Proc. Natl. Acad. Sci. USA
Vol. 82, pp. 2143–2147, April 1985
Medical Sciences

Sequence analysis of hepatitis A virus cDNA coding for capsid proteins and RNA polymerase

(hepatitis A virus RNA/picornavirus/sequence homology)

BAHIGE M. BAROUDY*, JOHN R. TICEHURST*, THOMAS A. MIELE*, JACOB V. MAIZEL, JR.†, ROBERT H. PURCELL*, AND STEPHEN M. FEINSTONE*

*Laboratory of Infectious Diseases, National Institute of Allergy and Infectious Diseases, and †Laboratory of Mathematical Biology, National Cancer Institute, National Institutes of Health, Bethesda, MD 20205

Communicated by Robert M. Chanock, November 19, 1984

ABSTRACT **We report here the nucleotide sequence corresponding to two large regions of the hepatitis A virus (HAV) genome. These comprise a sequence of 3274 bases corresponding to the 5′ end of the genome, which includes the putative capsid protein region of this picornavirus, and 1590 bases corresponding to the 3′ end of the genome, terminating in a 15-base poly(A) tract. These sequences revealed that HAV had the characteristic genomic organization of picornaviruses: an open reading frame beginning approximately 750 bases from the 5′ end of the RNA and a termination codon 60 bases from the 3′ poly(A) tract. The predicted amino acid sequences of both regions have been compared to analogous regions previously determined for other picornaviruses. There was sufficient homology to conclude that the 5′ region of HAV codes for capsid proteins and that the 3′ region codes for an RNA polymerase. However, these regions of HAV were not found to be closely related to analogous regions of poliovirus, encephalomyocarditis virus, and foot and mouth disease virus.**

Hepatitis A virus (HAV) has been classified as a picornavirus (1) belonging to the genus *Enterovirus* (2). Although picornaviruses express different pathogenic and epidemiologic characteristics, they share certain structural features. They all have an RNA genome that is single stranded and positive sense. Picornaviral genomes do not have the cap structure usually found at the 5′ end of eukaryotic mRNAs. There is a virus-encoded protein, VPg, covalently linked to the 5′ end of the genome. A 3′ poly(A) tract is virus encoded and is not added by cellular mechanisms. The genome of these viruses is translated into a single polypeptide from an open reading frame that begins approximately 750 bases from the 5′ end of the RNA molecule and spans almost the entire genome. This protein is cleaved subsequently by both host- and virus-encoded proteolytic enzymes to yield functional viral proteins. Surface antigens of picornaviruses are present in the virion capsid, which is composed of proteins VP1, VP2, VP3, and VP4 (3). While all of the described features of HAV resemble those of other picornaviruses, it has not yet been demonstrated that a VPg exists or that the genome is translated into a single precursor polypeptide.

Wild-type virus was amplified in marmosets, and greater than 99% of the genome was cloned as cDNA (4) to aid in the study of the molecular and antigenic structure of HAV. We report here the sequences corresponding to two large regions of the genome: a sequence of 3274 bases corresponding to the 5′ end and that of 1590 bases corresponding to the 3′ end of HAV. These regions of the HAV genome have been compared, using computer programs, to analogous regions previously determined in other laboratories for other picornaviruses.

MATERIALS AND METHODS

Plasmid Purification. HAV recombinant plasmids characterized by Ticehurst *et al.* (4) were propagated and purified by ethidium bromide/cesium chloride centrifugation (ref. 5, pp. 86–94).

End Labeling and Sequencing of DNA Fragments. DNA fragments (5–15 pmol) were labeled by using the Klenow fragment of *Escherichia coli* DNA polymerase I (6), T4 polynucleotide kinase (ref. 5, pp. 122–124), or terminal deoxynucleotidyltransferase (7). End-labeled DNA fragments were cleaved asymmetrically with the appropriate restriction enzymes, separated by electrophoresis on agarose gels, and purified subsequently by the glass extraction method (8). The chemical method of Maxam and Gilbert (9) was used to determine the sequence of these DNA fragments.

Computer Analysis. Nucleotide sequences were stored and analyzed in a VAX 11/750 computer. The SEQH program (10) was used to determine locally homologous areas in nucleotide sequences of HAV and other picornaviruses (11–14). Several programs (10, 15, 16) were used to compare predicted amino acid sequences of HAV with those of other picornaviruses (11–13, 17–19) and cowpea mosaic virus (20, 21). Visual alignment of multiple amino acid sequences was made from the outputs of these protein homology programs to identify possible post-translational cleavage sites within the predicted amino acid sequences of HAV.

RESULTS

Sequence of the Capsid Region and 3′ End of the HAV Genome. We have partially sequenced the HAV recombinant plasmids $pHAV_{LB}39$, $pHAV_{LB}58$, $pHAV_{LB}113$, $pHAV_{LB}228$, $pHAV_{L}1307$, and $pHAV_{L}1688$ to arrive at the sequence shown in Fig. 1. This sequence represents the region of the genome encoding the putative capsid proteins of HAV. Ninety-five percent of the sequence was obtained on complementary strands, while the remainder was obtained from multiple determinations on one strand.

The translation of the sequence shown in Fig. 1 from three different frames revealed the existence of a long open reading frame beginning with an isoleucine codon at position 707. Two methionine codons were found, beginning at nucleotides 713 and 719, from where the open reading frame extends to as far as sequence has been determined (position 3274). A polypyrimidine tract (nucleotides 690–704) is located on the 5′ side of the initiation codons.

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

Abbreviations: EMC, encephalomyocarditis virus; FMDV, foot and mouth disease virus; HAV, hepatitis A virus; PV, poliovirus.

```
      10         20         30         40         50         60         70         80         90         100        110        120
TTCCGGAGTCCCTCTTGGAAGTCCATGGTGAGGGGACTTGATACCTCACCGCCGTTTGCCTAGGCTATAGGCTAAATTTTCCCTTTCCCTTTTCCCTTTCCTATTCCCTTTGTTTTGCTTG

      130        140        150        160        170        180        190        200        210        220        230        240
TAAATATTAATTCCTGCAGGTTCAGGGTTCTTAAATCTGTTTCTCTATAAGAACACTCATTTTTCACGCTTTCTGTCTTCTTTCTTCCAGGGCTCTCCCCTTGCCCTAGGCTCTGGCCGT

      250        260        270        280        290        300        310        320        330        340        350        360
TGCGCCCGGCGGGGTCAACTCCATGATTAGCATGGAGCTGTAGGAGTCTAAATTGGGGACACAGATGTTTGGAACGTCACCTTGCAGTGTTAACTTGGCTTTCATGAATCTCTTTGATCT

      370        380        390        400        410        420        430        440        450        460        470        480
TCCACAAGGGGTAGGCTACGGGTGAAACCTCTTAGGCTAATACTTCTATGAAGAGATGCCTTGGATAGGGTAACAGCGGCGGATATTGGTGAGTTGTTAAGACAAAAACCATTCAACGCC

      490        500        510        520        530        540        550        560        570        580        590        600
GGAGGACTGACTCTCATCCAGTGGATGCATTGAGTGGATTGACTGTCAGGGCTGTCTTTAGGCTTAATTCCAGACCTCTCTGTGCTTAGGGCAAACATCATTTGGCCTTAAATGGGATTC

      610        620        630        640        650        660        670        680        690        700        710        720
TGTGAGAGGGGATCCCTCCATTGACAGCTGGACTGTTCTTTGGGGCCTTATGTGGTGTTTGCCTCTGAGGTACTCAGGGGCATTTAGGTTTTTCCTCATTCTTAAATAATAATGAACATG
                                                                                                                    MetAsnMet

      730        740        750        760        770        780        790        800        810        820        830        840
TCTAGACAAGGTATTTTCCAGACTGTTGGGAGTGGTCTTGACCACATCCTGTCTTTGGCAGACATTGAGGAAGAGCAAATGATTCAATCAGTTGATAGGACTGCAGTGACTGGTGCTTCT
SerArgGlnGlyIlePheGlnThrValGlySerGlyLeuAspHisIleLeuSerLeuAlaAspIleGluGluGluGlnMetIleGlnSerValAspArgThrAlaValThrGlyAlaSer

      850        860        870        880        890        900        910        920        930        940        950        960
TATTTTACTTCTGTGGATCAATCTTCAGTTCATACAGCTGAGGTTGGATCACACCAGGTTGAACCTTTGAGAACCTCTGTTGATAAACCCGGTTCAAAGAAGACTCAGGGAGAGAAATTT
TyrPheThrSerValAspGlnSerSerValHisThrAlaGluValGlySerHisGlnValGluProLeuArgThrSerValAspLysProGlySerLysLysThrGlnGlyGluLysPhe

      970        980        990        1000       1010       1020       1030       1040       1050       1060       1070       1080
TTCTTGATTCATTCTGCAGATTGGCTTACTACACATGCTCTTTTCCATGAAGTTGCAAAATTGGATGTGGTGAAATTATTATACAATGAGCAGTTTGCTGTTCAAGGGTTGTTGAGATAC
PheLeuIleHisSerAlaAspTrpLeuThrThrHisAlaLeuPheHisGluValAlaLysLeuAspValValLysLeuLeuTyrAsnGluGlnPheAlaValGlnGlyLeuLeuArgTyr

      1090       1100       1110       1120       1130       1140       1150       1160       1170       1180       1190       1200
CATACATATGCAAGATTTGGCATTGAAATTCAAGTTCAGATAAACCCTACACCTTTCCAACAGGGGGGATTGATCTGTGCTATGGTTCCTGGTGACCAGAGCTATGGTTCTATAGCATCA
HisThrTyrAlaArgPheGlyIleGluIleGlnValGlnIleAsnProThrProPheGlnGlnGlyGlyLeuIleCysAlaMetValProGlyAspGlnSerTyrGlySerIleAlaSer

      1210       1220       1230       1240       1250       1260       1270       1280       1290       1300       1310       1320
TTGACTGTTTATCCTCATGGTTTGTTAAATTGCAATATTAACAATGTGGTTAGAATAAAGGTTCCATTTATTTACACAAGAGGTGCTTACCACTTTAAAGATCCACAATACCCAGTTTGG
LeuThrValTyrProHisGlyLeuLeuAsnCysAsnIleAsnAsnValValArgIleLysValProPheIleTyrThrArgGlyAlaTyrHisPheLysAspProGlnTyrProValTrp

      1330       1340       1350       1360       1370       1380       1390       1400       1410       1420       1430       1440
GAATTGACAATTAGAGTTTGGTCAGAATTAAATATTGGGACAGGAACTTCAGCTTATACTTCACTCAATGTTTTAGCTAGATTTACAGATTTGGAGTTGCATGGATTAACTCCTCTTTCT
GluLeuThrIleArgValTrpSerGlnLeuAsnIleGlyThrGlyThrSerAlaTyrThrSerLeuAsnValLeuAlaArgPheThrAspLeuGluLeuHisGlyLeuThrProLeuSer

      1450       1460       1470       1480       1490       1500       1510       1520       1530       1540       1550       1560
ACACAAATGATGAGAAATGAATTTAGGGTCAGTACTACTGAGAATGTGGTGAATCTGTCAAATTATGAAGATGCAAGAGCAAAGATGTCTTTTGCTTTGGATCAGGAAGATTGGAAATCT
ThrGlnMetMetArgAsnGluPheArgValSerThrThrGluAsnValValAsnLeuSerAsnTyrGluAspAlaArgAlaLysMetSerPheAlaLeuAspGlnGluAspTrpLysSer

      1570       1580       1590       1600       1610       1620       1630       1640       1650       1660       1670       1680
GATCCGTCCCAGGGTGGTGGGATCAAAATTACTCATTTTACTACTTGGACATCTATTCCAACTTTGGCTGCTCAGTTTCCATTTAATGCTTCAGACTCAGTTGGTCAACAAATTAAAGTT
AspProSerGlnGlyGlyGlyIleLysIleThrHisPheThrThrTrpThrSerIleProThrLeuAlaAlaGlnPheProPheAsnAlaSerAspSerValGlyGlnGlnIleLysVal

      1690       1700       1710       1720       1730       1740       1750       1760       1770       1780       1790       1800
ATTCCAGTTGACCCATATTTTTTCCAAATGACAAATACGAATCCTGACCAAAAATGTATAACTGCTTTGGCTTCTATTTGTCAGATGTTTTGTTTTTGGAGAGGAGATCTTGTCTTTGAT
IleProValAspProTyrPhePheGlnMetThrAsnThrAsnProAspGlnLysCysIleThrAlaLeuAlaSerIleCysGlnMetPheCysPheTrpArgGlyAspLeuValPheAsp

      1810       1820       1830       1840       1850       1860       1870       1880       1890       1900       1910       1920
TTTCAAGTTTTTCCCACCAAATATCATTCAGGTAGATTACTGTTTTGTTTTGTTCCTGGCAATGAGCTAATAGATGTTTCTGGAATCACATTAAAGCAAGCAACTACTGCTCCTTGTGCA
PheGlnValPheProThrLysTyrHisSerGlyArgLeuLeuPheCysPheValProGlyAsnGluLeuIleAspValSerGlyIleThrLeuLysGlnAlaThrThrAlaProCysAla

      1930       1940       1950       1960       1970       1980       1990       2000       2010       2020       2030       2040
GTAATGGATATTACAGGAGTGCAGTCAACTTTGAGATTTCGTGTTCCCTGGATTTCTGACACTCCTTACAGAGTGAACAGGTATACAAAGTCAGCACATCAGAAAGGTGAGTACACTGCC
ValMetAspIleThrGlyValGlnSerThrLeuArgPheArgValProTrpIleSerAspThrProTyrArgValAsnArgTyrThrLysSerAlaHisGlnLysGlyGluTyrThrAla

      2050       2060       2070       2080       2090       2100       2110       2120       2130       2140       2150       2160
ATTGGGAAGCTTATTGTGTATTGTTATAACAGATTGACCTCTCCTTCTAACGTTGCTTCCCATGTCAGAGTGAATGTTTATCTTTCAGCAATTAACTTGGAATGTTTTGCTCCTCTTTAT
IleGlyLysLeuIleValTyrCysTyrAsnArgLeuThrSerProSerAsnValAlaSerHisValArgValAsnValTyrLeuSerAlaIleAsnLeuGluCysPheAlaProLeuTyr

      2170       2180       2190       2200       2210       2220       2230       2240       2250       2260       2270       2280
CATGCTATGGATGTTACTACACAAGTTGGAGATGATTCTGGAGGTTTTTCAACAACAGTTTCTACAGAACAGAATGTTCCAGATCCCCAAGTTGGTATAACAACCATGAAAGATTTGAAA
HisAlaMetAspValThrThrGlnValGlyAspAspSerGlyGlyPheSerThrThrValSerThrGluGlnAsnValProAspProGlnValGlyIleThrThrMetLysAspLeuLys

      2290       2300       2310       2320       2330       2340       2350       2360       2370       2380       2390       2400
GGAAAAGCTAACAGAGGGAAAATGGATGTTTCAGGAGTACAAGCACCTGTGGGAGCTATCACAACAATTGAGGATCCAGTTTTAGCAAAGAAAGTACCTGAGACATTTCCTGAATTGAAA
GlyLysAlaAsnArgGlyLysMetAspValSerGlyValGlnAlaProValGlyAlaIleThrThrIleGluAspProValLeuAlaLysLysValProGluThrPheProGluLeuLys

      2410       2420       2430       2440       2450       2460       2470       2480       2490       2500       2510       2520
CCTGGAGAATCCAGACATACATCAGATCATATGTCCATCTACAAGTTTATGGGAAGGTCTCATTTCTTGTGCACTTTTACATTCAATTCAAATAATAAAGAGTACACATTTCCTATAACC
ProGlyGluSerArgHisThrSerAspHisMetSerIleTyrLysPheMetGlyArgSerHisPheLeuCysThrPheThrPheAsnSerAsnAsnLysGluTyrThrPheProIleThr

      2530       2540       2550       2560       2570       2580       2590       2600       2610       2620       2630       2640
TTGTCTTCAACCTCTAATCCTCCTCATGGTTTGCCATCAACACTGAGGTGGTTTTTCAACTTGTTTCAGTTGTATAGAGGGCCTTTAGATCTGACAATTATTATTACAGGAGCAACTGAT
LeuSerSerThrSerAsnProProHisGlyLeuProSerThrLeuArgTrpPhePheAsnLeuPheGlnLeuTyrArgGlyProLeuAspLeuThrIleIleIleThrGlyAlaThrAsp

      2650       2660       2670       2680       2690       2700       2710       2720       2730       2740       2750       2760
GTAGATGGCATGGCCTGGTTCACTCCAGTAGGTCTTGCCGTTGATACTCCTTGGGTAGAGAAGGAGTCAGCTTTGTCTATTGACTACAAAACTGCTCTTGGAGCTGTCAGATTTAACACA
ValAspGlyMetAlaTrpPheThrProValGlyLeuAlaValAspThrProTrpValGluLysGluSerAlaLeuSerIleAspTyrLysThrAlaLeuGlyAlaValArgPheAsnThr

      2770       2780       2790       2800       2810       2820       2830       2840       2850       2860       2870       2880
AGGAGAACAGGGAACATTCAGATTAGATTACCATGGTATTCTTATTTATATGCTGTGTCTGGAGCACTGGATGGTTTGGGTGACAAGACAGATTCTACATTTGGATTGGTTTCTATTCAG
ArgArgThrGlyAsnIleGlnIleArgLeuProTrpTyrSerTyrLeuTyrAlaValSerGlyAlaLeuAspGlyLeuGlyAspLysThrAspSerThrPheGlyLeuValSerIleGln

      2890       2900       2910       2920       2930       2940       2950       2960       2970       2980       2990       3000
ATTGCAAATTACAATCATTCTGATGAATACTTGTCTTTTAGTTGTTATTTGTCTGTCACAGAACAATCAGAGTTTTATTTTCCCAGAGCTCCATTGAACTCAAATGCCATGTTATCCACT
IleAlaAsnTyrAsnHisSerAspGluTyrLeuSerPheSerCysTyrLeuSerValThrGluGlnSerGluPheTyrPheProArgAlaProLeuAsnSerAsnAlaMetLeuSerThr

      3010       3020       3030       3040       3050       3060       3070       3080       3090       3100       3110       3120
GAATCAATGATGAGCAGAATTGCAGCTGGAGACTTGGAGTCATCAGTGGATGATCCTAGATCAGAGGAAGATAAAAGATTTGAGAGTCATATAGAATGCAGGAAGCCATATAAAGAACTG
GluSerMetMetSerArgIleAlaAlaGlyAspLeuGluSerSerValAspAspProArgSerGluGluAspLysArgPheGluSerHisIleGluCysArgLysProTyrLysGluLeu

      3130       3140       3150       3160       3170       3180       3190       3200       3210       3220       3230       3240
AGATTAGAAGTTGGGAAACAAAGACTCAAGTATGCTCAGGAAGAATTGTCAAATGAAGTACTTCCACCCCCTAGGAAAATGAAGGGACTGTTTTCACAAGCCAAAATTTCTCTTTTTTAT
ArgLeuGluValGlyLysGlnArgLeuLysTyrAlaGlnGluGluLeuSerAsnGluValLeuProProProArgLysMetLysGlyLeuPheSerGlnAlaLysIleSerLeuPheTyr

      3250       3260       3270
ACTGAGGAGCATGAAATAATGAAGTTTTCCTGG
ThrGluGluHisGluIleMetLysPheSerTrp
```

FIG. 1. (*Legend appears at the bottom of the next page.*)

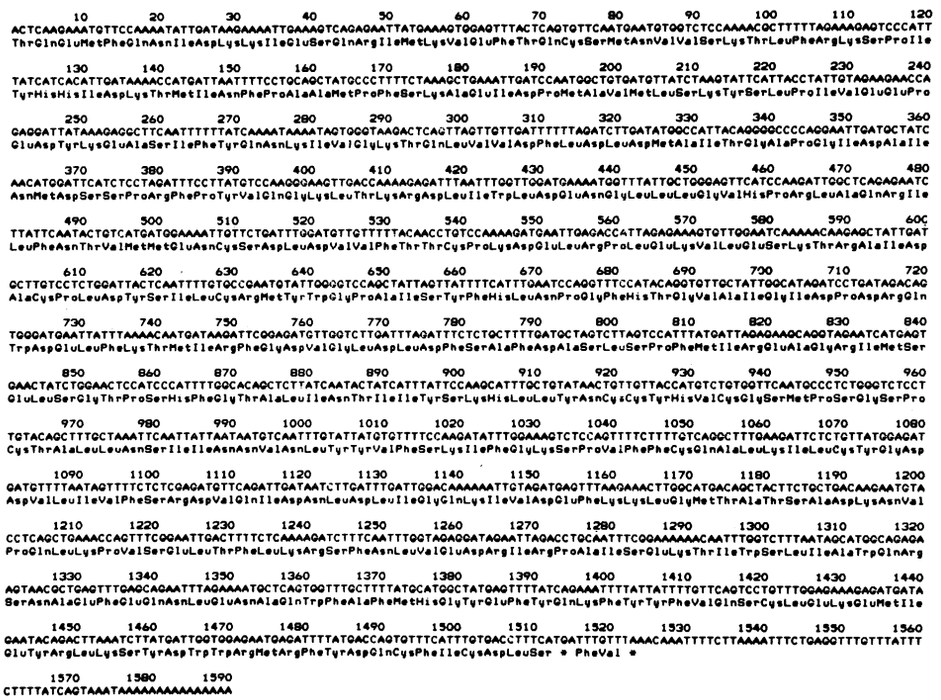

FIG. 2. Nucleotide and predicted amino acid sequence of cloned HAV cDNA corresponding to the 3′ end of HAV RNA. The poly(A) tract (nucleotides 1576–1590) was followed by poly(C) and the *Pst* I site of pBR322 in plasmids pHAV$_{LB}$24 and pHAV$_{LB}$93. The predicted amino acid sequence is shown below the nucleotide sequence until termination codons (*, nucleotides 1513–1515 and 1522–1524). A portion of this sequence was previously reported (4).

We have also obtained the entire sequence of pHAV$_{LB}$93, a plasmid that maps at the 3′ end of the genome (Fig. 2). The first 105 bases shown in Fig. 2 were obtained from the partial sequence of pHAV$_{LB}$24, a clone that contains sequences which extend upstream from those found in pHAV$_{LB}$93. A poly(A) tract of 15 bases is found at one end of pHAV$_{LB}$93 and presumably represents the 3′ end of HAV RNA. This poly(A) tract is 51 bases downstream from two closely spaced termination codons, which are preceded by 1512 bases in an open reading frame that is expected to be continuous with the one described above from the 5′ end of the genome. These termination codons are believed to be approximately 6700 bases downstream from the beginning of the reading frame.

Computer Analysis. Available nucleotide sequences from the 5′ untranslated regions of poliovirus (PV) type 1 (743 bases, ref. 12), encephalomyocarditis virus (EMC) (205 bases, ref. 13), and foot and mouth disease virus (FMDV) type O$_1$K (724 bases, ref. 14) were compared by using the SEQH program (10) with the analogous region determined for HAV. Locally homologous areas detected were no greater than 20 bases (data not shown).

The predicted amino acid sequence of the 5′ region of HAV (Fig. 1) was compared with the sequences of the capsid protein regions of PV type 1 (11, 12), EMC (13), and FMDV type A$_{10}$ (18) to search for areas of homology. By using graphic matrix analysis (15), homology was demonstrated between EMC and FMDV in VP1 and among PV, EMC, and FMDV in VP2, VP3, and the amino terminus of VP4 (ref. 22; data with EMC not shown). In contrast, graphic matrices or the SEQHP program (10) revealed less homology between HAV and PV, EMC, or FMDV. Homology was most evident between HAV and PV in VP3 (Fig. 3*A*) and between HAV and FMDV in VP2 (Fig. 3*B*). It was concluded from graphic matrices that VP4 of HAV is probably shorter than known VP4 proteins (69 amino acids) and that there is probably no leader protein like the leader proteins of EMC (13) and FMDV (14, 18, 23).

An attempt was made to position amino and carboxyl termini of putative HAV capsid proteins by amino acid sequence homology. In PV, the virus-encoded protease cleaves between glutamine and glycine at the VP2–VP3 and VP3–VP1 junctions (11, 12) and a similar mechanism has been proposed for maturation of these proteins in other picornaviruses (13, 14, 18, 23). However, none of the dipeptides that are known to demarcate the junctions VP2–VP3 and VP3–VP1 were found in the HAV sequences that aligned with those of PV, EMC, and FMDV. Although HAV sequences aligned with known cleavage sites so that glutamine could be proposed as the carboxyl terminus of both VP2 and VP3, the amino acids that follow, methionine and valine (Fig. 1, nucleotides 1448–1450 and 2186–2188), are atypical (data not shown). The two capsid proteins of cowpea mosaic virus, a comovirus that shares many genomic characteristics with the picornaviruses, are cleaved between glutamine and methionine (20), but there was no significant homology between these capsid proteins and those of the picornaviruses by graphic matrix analysis (data not shown).

The known cleavage sites for VP4–VP2 and VP1–P2 are

FIG. 1 (*on preceding page*). Nucleotide and predicted amino acid sequence of cloned HAV cDNA corresponding to the putative capsid protein region of HAV. Nucleotide 1 is 20–30 bases from the 5′ terminus of HAV RNA (estimated from cDNA primer extension, unpublished data). In the first 706 nucleotides 10 Met codons are found. The longest peptide encoded is 26 amino acids (nucleotides 306–383). After the Met codon at nucleotide 713 an open reading frame extends to nucleotide 3274. This region codes for a polypeptide, 854 amino acids in length, shown below the nucleotide sequence. The nucleotide C determined at position 849 in pHAV$_{LB}$39 was T in pHAV$_{LB}$113, resulting in an Ile codon. The nucleotide G determined at position 2174 in pHAV$_{LB}$39 and pHAV$_{LB}$58 was A in pHAV$_{L}$1307, resulting in an Ile codon.

A. PV (VP3, carboxy portion, and VP1, amino portion) and HAV

```
PV :  LSLSPASDPRLSHTMLGEILNYYTHWAGSLKFTF LFCGSMMATGKLLVSYAPPGA  DPPKKR KEAMLGTHVIWDI G
HAV:  FQMTNTNPDQKCITALASICQMFCFWRGDLVFDFQVF PTKYHSGRLLFCFVPGNELIDVSGITLKQATTAPCAVMDITG

      LQSSCTMVVPWISNSTYR QTIDDSFTE GGY  IS VF Y QTRIVVPLSTPREMDILGFVSACNDFSVRLLRDTTHIE
      VQSTLRFRVPWISDTPYRVNRYTKSAHQKGEYTAIGKLIVYCYNRLTSPSNVASHVRVNVYLSAINLECFAPLYHAMDVT

      QKALAQGLGQMLESM IDNTVRETVGAATSRDALPNTEASGPTHSKEIPALTAVETGATNPLVPSDTVQT RHVVQHRSR
      TQ VGDDSGGFSTTVSTEQNVPDPQVGITTMKDLKGKANRGKMDVSGVQAPVGAITTIEDPVLAKKVPETFPELKPGESR
```

B. FMDV (VP4, carboxy terminus, and VP2) and HAV

```
FMDV: TGLFGAL LADKKTEETTLLEDRILTTRNGHTTSTTQSSVGVTYGYSTEEDHVAGPNTSGLETRVVQAERFFKKFLFDW
HAV : SGLDHILSLADIEEEQMIQSVDRTAVTGASYFTSVDQSSVHTAEVGSHQVEPLRTSVDKP GSKKTQGEKFFLIHSADWL

      TTDKPFGYLTKLELPTD HHGVFG HLVDSY AYMRNGWDVEVCAVGNQFNGGCLLVAMVPEWKAFDTREKYQLTLFPHQ
      TTHALFHEVAKLDVVKLLYNEQFAVQGLLRYHTYARFGIEIQVQINPTPFQQGGLICAMVPGDQSYGSIAS  LTVYPHG

      FISPRTNMTAHITVPYL GVNRYDQYKKHK P WTLVVMVLSPLTVSN TAA PQIKVYANIAPTYVHVAGELPSK EG
      LLNCNINNVVRIKVPFIYTRGAY HFKDPQYPVWELTIRVWSELNIGTGTSAYTSLNVLARFTDLELHGLTPLSTQMMRN
```

FIG. 3. Comparison of selected amino acid sequences. Sequences were determined by the SEQHP computer program (10), using standard parameters (BIAS = 0, DEL = 8). Amino acids are denoted by single-letter abbreviations and conserved amino acids by asterisks. (*A*) PV type 1 sequence obtained from refs. 11 and 12, amino acids 425–650, with the cleavage site between VP3 and VP1 (Gln^{578}-Gly^{579}) indicated by a line over the letters. HAV sequence predicted from Fig. 1, nucleotides 1703–2416. (*B*) FMDV type A_{10} sequence obtained from ref. 18, amino acids 164–391, with the cleavage sites between VP4 and VP2 (Ala^{172}-Asp^{173}) and between VP2 and VP3 (Glu^{390}-Gly^{391}) indicated by lines over the letters. HAV sequence predicted from Fig. 1, nucleotides 752–1459.

thought to be cleaved by a host-encoded protease. The sequence Leu-Ala-Asp, found at the VP4–VP2 junction in EMC and FMDV, was identified in HAV (Fig. 1, nucleotides 776–784) within a region that aligned with the VP4–VP2 site of FMDV (Fig. 3*B*), PV, and EMC. If Ala-Asp is used to demarcate VP4 and VP2 of HAV, VP4 would be smaller in size than determined by electrophoretic analysis (1) but its predicted length of 23 amino acids would be consistent with the analysis of graphic matrices. Alignment for the VP1–P2 junction was more difficult to establish because of greater divergence in this region, but it is likely that the entire sequence encoding VP1 has been determined.

Graphic matrix analysis was also used to compare the predicted amino acid sequence of the 3′ region of HAV (Fig. 2) with analogous sequences of other picornaviruses (Fig. 4). This region is known to encode an RNA polymerase, $3D^{pol}$ (terminology of Rueckert and Wimmer, ref. 25). Sequences of PV, EMC, and FMDV (Fig. 4 *D*, *E*, and *F*) demonstrated extensive homology among these viruses throughout this region. On the other hand, this region of HAV showed less homology to the previously determined $3D^{pol}$ sequences (Fig. 4 *A*, *B*, and *C*). Although data from the computer programs allowed us to say that the complete sequence for the putative HAV RNA polymerase had been determined, alignment of this region demonstrated that HAV does not have either of the dipeptides (Gln-Gly or Glu-Gly, see legend to Fig. 4) thought to be cleaved by the viral protease between $3C^{pro}$ and $3D^{pol}$. It is possible that an atypical dipeptide, either Glu-Ser or Gln-Arg (Fig. 2, bases 37–48), is recognized in HAV. Within the putative $3D^{pol}$ sequence of HAV, there is an amino acid segment (Ile-Leu-Cys-Tyr-Gly-Asp-Asp-Val-Leu-Ile, predicted from nucleotides 1063–1092 in Fig. 2) that is highly conserved in picornaviruses and among RNA-dependent nucleic acid polymerases in general (A. C. Palmenberg, personal communication; ref. 26).

DISCUSSION

The size and organization of the HAV genome resemble those of other picornaviruses. In our laboratory HAV RNA has consistently comigrated with PV RNA and 7.5-kilobase marker DNA when analyzed by electrophoresis after denaturation with glyoxal/dimethyl sulfoxide, in agreement with the size estimated from the sum of overlapping cloned and primer-extended cDNAs (4). Unlike the larger genomes of aphthoviruses and cardioviruses, HAV RNA lacks the poly(C) region found near the 5′ end and probably has insufficient coding capacity for a nonstructural protein at the amino terminus of the capsid protein region.

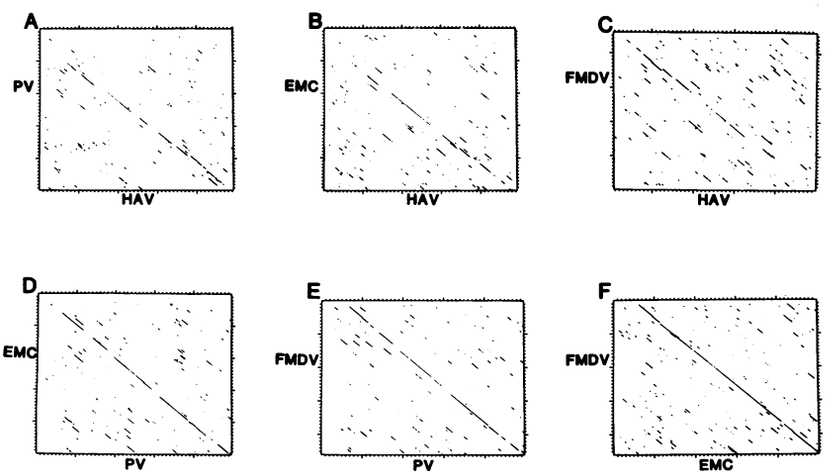

FIG. 4. Comparison of predicted amino acid sequences from the picornavirus genome regions coding for RNA polymerase. Graphic analysis (15) was performed by using the mutation data matrix of Dayhoff (24) with a window size of 25 and a minimum score of 15. The marks on the axes indicate intervals of 10 amino acid residues. The origin of each graph is located in the upper left corner. Sequences used were as follows: PV type 1 (*A*, *D*, and *E*), residues 1700–2207 (11, 12); EMC (*B*, *D*, and *F*), residues 1780–2290 (13); FMDV type A_{12} (*C*, *E*, and *F*), residues 1–471 (19); and HAV (*A*, *B*, and *C*), amino acids 1–504 predicted from bases 1–1512 in Fig. 2. The known or proposed cleavage sites between $3C^{pro}$ and $3D^{pol}$ are as follows: in PV, Gln^{1747}-Gly^{1748}; in EMC, Gln^{1830}-Gly^{1831}; and in FMDV type A_{12}, Glu^{1}-Gly^{2}. Sequence from the analogous region of rhinovirus type 2 (17), which has a proposed cleavage site of Gln-Gly, was highly homologous to PV and, similar to the patterns shown here for PV, was also homologous to the other picornaviral sequences (22). Less, but probably significant, relatedness was detected when this analysis was used to compare the predicted amino acid residues of cowpea mosaic virus bottom-component RNA (21) with those of PV type 1 or HAV (data not shown). The nucleotide sequence of HAV from Fig. 2 was also translated to the other two reading frames and compared by graphic analysis with PV and FMDV; there was no homology detected (data not shown).

As in other picornavirus genomes, a single long open reading frame begins over 700 nucleotides from the 5′ terminus of HAV. No function has yet been discovered for the untranslated region, although its presence is required for production of PV after transfection of cells with cloned PV type 1 cDNA (27). Mutations in this area appear to affect the virulence of PV type 3.‡ The sequence of this region is highly conserved among the serotypes of PV (28) and, in addition, between PV and coxsackievirus B3.§ There was no significant homology between this region of the HAV genome and analogous sequences available from PV, EMC, and FMDV. Approximately 740 bases from the 5′ end of the genome, two initiation codons were found in the sequence *A*TA ATG *A*AC ATG *T* (Fig. 1, nucleotides 710–722); these initiation codons are surrounded by consensus nucleotides (italicized), described by Kozak (29) as preferred for initiation of eukaryotic translation. In addition, the pyrimidine-rich region near the methionine codons is similar to that described in FMDV (23) and found in PV (11, 12) and EMC (13).

Because the HAV genome appears to be translated as a large polyprotein, comparison of known picornaviral amino acid sequences with those predicted for HAV was used in an attempt to define amino and carboxyl termini of HAV proteins. Greater sequence divergence in areas analogous to those that encode capsid proteins as compared with the putative RNA polymerase region was not surprising, because the viral capsid defines immunologic differences between picornaviruses. HAV may utilize Ala-Asp for host-mediated cleavage between VP4 and VP2 in a manner similar to EMC and FMDV. The dipeptides previously shown (or thought) to be cleaved by picornaviral proteases were absent from the predicted amino acid sequences of HAV that were aligned to sequences in the vicinity of previously established picornaviral cleavage sites. These data suggest that HAV uses different amino acid sequences or mechanisms in its post-translational cleavage scheme. The possible use of an atypical dipeptide, Gln-Val, at the VP3–VP1 cleavage site is supported by limited amino acid sequence data (30).

In general, the HAV genome is organized most like the genomes of the enteroviruses and rhinoviruses. However, the nucleotide sequence at the 5′ end is unlike any of the sequences already determined, with the exception of the pyrimidine-rich region found near the initiation codons as in PV, EMCV, and FMDV. Although there is sufficient homology in the predicted proteins to allow alignment and prediction of putative cleavage areas, the proteins of HAV appear to be more different from those of other picornaviruses than the latter appear to be when compared with each other.

After this manuscript was submitted, an additional sequence extending to the 5′ end of the HAV genome was determined from a cloned cDNA that was obtained by extension of a primer from $pHAV_{LB}113$ (*Nco* I to *Nci* I, nucleotides 23–246 in Fig. 1). The 22 nucleotides from the HAV 5′ terminus that were missing from Fig. 1 are 5′-T-T-C-A-A-G-A-G-G-G-G-T-C-T-C-C-G-G-G-A-A-T-3′. This sequence is in close agreement with, but not identical to, that obtained by D. Dina (personal communication) from a different strain of HAV.

We greatly appreciate John Owens' participation in developing and running computer programs, the technical assistance of Taylor Chestnut, editorial preparation by Marianne Guiler and Linda Jordan, and critical reading of our manuscript by Robert Chanock and Jeffrey Cohen. We thank Ann Palmenberg and also Patrick Argos and Eckard Wimmer for helpful discussions and information related to sequence alignment and for sending copies of their papers prior to publication.

1. Gust, I. D., Coulepis, A. G., Feinstone, S. M., Locarnini, S. A., Moritsugu, Y., Najera, R. & Siegl, G. (1983) *Intervirology* **20,** 1–7.
2. Melnick, J. L. (1982) *Intervirology* **18,** 105–106.
3. Putnak, J. R. & Phillips, B. A. (1981) *Microbiol. Rev.* **45,** 287–315.
4. 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.
5. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY).
6. Yang, R. C. A. & Wu, R. (1979) *Virology* **92,** 340–352.
7. Tu, C.-P. & Cohen, S. N. (1980) *Gene* **10,** 177–183.
8. Vogelstein, B. & Gillespie, D. (1979) *Proc. Natl. Acad. Sci. USA* **76,** 615–619.
9. Maxam, A. M. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 560–564.
10. Goad, W. B. & Kanehisa, M. I. (1982) *Nucleic Acids Res.* **10,** 247–263.
11. 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.
12. Rancaniello, V. R. & Baltimore, D. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 4887–4891.
13. Palmenberg, A. C., Kirby, E. M., Janda, M. R., Drake, H. L., Duke, G. M., Potratz, K. F. & Collett, M. S. (1984) *Nucleic Acids Res.* **12,** 2969–2985.
14. Forss, S., Strebel, K., Beck, E. & Schaller, H. (1984) *Nucleic Acids Res.* **12,** 6587–6601.
15. Maizel, J. V., Jr., & Lenk, R. P. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 7665–7669.
16. Murata, M., Richardson, J. S. & Sussman, J. L. (1985) *Proc. Natl. Acad. Sci. USA* **82,** in press.
17. Skern, T., Sommergruber, W., Blaas, D., Pieler, C. & Kuechler, E. (1984) *Virology* **136,** 125–132.
18. Boothroyd, J. C., Harris, T. J. R., Rowlands, D. J. & Lowe, P. A. (1982) *Gene* **17,** 153–161.
19. Robertson, B. H., Morgan, D. O., Moore, D. M., Grubman, M. J., Card, J., Fischer, T., Weddell, G., Dowbenko, D. & Yansura, D. (1983) *Virology* **126,** 614–623.
20. van Wezenbeek, P., Verver, J., Harmsen, J., Vos, P. & van Kammen, A. (1983) *EMBO J.* **2,** 941–946.
21. Lomonossoff, G. P. & Shanks, M. (1983) *EMBO J.* **2,** 2253–2258.
22. Baroudy, B. M., Ticehurst, J., Miele, T., Maizel, J. V., Purcell, R. H. & Feinstone, S. M. (1985) in *Modern Approaches to Vaccines*, eds. Lerner, R., Brown, F. & Chanock, R. M. (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY), in press.
23. Beck, E., Forss, S., Strebel, K., Cattaneo, R. & Feil, G. (1984) *Nucleic Acids Res.* **12,** 7873–7885.
24. Dayhoff, M. O. (1969) *Atlas of Protein Sequences and Structure* (National Biomedical Research Foundation, Silver Spring, MD).
25. Rueckert, R. R. & Wimmer, E. (1984) *J. Virol.* **50,** 957–959.
26. Kamer, G. & Argos, P. (1984) *Nucleic Acids Res.* **12,** 7269–7282.
27. Racaniello, V. R. & Baltimore, D. (1981) *Science* **214,** 916–919.
28. Toyoda, H., Kohara, M., Kataoka, Y., Suganuma, T., Omata, T., Imura, N. & Nomoto, A. (1984) *J. Mol. Biol.* **174,** 561–585.
29. Kozak, M. (1983) *Microbiol. Rev.* **47,** 1–45.
30. Hughes, J., Bennett, C., Stanton, L., Linemeyer, D. & Mitra, S. (1985) in *Modern Approaches to Vaccines*, eds. Lerner, R., Brown, F. & Chanock, R. M. (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY), in press.

‡Almond, J. W., Westrop, G. D., Stanway, G., Cann, A. J., Minor, P. D., Evans, D. M. A. & Schild, G. C. Sixth International Congress of Virology, Sept. 1–7, 1984, Sendai, Japan, p. 138 (abstr. no. P8-13).

§Tracy, S., Chapman, N. C. & Liu, H. Sixth International Congress of Virology, Sept. 1–7, 1984, Sendai, Japan, p. 204 (abstr. no. P31-5).