Nucleotide sequence of the genomic RNA of hepatitis C virus isolated from a human carrier: comparison with reported isolates for conserved and divergent regions

Hiroaki Okamoto, 1 Shunichi Okada, 2 Yasushi Sugiyama, 3 Kiyohiko Kurai, 1 Hisao Izuka, 4 Atsuhiko Machida, 5 Yuzo Miyakawa 6 and Makoto Mayumi 1*

1 Immunology Division, Jichi Medical School, Minamikawachi-Machi, Tochigi-Ken 329-04, 2 First Department of Internal Medicine, Yamanashi Medical College, Yamanashi-Ken 409-38, 3 Department of Hygiene, University of Hiroshima, Hiroshima-Ken 734, 4 Japanese Red Cross Blood Center, Saitama-Ken 362, 5 Department of Immunology, the Kitasato Institute, Tokyo 108 and 6 Mita Institute, Tokyo 108, Japan

The complete nucleotide sequence of a hepatitis C virus derived from plasma of a human carrier in Japan was determined. The cDNA of the isolate (HC-J6) contained 9481 nucleotides and an additional T stretch of 30 to 108 nucleotides at the 3' end, and had one large open reading frame coding for a polyprotein of 3033 amino acids. It differed by 31.8 to 32.1% in the nucleotide sequence and by 27-4 to 27-7% in the amino acid sequence from an American isolate and two Japanese isolates previously reported. Among these four isolates, the 5' non-coding region of 329 to 341 nucleotides was well conserved (>93% identity), whereas the 3' non-coding region of 39 to 45 nucleotides (T stretches not included) was more variable (>30% identity). An excellent degree of conservation of the 5' non-coding region would reflect its pivotal role in replication, and primers deduced from this region could be applied for the sensitive and specific detection of viral RNA by polymerase chain reaction. Due to a high degree of similarity in the amino acid sequence of the putative core protein (>90%), antigen probes deduced from it would be suitable for the serological diagnosis of HCV infection. Low sequence similarity in the putative envelope protein (>53% identity), however, would have to be taken into account in considering the immunoprophylaxis of HCV infection.

Introduction

Hepatitis caused by virus(es) other than hepatitis A or B viruses, referred to as non-A, non-B (NANB) hepatitis, currently accounts for the great majority of post-transfusion hepatitis cases, and has attracted increasing concern because of a high rate of chronic disease, with serious consequences leading to hepatocellular carcinoma (reviewed by Dienstag, 1983). Recognized in the early 1970s (Prince et al., 1974; Feinstone et al., 1975), the viral agent responsible for NANB hepatitis remained elusive for almost 2 decades.

In 1989, Choo et al. cloned the genome of an RNA virus from the plasma of a chimpanzee infected with material from a patient with NANB hepatitis, and designated it hepatitis C virus (HCV). An enzyme-linked immunosorbent assay has been developed, involving the translation product of non-structural (NS) regions (Kuo et al., 1989). A close association of NANB hepatitis with the antibody, known as anti-HCV, has been confirmed world-wide (Alter et al., 1989; Esteban et al., 1989; van der Poel et al., 1990). It has become evident that HCV is responsible for the majority of NANB hepatitis cases, and that testing for anti-HCV helps in the diagnosis of an HCV aetiology in patients with NANB hepatitis and is useful for excluding blood units contaminated with HCV.

Presently, the entire nucleotide sequence of HCV is available for at least three isolates (Kato et al., 1990; Choo et al., 1991; Takamizawa et al., 1991). The complete nucleotide sequence has now been determined for another HCV isolate, derived from a human asymptomatic carrier in Japan, and compared with these three isolates. Conserved and divergent regions of the HCV genome, outlined in the comparison, will contribute toward virological analyses of HCV, and may have implications in serological diagnosis and immunoprophylaxis of HCV infection.

The nucleotide sequence data reported in this paper are deposited in the DDBJ, GenBank and EMBL Nucleotide Sequence Databases under the accession number D00944.
Methods

Preparation of nucleic acids and synthesis of cDNA. HCV RNA was extracted as described elsewhere (Okamoto et al., 1990c), from the plasma of an asymptomatic carrier in Japan (HC-J6) which contained anti-HCV (HCV Ab ELISA test; Ortho Diagnostic Systems, Tokyo, Japan) and HCV RNA detectable by polymerase chain reaction (PCR) (Okamoto et al., 1990b). HCV RNA was denatured at 70 °C for 1 min, and reverse-transcribed using appropriate HCV-specific primers and Moloney murine leukemia virus reverse transcriptase (Superscript; BRL). Oligonucleotide primers (20-mers) deduced from isolates HC-J1, HC-J4 (Okamoto et al., 1990c) or HCV-1 (Choo et al., 1991) were synthesized in a Cyclone Plus DNA synthesizer (Milligen/Biosearch).

Amplification of cDNA by PCR and construction of cDNA libraries. cDNA was amplified by PCR with a GeneAmp DNA amplification reagent kit (Perkin-Elmer Cetus), or seeded for constructing cDNA libraries in the bacteriophage λgt10 with a cDNA synthesis kit (Amersham). Clones were selected by plaque hybridization as described previously (Okamoto et al., 1990c). The complete nucleotide sequence was determined for HC-J6. In addition, the 5' and 3' end sequences were determined for isolates HC-J1 and HC-J4 (Okamoto et al., 1990c), as well as for isolate HC-J7 obtained from the serum of a patient on maintenance haemodialysis who was positive for anti-HCV.

Determination of 5' and 3' end sequences. cDNA was synthesized with an antisense primer [no. 36: nucleotides (nt) 245 to 264] deduced from the 5' non-coding regions of isolates HC-J1 and HC-J4 (Okamoto et al., 1990b). Excess primers were removed, and the cDNA was targeted with dATP by terminal deoxynucleotidyl transferase in a DNA tailing kit (Boehringer Mannheim). It was amplified by one-sided PCR with oligo(dT)20 and an antisense primer (no. 109: nt 140 to 159) positioned within no. 36 following the method of Ohara et al., 1989), and the 5' end sequence was determined. HCV RNA was polyadenylated using Escherichia coli poly(A) polymerase (Takara Biochemicals) and reverse-transcribed to cDNA in the presence of oligo(dT)20. The cDNA was amplified by one-sided PCR with oligo(dT)20 and a sense primer (no. 90: nt 9221 to 9240), and the 3' end sequence was determined.

Cloning and sequencing of cDNA. PCR products were separated by electrophoresis on a composite gel of 0.5 to 1.5% (w/v) NuSieve and 0.5% agarose (FMC BioProducts) and a portion of gel was cut out, corresponding to the position of PCR products with the expected size. DNA was eluted from it and treated with T4 DNA polymerase and T4 polynucleotide kinase (Takara Biochemicals). The 5'-phosphorylated PCR products, as well as cDNA inserts excised from λgt10, were cloned into M13mp18 or -mpl9 phage vectors, and overlapping cDNA species sequenced for both plus and minus strands by the dideoxynucleotide chain termination method (Sanger et al., 1977).

The strategy for sequencing HC-J6 is depicted in Fig. 1. The extreme 5' and 3' end sequences were determined on 13 and 19 cDNA clones, respectively, and consensus sequences were obtained. Three clones each of the PCR products were independently isolated, corresponding to the 5'-terminal region (nt 44 to 846, nt 731 to 1905 and nt 1847 to 2571) or 3'-terminal region (nt 8324 to 9261), and the consensus nucleotide sequence was determined. They had < 2.4% silent and ≤ 0.9% missense mutations. Although the 5' and 3'-terminal sequences of HC-J6 were amplified by PCR with primers deduced from reported isolates, this approach was not feasible for the central region of HC-J6, probably because of a high degree of sequence variation. The central region (nt 2552 to 8680), therefore, was determined from the position of PCR products with the expected size. DNA was eluted from it and treated with T4 DNA polymerase and T4 polynucleotide kinase (Takara Biochemicals). The overlapping cDNA species sequenced for both plus and minus strands by the dideoxynucleotide chain termination method (Sanger et al., 1977). Among the previously reported isolates, HCV-J (Kato et al., 1990) has 9413 nt for 3010 aa, HCV-1 (Choo et al., 1991) has 9379 nt for 3011 aa, and HCV-BK (Takamizawa et al., 1991) has 9416 nt for 3010 aa; HCV-J and HCV-BK possess a deletion of 3 nt in the NS5 region. The ORF of HC-J6 was longer than any of them. In comparison with the three reported isolates, HC-J6 had three insertions of 3, 3 and 6 nt, respectively, and two insertions of 12 and 60 nt as well as two deletions of 12 and 3 nt (6 nt compared with HCV-1) in the NS5 region (Fig. 1 and 2).

The four isolates were compared to find similarity in nt and aa sequences (Table 1). HC-J6 was different from each of the other three isolates, sharing > 70% of its nt and aa sequences with them. HCV-J was close to HCV-BK, similar in > 90% of both nt and aa sequences.

Results and Discussion

The nucleotide sequence of the cDNA of an HCV isolate, named HC-J6, is shown in Fig. 2. The HC-J6 cDNA contained 9481 nt and an additional T stretch of 30 to 108 nt at the 3' end. It had a large, single open reading frame (ORF) of 9099 nt, capable of coding for 3033 amino acids (aa). It had a base composition of 20.8% adenine, 21.7% thymine, 29.7% cytosine and 27.8% guanine.

Among the previously reported isolates, HCV-J (Kato et al., 1990) has 9413 nt for 3010 aa, HCV-1 (Choo et al., 1991) has 9379 nt for 3011 aa, and HCV-BK (Takamizawa et al., 1991) has 9416 nt for 3010 aa; HCV-J and HCV-BK possess a deletion of 3 nt in the NS5 region. The ORF of HC-J6 was longer than any of them. In comparison with the three reported isolates, HC-J6 had three insertions of 3, 3 and 6 nt, respectively, in the NS1 region, and two insertions of 12 and 60 nt as well as two deletions of 12 and 3 nt (6 nt compared with HCV-1) in the NS5 region (Fig. 1 and 2).

The four isolates were compared to find similarity in nt and aa sequences (Table 1). HC-J6 was different from each of the other three isolates, sharing > 70% of its nt and aa sequences with them. HCV-J was close to HCV-BK, similar in > 90% of both nt and aa sequences.
HCV-1 was closer to HCV-J and HCV-BK (>80% identity) than to HC-J6. On the basis of sequence subgroups of the HCV genome, similarity therefore, there would be at least three pestiviruses (Miller & Purcell, 1990; Okamoto et al., 1990c), and have now been extended to each region was compared for its degree of similarity (Table 1).

Sequence divergences were not evenly distributed over the entire genome. Nucleotide sequences of the four isolates were divided into the 5' non-coding region, C (core) and E (envelope) genes, NS1-NS5 regions and the 3' non-coding region (Fig. 1), as has been proposed due to the high degree of sequence conservation of the 5' non-coding region (>93%), primers deduced from this area could be used for the sensitive detection of HCV RNA by PCR (Okamoto et al., 1990b).

HC-J6 possessed a 5' non-coding region of 340 nt.All 13 clones of HC-J6, covering the extreme 5' end (nt 1 to 159), displayed a deletion of one C from a series of C residues at nt 4 to 8. Eight of them displayed an identical sequence, which was adopted as the consensus sequence. Minor sequence variations were observed in the remaining five clones. Two of them had a C-to-T mutation at nt 100, A-to-G mutation at nt 36, or insertion of one C into a C stretch at nt 118. The Y-terminal nt of HC-J6 was invariably G.

Characterization of the coding regions of HC-J6 RNA is given in Table 1. The NS1 and NS2 protein-encoding regions were predominantly conserved in HC-J6 RNA, as shown by the degree of identity among the four isolates and the consensus sequence. For the NS3 protein-encoding region, which consists of 609 aa, the degree of identity among the four isolates was 92% (accession numbers D00830, D00831, and D00832), and those of HCV-J and HC-J6 RNA were 78%, respectively.

The 13 clones of HC-J6 RNA with the highest degree of identity to the consensus sequence could be divided into five groups, as shown in Table 1. These groups were designated according to the 10 to 20 aa residues at nt 4 to 8. Eight of them displayed an identical sequence, which was adopted as the consensus sequence for the 5' non-coding region of the HCV genome would be postulated as being 341 nt. Due to the high degree of sequence conservation of the 5' non-coding region (>93%), primers deduced from this area could be used for the sensitive detection of HCV RNA by PCR (Okamoto et al., 1990b).

HC-J6 possessed a 5' non-coding region of 340 nt. All 13 clones of HC-J6, covering the extreme 5' end (nt 1 to 159), displayed a deletion of one C from a series of C residues at nt 4 to 8. Eight of them displayed an identical sequence, which was adopted as the consensus sequence. Minor sequence variations were observed in the remaining five clones. Two of them had a C-to-T mutation at nt 100, A-to-G mutation at nt 36, or insertion of one C into a C stretch at nt 118. The Y-terminal nt of HC-J6 was invariably G.

HCV-1 was closer to HCV-J and HCV-BK (>80% identity) than to HC-J6. On the basis of sequence similarity therefore, there would be at least three subgroups of the HCV genome.

Sequences divergences were not evenly distributed over the entire genome. Nucleotide sequences of the four isolates were divided into the 5' non-coding region, C (core) and E (envelope) genes, NS1-NS5 regions and the 3' non-coding region (Fig. 1), as has been proposed due to the high degree of sequence similarity of the 5' non-coding region (>93%), primers deduced from this area could be used for the sensitive detection of HCV RNA by PCR (Okamoto et al., 1990b).

HC-J6 possessed a 5' non-coding region of 340 nt. All 13 clones of HC-J6, covering the extreme 5' end (nt 1 to 159), displayed a deletion of one C from a series of C residues at nt 4 to 8. Eight of them displayed an identical sequence, which was adopted as the consensus sequence. Minor sequence variations were observed in the remaining five clones. Two of them had a C-to-T mutation at nt 100, A-to-G mutation at nt 36, or insertion of one C into a C stretch at nt 118. The Y-terminal nt of HC-J6 was invariably G.
The 3′ non-coding regions of HC-J6, HCV-J, HCV-1 and HCV-BK, along with those of HC-J1, HC-J4 and HC-J7 (accession numbers D00825, D00826 and D00828, respectively), were compared in Fig. 3. The 3′ non-coding region is markedly diverse both in length (39 to 45 nt) and sequence (>29-63% identity), in remarkable contrast to the high sequence conservation observed in the 5′ non-coding region. The number of nt upstream of the T stretch was uniform among clones from each isolate: 42 nt in 19 clones of HC-J6, 45 nt in 16 clones from HC-J1, 41 nt in 13 clones from HC-J4, and 39 nt in 9 clones from HC-J7.

No preferences were observed for the termination codon. It was TAG for HC-J6 and HC-J7, TGA for HC-J4, HCV-J, HCV-1 and HCV-BK, and TAA for HC-J1. A second in-phase termination codon was displayed by all isolates. This was positioned at 12 nt after the first termination codon in HC-J6 and HC-J7, and at 9 nt after the first in HC-J1, HC-J4, HCV-J, HCV-1 and HCV-BK.

The T stretch at the extreme 3′ end was variable in length, with occasional T-to-C point mutations. The extreme 3′ end of HCV-J has a stretch of 12 T residues ending with C, which would represent such a T-to-C

Table 1. Two-by-two comparison for the identity (%) in nucleotide and amino acid sequences within nine different regions among four HCV isolates*

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>5′ Non-coding</td>
<td>93.0</td>
<td>94.4</td>
<td>93.7</td>
<td>97.9</td>
<td>99.1</td>
<td>98.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>81.7</td>
<td>81.0</td>
<td>82.5</td>
<td>90.8</td>
<td>95.8</td>
<td>90.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>59.4</td>
<td>62.5</td>
<td>59.2</td>
<td>75.5</td>
<td>90.6</td>
<td>74.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NS1</td>
<td>68.8</td>
<td>66.9</td>
<td>69.6</td>
<td>72.5</td>
<td>86.6</td>
<td>71.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NS2</td>
<td>59.1</td>
<td>56.6</td>
<td>58.7</td>
<td>71.4</td>
<td>91.6</td>
<td>71.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NS3</td>
<td>70.4</td>
<td>70.3</td>
<td>70.2</td>
<td>79.8</td>
<td>90.6</td>
<td>78.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NS4</td>
<td>67.1</td>
<td>66.6</td>
<td>67.3</td>
<td>79.1</td>
<td>92.0</td>
<td>79.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NS5</td>
<td>66.5</td>
<td>66.3</td>
<td>65.8</td>
<td>79.0</td>
<td>92.9</td>
<td>79.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3′ Non-coding</td>
<td>38.1</td>
<td>29.6</td>
<td>42.9</td>
<td>66.7</td>
<td>90.5</td>
<td>70.4</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* HC-J6, HCV-J (Kato et al., 1990), HCV-1 (Choo et al., 1991) and HCV-BK (Takamizawa et al., 1991) were compared for nucleotide sequence identities within the whole genome (a stretch of U residues at the extreme 3′ end excluded), as well as in various regions of the genome. Percentages of amino acids identical within the polyprotein precursor and seven putative proteins processed from it (presumed from the similarity of HCV to flaviviruses) are given in parentheses.

The 3′ non-coding regions of HC-J6, HCV-J, HCV-1 and HCV-BK, along with those of HC-J1, HC-J4 and HC-J7 (accession numbers D00825, D00826 and D00828, respectively), are compared in Table 1. The T stretch at the extreme 3′ end was variable in length, with occasional T-to-C point mutations. The T stretch of HCV-J has a stretch of 12 T residues ending with C, which would represent such a T-to-C

Fig. 3. Nucleotide sequences of cDNAs for the 3′ non-coding regions of HCV isolates. The sequences are shown for the four isolates whose entire nucleotide sequences have been determined, and also for the HC-J1, HC-J4 and HC-J7 isolates, sequences of which have been partially determined. Stretches of T residues are indicated in parentheses, with numbers that vary even among clones from the same isolate.

Mutation. The length of the T stretch varied even among clones from the same isolate. It spanned 30 to 108 nt in 19 clones from HC-J6, 11 to 83 nt in 16 clones from HC-J1, 9 to 36 nt in 13 clones from HC-J4 and 21 to 92 nt in nine clones from HC-J7. Poly(A) tails proposed for HCV-1 (Choo et al., 1991; Han et al., 1991) were not confirmed in any of these other isolates. Despite the resemblance of HCV to flaviviruses and pestiviruses, no significant homology has been observed between the 3′ non-coding sequence of HCV and that of flaviviruses (Chambers et al., 1990) or pestiviruses (Collett et al., 1988; Meyers et al., 1989; Moormann et al., 1990), either in length or in sequence. In addition, flaviviruses and pestiviruses do not possess such a long poly(U) sequence at the 3′ end as that observed in HCV.

Among amino acid sequences of seven putative structural and non-structural proteins, the conservation was highest for the C protein (> 90%), followed by the NS3 protein (> 81%). The sequence of NS2 protein was the most divergent, being only 58% identical. This low level of variation in the C protein will allow the design of antigen probes, and could become very important in the development of serological tests for the diagnosis of HCV infection, using synthetic oligopeptides (Okamoto et al., 1990a) and proteins obtained by recombinant gene technology (Muraiso et al., 1990; van der Poel et al., 1991).

The putative E protein was divergent, with a sequence identity of > 53%, and the sequence of the NS1 protein was > 71% identical. Different subgroups of HCV could be classified according to their sequence divergences (up to 47%) in the E protein, which exceed those (up to 40%) found among the four serotypes of dengue virus (Chambers et al., 1990). HCV subgroups would be expected to have distinct epidemiological distributions, and these would influence the choice of materials best suited for diagnosis and immunoprophylaxis in the geographical area at issue.
The E protein would be of practical importance, since it presents a target in the immunoprophylaxis of HCV infection, whether by specific immunoglobulins or by vaccines. Eight cysteine residues are conserved in the E protein of the four isolates, suggesting that the epitopes borne by it may well be dependent on the molecular conformation. Should this be the case, synthetic oligopeptides deduced from the E protein sequence might not work as efficiently, as antigen probes, as those from the C protein, in which only three cysteine residues are conserved (Okamoto et al., 1990c). Future efforts to characterize the envelope protein, as well as the product of the NS1 region, will be instrumental in the development of HCV vaccines, as has been shown for flavivirus vaccines (Heinz, 1986).

Sequence similarities to flaviviruses and pestiviruses, as pointed out in previous studies (Kato et al., 1990; Miller & Purcell, 1990; Choo et al., 1991; Takamizawa et al., 1991; Han et al., 1991), were recognized in HC-J6 also (Fig. 2). They included a potential hairpin structure in the 5′ non-coding region, palindromic structures in the 3′-terminal region of the ORF, and amino acid motifs conserved among putative trypsin-like serine proteases or NTP-binding helicases in the NS3 protein (Bazan & Fletterick, 1989; Lain et al., 1989; Moermann et al., 1990). Also observed was the Gly-Asp-Asp triplet in the NS5 protein that is commonly displayed by viral RNA-dependent RNA polymerases (Kamer & Argos, 1984).

The nucleotide sequence of HC-J6 presented here has revealed a degree of variation of the HCV genome which is much higher than that seen in a comparison of sequences of the three isolates reported previously (Kato et al., 1990; Choo et al., 1991; Takamizawa et al., 1991). The wide range of HCV sequence variation would reflect its nature as an RNA virus, typified by influenza A and human immunodeficiency viruses (Buonaguro et al., 1986; Hahn et al., 1986; Steinhauser & Holland, 1987). Obviously, additional HCV isolates need to be sequenced to characterize fully the range of variations of this virus. The knowledge gathered will shed light on the global distribution of various HCV types, and help to establish routes of transmission, besides those of transfusion and drug abuse, which maintain HCV in the community.

References
A structural protein of hepatitis C virus expressed in *E. coli* facilitates accurate detection of hepatitis C virus. Biochemical and Biophysical Research Communications 172, 511-516.


Infectivity of blood seropositive for hepatitis C virus antibodies. Lancet 335, 558-560.

Confirmation of hepatitis C virus infection by new four-antigen recombinant immunoblot assay. Lancet 337, 317-319.

(Received 20 May 1991; Accepted 29 July 1991)