                    1430                                    1440
Glu Ile Arg Lys Gln Asn Met Ser Glu Phe Met Glu Leu Trp Ser Gln Gly Ile Ser Asp Asp Asp Asn Asp Ser Ala Val Ala Glu Phe Phe Gln Ser Phe Pro Ser Gly
GAA ATT AGG AAA CAG AAT ATG AGT GAA TTC ATG GAG TTG TGG TCT CAG GGA ATT TCA GAT GAT GAC AAT GAT AGT GCA GTG GCT GAG TTT TTC CAG TCT TTT CCA TCT GGT (5062)

                        1450                                    1460                                    1470                                    1480
Glu Pro Ser Asn Trp Lys Leu Ser Ser Phe Phe Gln Ser Val Thr Asn His Lys Trp Val Ala Val Gly Ala Ala Val Gly Ile Leu Gly Val Leu Val Gly Gly Trp Phe
GAA CCA TCA AAT TGG AAG TTA TCT AGT TTT TTC CAA TCT GTC ACT AAT CAC AAG TGG GTT GCT GTG GGA GCT GCA GTT GGC ATT CTT GGA GTG CTT GTG GGA GGA TGG TTT (5173)
                3B                                                                                          3C
                                    1490                                    1500                                    1510
Val Tyr Lys His Phe Ser Arg Lys Glu Glu Glu Pro Ile Pro Ala Glu Gly Val Tyr His Gly Val Thr Lys Pro Lys Gln Val Ile Lys Leu Asp Ala Asp Pro Val Glu
GTG TAT AAG CAT TTT TCC CGC AAA GAG GAA GAA CCA ATT CCA GCT GAA GGG GTT TAT CAT GGC GTG ACT AAG CCC AAA CAA GTG ATT AAA TTG GAT GCA GAT CCA GTA GAG (5284)

        1520                                    1530                                    1540                                    1550
Ser Gln Ser Thr Leu Glu Ile Ala Gly Leu Val Arg Lys Asn Leu Val Gln Phe Gly Val Gly Glu Lys Asn Gly Cys Val Arg Trp Val Met Asn Ala Leu Gly Val Lys
TCC CAG TCA ACT CTA GAA ATA GCA GGA TTA GTT AGG AAA AAT CTG GTT CAG TTT GGA GTT GGT GAG AAA AAT GGA TGT GTG AGA TGG GTC ATG AAT GCC TTA GGA GTG AAG (5395)

                    1560                                    1570                                    1580                                    1590
Asp Asp Trp Leu Leu Val Pro Ser His Ala Tyr Lys Phe Glu Lys Asp Tyr Glu Met Met Glu Phe Tyr Phe Asn Arg Gly Gly Thr Tyr Tyr Ser Ile Ser Ala Gly Asn
GAT GAT TGG TTG TTA GTA CCT TCT CAT GCT TAT AAA TTT GAA AAG GAT TAT GAA ATG ATG GAG TTT TAC TTC AAT AGA GGT GGA ACT TAC TAT TCA ATT TCA GCT GGT AAT (5506)

                                1600                                    1610                                    1620
Val Val Ile Gln Ser Leu Asp Val Gly Phe Gln Asp Val Val Leu Met Lys Val Pro Thr Ile Pro Lys Phe Arg Asp Ile Thr Gln His Phe Ile Lys Lys Gly Asp Val
GTT GTT ATT CAA TCT TTA GAT GTG GGA TTT CAA GAT GTT GTT TTA ATG AAG GTT CCT ACA ATT CCC AAG TTT AGA GAT ATT ACT CAA CAC TTT ATT AAG AAA GGA GAT GTG (5617)

    1630                                    1640                                    1650                                    1660
Pro Arg Ala Leu Asn Arg Leu Ala Thr Leu Val Thr Thr Val Asn Gly Thr Pro Met Leu Ile Ser Glu Gly Pro Leu Lys Met Glu Glu Lys Ala Thr Tyr Val His Lys
CCT AGA GCC TTA AAT CGC TTG GCA ACA TTA GTG ACA ACC GTT AAT GGA ACT CCT ATG TTA ATT TCT GAG GGA CCA TTA AAG ATG GAA GAA AAA GCC ACT TAT GTT CAT AAG (5728)
                                                    3D
                1670                                    1680                                    1690                                    1700
Lys Asn Asp Gly Thr Thr Val Asp Leu Thr Val Asp Gln Ala Trp Arg Gly Lys Gly Glu Gly Leu Pro Gly Met Cys Gly Gly Ala Leu Val Ser Ser Asn Gln Ser Ile
AAG AAT GAT GGT ACT ACA GTT GAT TTG ACT GTA GAT CAG GCA TGG AGA GGA AAA GGT GAA GGT CTT CCT GGA ATG TGT GGT GGG GCC CTA GTG TCA TCA AAT CAG TCC ATA (5839)

                            1710                                    1720                                    1730
Gln Asn Ala Ile Leu Gly Ile His Val Ala Gly Gly Asn Ser Ile Leu Val Ala Lys Leu Val Thr Gln Glu Met Phe Gln Asn Ile Asp Lys Lys Ile Glu Ser Gln Arg
CAG AAT GCA ATT TTG GGT ATT CAT GTT GCT GGA GGA AAT TCA ATT CTT GTG GCA AAG CTG GTT ACT CAA GAA ATG TTT CAA AAC ATT GAT AAG AAA ATT GAA AGT CAG AGA (5940)

1740                                    1750                                    1760                                    1770
Ile Met Lys Val Glu Phe Thr Gln Cys Ser Met Asn Val Val Ser Lys Thr Leu Phe Arg Lys Ser Pro Ile His His His Ile Asp Lys Thr Met Ile Asn Phe Pro Ala
ATA ATG AAA GTG GAA TTT ACT CAA TGT TCA ATG AAT GTA GTC TCC AAA ACG CTT TTT AGA AAG AGT CCC ATT CAT CAC CAC ATT GAT AAA ACC ATG ATT AAT TTT CCT GCA (6051)

            1780                                    1790                                    1800                                    1810
Ala Met Pro Phe Ser Lys Ala Glu Ile Asp Pro Met Ala Met Met Leu Ser Lys Tyr Ser Leu Pro Ile Val Glu Glu Pro Glu Asp Tyr Lys Glu Ala Ser Val Phe Tyr
GCT ATG CCT TTC TCT AAA GCT GAA ATT GAT CCA ATG GCT ATG ATG TTG TCC AAA TAT TCA TTA CCT ATT GTG GAG GAA CCA GAG GAT TAC AAG GAA GCT TCA GTT TTT TAT (6162)

                        1820                                    1830                                    1840                                    1850
Gln Asn Lys Ile Val Gly Lys Thr Gln Leu Val Asp Asp Phe Leu Asp Leu Asp Met Ala Ile Thr Gly Ala Pro Gly Ile Asp Ala Ile Asn Met Asp Ser Ser Pro Gly
CAA AAC AAA ATA GTA GGC AAG ACT CAG CTA GTT GAT GAC TTT TTA GAT CTT GAT ATG GCT ATT ACA GGG GCT CCA GGC ATT GAT GCT ATC AAT ATG GAT TCA TCT CCT GGG (6273)

                                    1860                                    1870                                    1880
Phe Pro Tyr Val Gln Glu Lys Leu Thr Lys Arg Asp Leu Ile Trp Leu Asp Glu Asn Gly Leu Leu Leu Gly Val His Pro Arg Leu Ala Gln Arg Ile Leu Phe Asn Thr
TTT CCT TAT GTT CAA GAA AAA TTG ACC AAA AGA GAT TTA ATT TGG TTG GAT GAA AAT GGT TTG CTG TTA GGA GTT CAC CCA AGA TTG GCC CAG AGA ATT TTA TTT AAT ACT (6384)

        1890                                    1900                                    1910                                    1920
Val Met Met Glu Asn Cys Ser Asp Leu Asp Val Val Phe Thr Thr Cys Pro Lys Asp Glu Leu Arg Pro Leu Glu Lys Val Leu Glu Ser Lys Thr Arg Ala Ile Asp Ala
GTC ATG ATG GAA AAT TGT TCT GAC TTA GAT GTT GTT TTT ACA ACT TGT CCA AAA GAT GAA TTG AGA CCA TTA GAG AAA GTT TTG GAA TCA AAA ACA AGA GCC ATT GAT GCT (6495)

                    1930                                    1940                                    1950                                    1960
Cys Pro Leu Asp Tyr Thr Ile Leu Cys Arg Met Tyr Trp Gly Pro Ala Ile Ser Tyr Phe His Leu Asn Pro Gly Phe His Thr Gly Val Ala Ile Gly Ile Asp Pro Asp
TGT CCT TTG GAT TAT ACA ATT CTA TGT CGA ATG TAT TGG GGT CCA GCT ATC AGT TAT TTC CAT TTG AAT CCA GGG TTT CAC ACA GGT GTT GCT ATT GGC ATA GAT CCT GAT (6606)

                                1970                                    1980                                    1990
Arg Gln Trp Asp Glu Leu Phe Lys Thr Met Ile Arg Phe Gly Asp Val Gly Leu Asp Leu Asp Phe Ser Ala Phe Asp Ala Ser Leu Ser Pro Phe Met Ile Arg Glu Ala
AGA CAG TGG GAT GAA TTA TTT AAA ACA ATG ATA AGA TTT GGA GAT GTT GGT CTT GAT TTA GAT TTC TCT GCT TTT GAT GCC AGT CTT AGT CCA TTT ATG ATT AGG GAA GCA (6717)

    2000                                    2010                                    2020                                    2030
Gly Arg Ile Met Ser Glu Leu Ser Gly Thr Pro Ser His Phe Gly Thr Ala Leu Ile Asn Thr Ile Ile Tyr Ser Lys His Leu Leu Tyr Asn Cys Cys Tyr His Val Cys
GGT AGA ATC ATG AGT GAA TTA TCT GGA ACA CCA TCT CAT TTT GGA ACA GCT CTT ATC AAT ACT ATC ATT TAT TCT AAA CAT CTG CTG TAC AAC TGT TGT TAT CAT GTT TGT (6828)

                2040                                    2050                                    2060                                    2070
Gly Ser Met Pro Ser Gly Ser Pro Cys Thr Ala Leu Leu Asn Ser Ile Ile Asn Asn Ile Asn Leu Tyr Tyr Val Phe Ser Lys Ile Phe Gly Lys Ser Pro Val Phe Phe
GGT TCA ATG CCT TCT GGG TCT CCT TGC ACA GCT TTG TTG AAT TCA ATT ATT AAT AAT ATT AAT CTG TAT TAT GTG TTT TCT AAA ATA TTT GGA AAG TCT CCA GTT TTC TTT (6939)

                            2080                                    2090                                    2100
Cys Gln Ala Leu Arg Ile Leu Cys Tyr Gly Asp Asp Val Leu Ile Val Phe Ser Arg Asp Val Gln Ile Asp Asn Leu Asp Leu Ile Gly Gln Lys Ile Val Asp Glu Phe
TGT CAA GCT TTG AGG ATC CTT TGT TAC GGA GAT GAT GTT TTG ATA GTT TTT TCC AGA GAT GTT CAA ATT GAC AAT CTT GAC TTG ATT GGA CAG AAA ATT GTA GAT GAG TTC (7050)

2110                                    2120                                    2130                                    2140
Lys Lys Leu Gly Met Thr Ala Thr Ser Ala Asp Lys Asn Val Pro Gln Leu Lys Pro Val Ser Glu Leu Thr Phe Leu Lys Arg Ser Phe Asn Leu Val Glu Asp Arg Ile
AAA AAA CTT GGC ATG ACA GCC ACC TCA GCT GAT AAA AAT GTG CCT CAA CTG AAG CCA GTT TCA GAA TTG ACT TTT CTC AAA AGA TCT TTC AAT TTG GTG GAG GAT AGA ATT (7161)

            2150                                    2160                                    2170                                    2180
Arg Pro Ala Ile Ser Glu Lys Thr Ile Trp Ser Leu Met Ala Trp Gln Arg Ser Asn Ala Glu Phe Glu Gln Asn Leu Glu Asn Ala Gln Trp Phe Ala Phe Met His Gly
AGA CCT GCA ATT TCA GAA AAG ACA ATT TGG TCT TTG ATG GCT TGG CAG AGA AGT AAC GCT GAG TTT GAG CAG AAT TTA GAA AAT GCT CAG TGG TTT GCT TTT ATG CAT GGC (7272)

                        2190                                    2200                                    2210                                    2220
Tyr Glu Phe Tyr Gln Lys Phe Tyr Tyr Phe Val Gln Ser Cys Leu Glu Lys Glu Met Ile Glu Tyr Arg Leu Lys Ser Tyr Asp Trp Trp Arg Met Arg Phe Tyr Asp Gln
TAT GAG TTC TAT CAG AAA TTT TAT TAT TTT GTT CAG TCC TGT TTG GAG AAA GAG ATG ATA GAA TAT AGA CTT AAA TCT TAT GAT TGG TGG AGA ATG AGA TTT TAT GAC CAG (7383)

                        2227
Cys Phe Ile Cys Asp Leu Ser OP
TGT TTC ATT TGT GAC CTT TCA TGA TTTGTTTAAACAAATTTTCTTACTCTTTCTGAGGTTTGTTTATTTCTTTTGTCCGCTAACTAAAAAAAAAAAAAAAA (7484)

Translated Mol. Weight = 251,940
```

FIG. 2. Nucleotide sequence of the HAV genome and the predicted amino acid sequence. Arrows indicate the putative cleavage sites for the HAV polyprotein as determined by comparative structural analysis of poliovirus and HAV. The putative cleavage site between regions 1C and 1D (VP1 and VP4) could not be predicted by this analysis. The nomenclature used for the HAV putative polypeptides is the one recommended by Rueckert and Wimmer (13). Clone pHAV1 contains a deletion of 21 nucleotides. The sequence between nucleotide 3529 and nucleotide 3551 was independently derived from another HAV clone.

in poliovirus (14, 15). The role of these sequences in the viral life cycle has not been determined.

The Coding Sequence. A leader sequence, 733 nucleotides long, precedes the first candidate AUG start codon. This triplet opens a 2227-amino-acid-long reading frame. A second AUG, two amino acids downstream, is part of the same reading frame and also could serve as an initiation codon (18). The putative HAV polyprotein translated from the nucleotide sequence has an M_r of 251,940, once more in good agreement with data reported for poliovirus (14, 15).

The HAV Polyprotein. Since no precise information is available on HAV-coded structural and nonstructural proteins, we have attempted to analyze the HAV coding sequence and derive some mapping information by comparative studies with poliovirus. A direct comparison of the nucleotide sequence of HAV and poliovirus failed to reveal

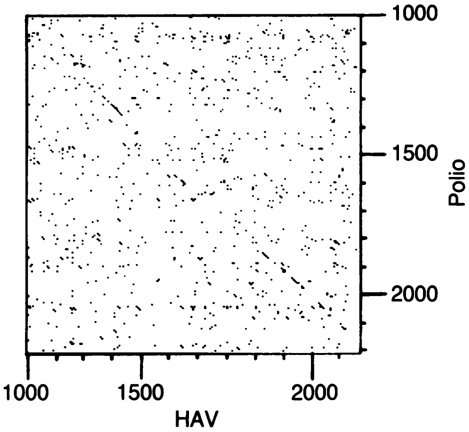

FIG. 3. Dot-matrix analysis of homology between the amino acid sequence of HAV and of poliovirus. The amino acid sequences of the two viruses were compared, starting at amino acid 1000. The computer program used a window of five amino acids with a filter of four matches out of five. Amino acids having similar properties were equated as follows: Thr=Ser; Tyr=Phe; Val=Leu=Ile; Asp=Glu; Lys=Arg; and Asn=Gln.

any detectable homology.

To verify the accuracy of the putative reading frame throughout the HAV genome, two separate analyses of the sequence were conducted. First, the codon usage in the HAV open reading frame was examined and found to be consistent with that reported for poliovirus (15) and other human genes. In addition, the codon usage within the HAV sequence itself was checked by comparing consecutive blocks of 1,000 nucleotides through the entire genome. No inconsistencies were observed in this internal comparison, leading to the conclusion that the nucleotide sequence encodes the correct polyprotein in every part of the HAV genome, and no substantial areas of frameshift have been introduced by sequencing errors. Second, the amino acid sequences of HAV and poliovirus were compared. A single area of homology was detected within the amino-terminal 1000 amino acids of the two viruses. The amino acid sequence of this region from the two viruses is:

Poliovirus Val-Pro-Trp-Ile-Ser-Asn-Ser-Thr-Xaa-Tyr-Arg
507
HAV Val-Pro-Trp-Ile-Ser-Asp-Xaa-Thr-Pro-Tyr-Arg
418

The positions of this homology stretch in the protein sequence (poliovirus amino acid 508, HAV amino acid 418) indicates that HAV has undergone approximately a 300-nucleotide shift in the position of this sequence compared to

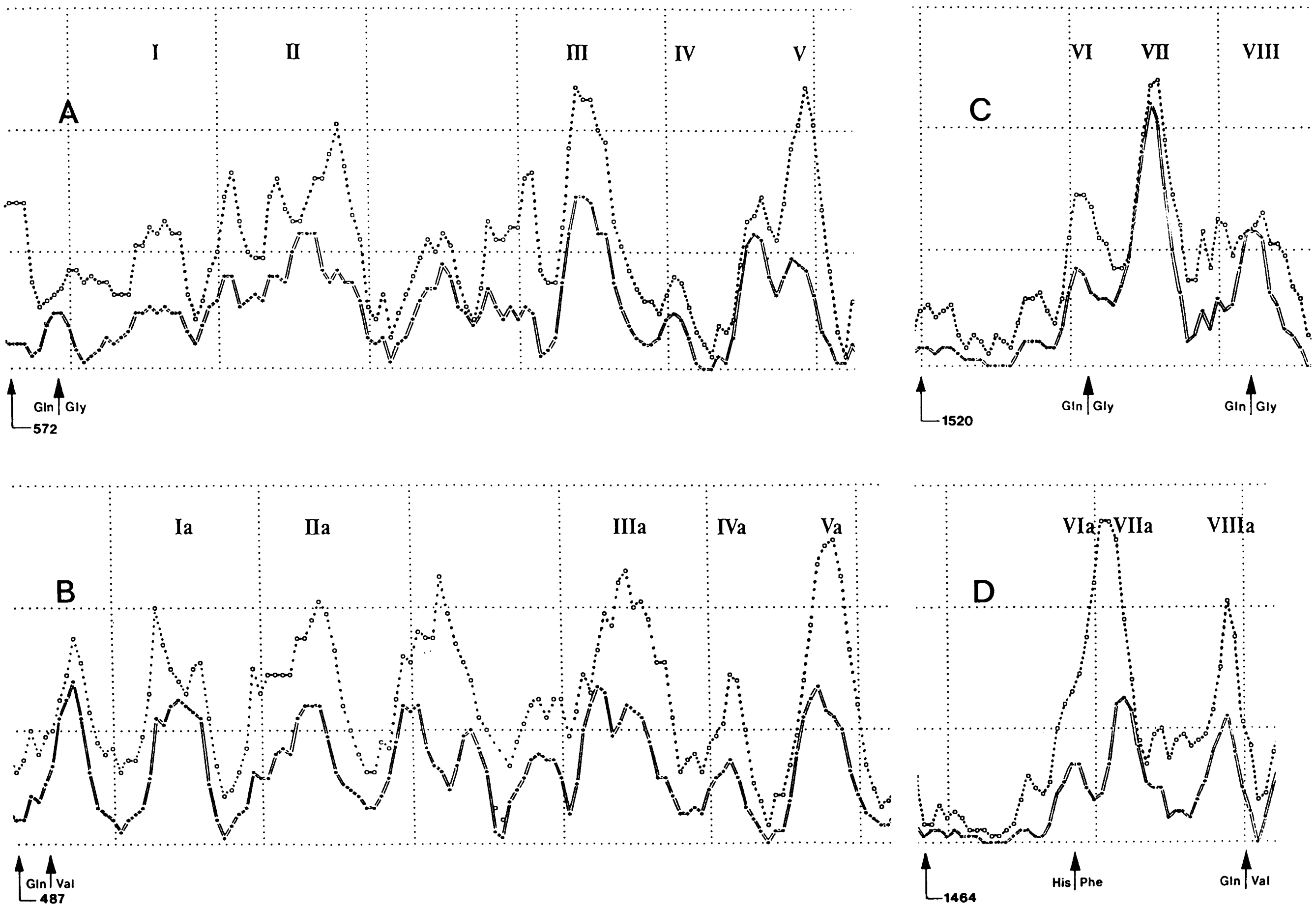

FIG. 4. Structural analysis of selected regions of HAV and poliovirus. The structural profile was computed by using a combination of structure and hydrophilicity. Three parameters were mixed: (*i*) the potential for helix or the β-sheet formation of the individual residues (19, 21) (values greater than 1.0 were scored as structure positive, and the higher one was selected; below 1.0, the lower value was selected), (*ii*) the hydrophilicity index of the residue (20), and (*iii*) the average hydrophilicity index of the four surrounding residues. Each parameter was normalized to affect the final curve evenly. Curves were corrected by averaging over seven residues. ......, Curve showing the mixing of parameters *ii* and *iii* such that a peak represents a region of high hydrophilicity; = = =, curve showing the mixing of the three parameters such that a peak represents a region of high hydrophilicity and lack of structural features. Arrows with numbers indicate the position of the first amino acid shown on the viral polyprotein. Each dot of the grid represents one residue, and each grid space contains 20 residues. Known cleavage positions for poliovirus and putative ones for HAV are indicated by arrows and amino acid pairs. (*A*) Poliovirus VP1. (*B*) HAV VP1. (*C*) Poliovirus VPg. (*D*) HAV VPg.

poliovirus. Taking into account that both polyproteins start within a few nucleotides on the respective genomes, we conclude that the initial portion of region P1 in HAV has undergone a small deletion.

A comparison of the second half of HAV and poliovirus polyproteins (amino acids 1000–2200) is shown in Fig. 3. Here a number of homologous stretches were identified, indicating a higher degree of conservation in this part of the viral genome. Our data indicate that the portions of the HAV genome that, by analogy with poliovirus, are expected to code for the enzymes involved in viral replication are more highly conserved than the regions coding for the virion proteins and viral antigenic determinants. The finding of short regions of homology throughout the second part of the two genomes supports the notion that the HAV open reading frame is correctly identified by the nucleotide sequence.

Structural Comparison of HAV and Poliovirus Polyproteins. A number of recent reports on protein structure have indicated that proteins carrying out homologous functions but with little or no sequence homology do display considerable "structural" similarities. Among these, areas of high hydrophilicity devoid of structural features (i.e., α helices and β sheets) have been found to correspond often to major antigenic sites (13, 19, 20).

HAV and poliovirus polyproteins were compared by using a computer program that displays nonstructural and hydrophilic properties of polypeptides (Fig. 4). Highly similar profiles were obtained throughout the entire polyprotein. Examples of these similarities for regions of interest are shown in Fig. 4. The short region of amino acid homology between HAV and poliovirus (poliovirus, amino acid 507; HAV, amino acid 418) was used to match the two sequences.

Fig. 4*A* shows the amino-terminal portion of poliovirus VP1. The cleavage site between VP3 and VP1 is identified by the Gln-Gly amino acid pair. Fig. 4*B* represents the matched homologous region on HAV. In a position quite close to the Gln-Gly cleavage pair on poliovirus, a Gln-Val pair is found on HAV. We think that this is a likely candidate for the cleavage site between VP3 and VP1 of HAV. The overall structural and hydrophilicity profiles of the two proteins are quite similar. "Peaks" representing putative areas of structural homology have been identified with corresponding roman numerals. Areas III and V of poliovirus represent major antigenic sites of VP1 (22–25). One tentative conclusion that can be drawn from this analysis is that, if these structurally homologous regions serve similar functions in the two viruses, then areas IIIa and Va of HAV should represent potential major antigenic sites for this virus.

Another area of striking similarity is shown in Fig. 4 *C* and *D*. These areas represent polio's VPg and the homologous HAV region. A long hydrophobic stretch (20 amino acids) precedes the Gln-Gly cleavage site in poliovirus (Fig. 4*C*). A similar region is found in HAV; by matching this region and areas VI–VIa in the two viruses, one can infer that the HAV VPg should start around the His-Phe residues shown in Fig. 4*D*. This is followed by areas VII–VIIa and VIII–VIIIa, also displaying a high degree of structural homology. The end of VPg is marked by a Gln-Gly pair in poliovirus and possibly by the Gln-Val pair in HAV. If these cleavage sites of HAV VPg are correct, the resulting protein is 23 amino acids long (poliovirus VPg is 22 amino acids long) and contains a tyrosine residue at position 1499 (see Fig. 2), which can mediate the protein nucleic acid linkage at the 5′ end of the genomic RNA.

Using the same criteria of structural homology for comparison, we have attempted to map the rest of the HAV polyprotein with respect to the known poliovirus polyprotein-processing sites. Tentative boundaries for the P1, P2, and P3 regions have been identified and are shown in Fig. 2. Similarly, major putative cleavage sites are shown for the virion structural and nonstructural proteins. All gene products identified for poliovirus appear to have homologs in HAV at very similar positions except for VP4, which as a consequence of the 100-amino acid shift in the VP1 position as discussed, may not be encoded as the first protein in P1 but rather as the last one, following VP1. In general, with the exception of one VPg cleavage site (His-Phe), potential cleavage sites have a glutamine/hydrophobic amino acid pair.

It is important to stress that all of the cleavage sites indicated in Fig. 2 are tentative and that, in the absence of specific protein size and sequence data, no firm identifications are possible. However, we believe that this kind of analysis provides us with good working hypotheses for future experimentation. For example, tentative identification of VP1 coding sequences and putative antigenic sites can now be experimentally tested by the use of synthetic peptides in a manner analogous to that used for poliovirus. In addition, expression of individual portions of the HAV genome in microorganisms may provide us with specific gene products for detailed immunological studies.

We thank our colleagues at Chiron for all their useful suggestions and Drs. J. Ticehurst, B. Baroudy, and R. Purcell for providing materials and results prior to publication. We would also like to thank Becky Toman and KB Graphics for preparing the figures and Pauline Montes and Dana Topping for preparation of the typescript.

1. Siegl, G. & Frosner, G. G. (1978) *J. Virol.* **26,** 48–53.
2. Coulepis, A. G., Tannock, G. A., Locarnini, S. A. & Gust, I. D. (1981) *J. Virol.* **37,** 473–477.
3. Ticehurst, J. R., Racaniello, V. R., Baroudy, B. M., Baltimore, D., Purcell, R. H. & Feinstone, S. M. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 5885–5889.
4. Coulepis, A. G., Locarnini, S. A., Ferris, A. A., Lehmann, N. & Gust, I. D. (1978) *Intervirology* **10,** 24–31.
5. Coulepis, A. G., Locarnini, S. A. & Gust, I. D. (1980) *J. Virol.* **35,** 572–574.
6. Hughes, J. W., Stanton, L. W., Tomassini, J. E., Long, W. J. & Scolnick, E. M. (1984) *J. Virol.* **52,** 465–473..
7. von der Helm, K., Winnacker, E. L., Deinhardt, F., Frosner, G., Gauss-Muller, V., Bayerl, B., Scheid, R. & Siegl, G. (1981) *J. Virol. Methods* **3,** 37–44.
8. Daemer, R. J., Feinstone, S. M., Gust, I. D. & Purcell, R. H. (1981) *Infect. Immun.* **32,** 388–393.
9. Flehmig, B. (1980) *Med. Microbiol. Immunol.* **168,** 239–248.
10. Frosner, G. G., Denhardt, F., Scheid, R., Gauss-Muller, V., Holmes, N., Messelberger, V., Siegl, G. & Alexander, J. J. (1979) *Infection* **7,** 303–350.
11. Locarnini, S. A., Coulepis, A. G., Westaway, E. G. & Gust, I. D. (1981) *J. Virol.* **37,** 216–225.
12. Provost, P. J. & Hilleman, M. R. (1979) *Proc. Soc. Exp. Biol. Med.* **160,** 213–221.
13. Rueckert, R. R. & Wimmer, E. (1984) *J. Virol.* **50,** 957–959.
14. Kitamura, N., Semler, B. L., Rothberg, P. G., Larsen, G. R., Adler, C. J., Dorner, A. J., Emini, E. A., Hanecak, R., Lee, J. J., van der Werf, S., Anderson, C. W. & Wimmer, E. (1981) *Nature (London)* **291,** 547–553.
15. Racaniello, V. R. & Baltimore, D. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 4887–4891.
16. Lee, Y. F., Nomoto, A., Detjen, B. M. & Wimmer, E. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 59–63.
17. Nomoto, A., Detjen, B., Pozzatti, R. & Wimmer, E. (1977) *Nature (London)* **268,** 208–212.
18. Kozak, M. (1984) *Nucleic Acids Res.* **12,** 857–872.
19. Chou, P. Y. & Fasman, G. D. (1974) *Biochemistry* **13,** 211–222.
20. Hopp, T. P. & Woods, K. R. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 3824–3828.
21. Chou, P. Y. & Fasman, G. D. (1974) *Biochemistry* **13,** 222–245.
22. Emini, E. A., Jameson, B. A., Lewis, A. J., Larsen, G. R. & Wimmer, E. (1982) *J. Virol.* **43,** 997–1005.
23. Emini, E. A., Jameson, B. A., Lewis, A. J. & Wimmer, E. (1983) *Nature (London)* **304,** 699–702.
24. Evans, D. M. A., Minor, P. D., Schild, G. S. & Almond, J. W. (1983) *Nature (London)* **304,** 459–462.
25. Minor, P. D., Schild, G. C., Bootman, J., Evans, D. M. A., Ferguson, M., Reeve, P., Spitz, M., Stanway, G., Cann, A. J., Hauptmann, R., Clarke, L. D., Mountford, R. C. & Almond, J. W. (1983) *Nature (London)* **301,** 674–679.