Proc. Natl. Acad. Sci. USA
Vol. 91, pp. 8239–8243, August 1994
Biochemistry

Sequence analysis of the core gene of 14 hepatitis C virus genotypes

(non-A, non-B hepatitis/genetic heterogeneity/polymerase chain reaction/phylogenetic tree/taxonomy)

JENS BUKH, ROBERT H. PURCELL, AND ROGER H. MILLER

Hepatitis Viruses Section, Laboratory of Infectious Diseases, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD 20892

Contributed by Robert H. Purcell, May 17, 1994

ABSTRACT **We previously sequenced the 5′ noncoding region of 44 isolates of hepatitis C virus (HCV), as well as the envelope 1 (E1) gene of 51 HCV isolates, and provided evidence for the existence of at least 6 major genetic groups consisting of at least 12 minor genotypes of HCV (i.e., genotypes I/1a, II/1b, III/2a, IV/2b, 2c, V/3a, 4a–4d, 5a, and 6a). We now report the complete nucleotide sequence of the putative core (C) gene of 52 HCV isolates that represent all of these 12 genotypes as well as two additional genotypes provisionally designated 4e and 4f that we identified in this study. The phylogenetic analysis of the C gene sequences was in agreement with that of the E1 gene sequences. A major division in the genetic distance was observed between HCV isolates of genotype 2 and those of the other genotypes in analysis of both the E1 and C genes. The C gene sequences of 9 genotypes have not been reported previously (i.e., genotypes 2c, 4a–4f, 5a, and 6a). Our analysis indicates that the C gene-based methods currently used to determine the HCV genotype, such as PCR with genotype-specific primers, should be revised in light of these data. We found that the predicted C gene was exactly 573 nt long in all 52 HCV isolates, with an N-terminal start codon and no in-frame stop codons. The nucleotide and predicted amino acid identities of the C gene sequences were in the range of 79.4–99.0% and 85.3–100%, respectively. Furthermore, we mapped universally conserved, as well as genotype-specific, nucleotide and deduced amino acid sequences of the C gene. The predicted C proteins of the different HCV genotypes shared the following features: (*i*) high content of proline residues, (*ii*) high content of arginine and lysine residues located primarily in three domains with 10 such residues invariant at positions 39–62, (*iii*) a cluster of 5 conserved tryptophan residues, (*iv*) two nuclear localization signals and a DNA-binding motif, (*v*) a potential phosphorylation site with a serine-proline motif, and (*vi*) three conserved hydrophilic domains that have been shown by others to contain immunogenic epitopes. Thus, we have extended analysis of the predicted C protein of HCV to all of the recognized genotypes, confirmed the existence of highly conserved regions of this important structural protein, and demonstrated that the genetic relatedness of HCV isolates is equivalent when analyzing the most conserved (i.e., C) and the most variable (i.e., E1) genes of the HCV genome.**

Hepatitis C virus (HCV) is an important human pathogen that can cause acute and chronic hepatitis, liver cirrhosis, and, possibly, hepatocellular carcinoma. The virus particles contain a positive polarity, single-stranded RNA genome with 5′ and 3′ noncoding (NC) regions. The core (C), envelope 1 (E1), and envelope 2 proteins are encoded at the 5′ terminus and the nonstructural proteins are encoded at the 3′ terminus of the single open reading frame of the genome (1, 2). It is now well established that there are a number of different genotypes of HCV, which may have important implications for pathogenesis, diagnosis, and vaccine development. Based on analysis of HCV isolates sequenced in their entirety, Okamoto *et al.* (3) demonstrated that all previously published sequences could be grouped into four such genotypes [i.e., genotypes I/1a (2), II/1b (4), III/2a (5), and IV/2b (3)]. At approximately the same time, our analysis of the highly conserved 5′ NC region of 44 HCV isolates from around the world suggested the existence of additional genetic groups of HCV (6). Our subsequent analysis of the highly variable E1 gene from 51 HCV isolates (7) confirmed the presence of 12 genotypes divided into at least six major genetic groups and accompanying subgroups (i.e., genotypes I/1a, II/1b, III/2a, IV/2b, 2c, V/3a, 4a–4d, 5a, and 6a). Genotype V/3a isolates have also been found by others (8–12). Subsequently, Simmonds *et al.* (13) confirmed the existence of multiple major genetic groups of HCV by sequence analysis of a short region within the NS-5 gene. However, until definitive overlapping sequences are available, the genetic relatedness of HCV isolates designated genotypes 4, 5, and 6 by Simmonds and coworkers to our isolates of genotypes 4a–4d, 5a, and 6a cannot be determined. Taken together, it is clear that there are at least 12 genotypes of HCV.

In this report, we have designated the various genotypes by the nomenclature proposed by Okamoto *et al.* (3) and by Chan *et al.* (9). However, as we have pointed out (7), the proposed classification schemes should be considered provisional until more data are obtained. In the present study, we have determined the complete nucleotide sequence of the C gene in 52 HCV isolates* that represent the 12 recognized HCV genotypes as well as two additional genotypes, 4e and 4f, identified in this study.

MATERIALS AND METHODS

Sera analyzed in this study were from 52 individuals from 12 countries, who were positive for antibodies to HCV (anti-HCV) by a first-generation test (14). The consensus 5′ NC and E1 gene sequences of the HCV RNA from these sera were previously analyzed (6, 7). In this study we have analyzed the consensus C gene sequence of these same HCV isolates. The procedures that were used for viral RNA extraction, cDNA synthesis, and nested PCR have been described (14). For the cDNA PCR assay, we used HCV-specific synthetic oligonucleotides deduced from previously determined sequences that flank the C gene (6, 7). In 51 of the 52 HCV isolates studied, we amplified the entire C gene and adjacent 5′ NC and E1 sequences. However, in isolate Z7 the 36 nt at the 3′ end of C were from a second DNA fragment that we obtained previ-

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: HCV, hepatitis C virus; NC, noncoding; C, core; E1, envelope 1.

*The sequences reported in this paper have been deposited in the GenBank data base (accession nos. U10189–U10240).

ously (7). Amplified DNA was purified by gel electrophoresis followed by glass-milk extraction (7) or electroelution and both strands were sequenced directly. In 44 of the 52 HCV isolates studied, we used the procedures for direct sequencing described previously (7). For a number of the HCV isolates, confirmatory sequencing was performed with the Applied Biosystems automated DNA sequencer (model 373A) and 8 HCV isolates of genotype I/1a or II/1b were sequenced exclusively by this method. Multiple sequence alignments were performed with the program GENALIGN (15). Phylogenetic trees were constructed by the unweighted pair-group method with arithmetic mean that is based on the assumption of a constant rate of evolution (16).

RESULTS AND DISCUSSION

In this study, we successfully reverse-transcribed and, by PCR, amplified the entire C gene from HCV isolates representing the 12 genotypes identified by analysis of E1 sequences (7) and from all of the HCV isolates with unique 5′ NC sequences (6) that we could not amplify in the previous study of E1. All 73 negative control samples interspersed among the test samples were negative for HCV RNA. The amplified DNA fragment obtained in 50 of the 52 HCV isolates was specifically designed to overlap with our previously obtained 5′ NC-C sequences (6) and C-E1 sequences (7) at ≈80 nt positions each. A complete match was observed in 6033 of 6035 overlapping nt. Two discrepancies were observed in isolate US6 at nt 552 (C and T) and 561 (C and T). This may have been due to microheterogeneity at these nucleotide positions, since the remaining overlapping sequence was unique for isolate US6. In addition, there were three confirmed instances of microheterogeneity: nt 33 in isolate SA11 (C,T, and T), nt 36 in isolate S45 (A,C, and A), and nt 552 in isolate P10 (C,T, and T). Overall, the excellent agreement in these overlapping sequences in this study with that of the two previous studies definitively ruled out contamination as a source of nonauthentic HCV sequences. Furthermore, this analysis proved that the sequences obtained were from a single population and not from different populations as could happen in mixed infections.

Analysis of the Nucleotide Sequence of the C Gene. We now report the nucleotide (nt 1–573) and deduced amino acid (aa 1–191) sequences of the putative C gene of 52 HCV isolates. Relative to the prototype sequence (1, 2), we found that the C gene was exactly 573 nt long in all 52 HCV isolates with an N-terminal start codon and no in-frame stop codons. Microheterogeneity, defined previously (7), was observed in 26 of the 52 HCV isolates at 0.2–1.4% of the 573 nucleotide positions of the C gene and resulted in changes in 0.5–1.0% of the 191 predicted aa in 12 of these isolates. We performed a multiple sequence alignment (data not shown) and found that the nucleotide identities of the C gene among these HCV isolates were in the range 79.4–99.0%. Since we were interested in comparing the genetic relatedness of HCV isolates in different gene regions we constructed phylogenetic trees of the C gene of all 52 HCV isolates from this study and the E1 gene of 51 HCV isolates from our previous study (7) using the unweighted pair-group method with arithmetic mean (16) (Fig. 1). In both dendrograms we observed a division of the

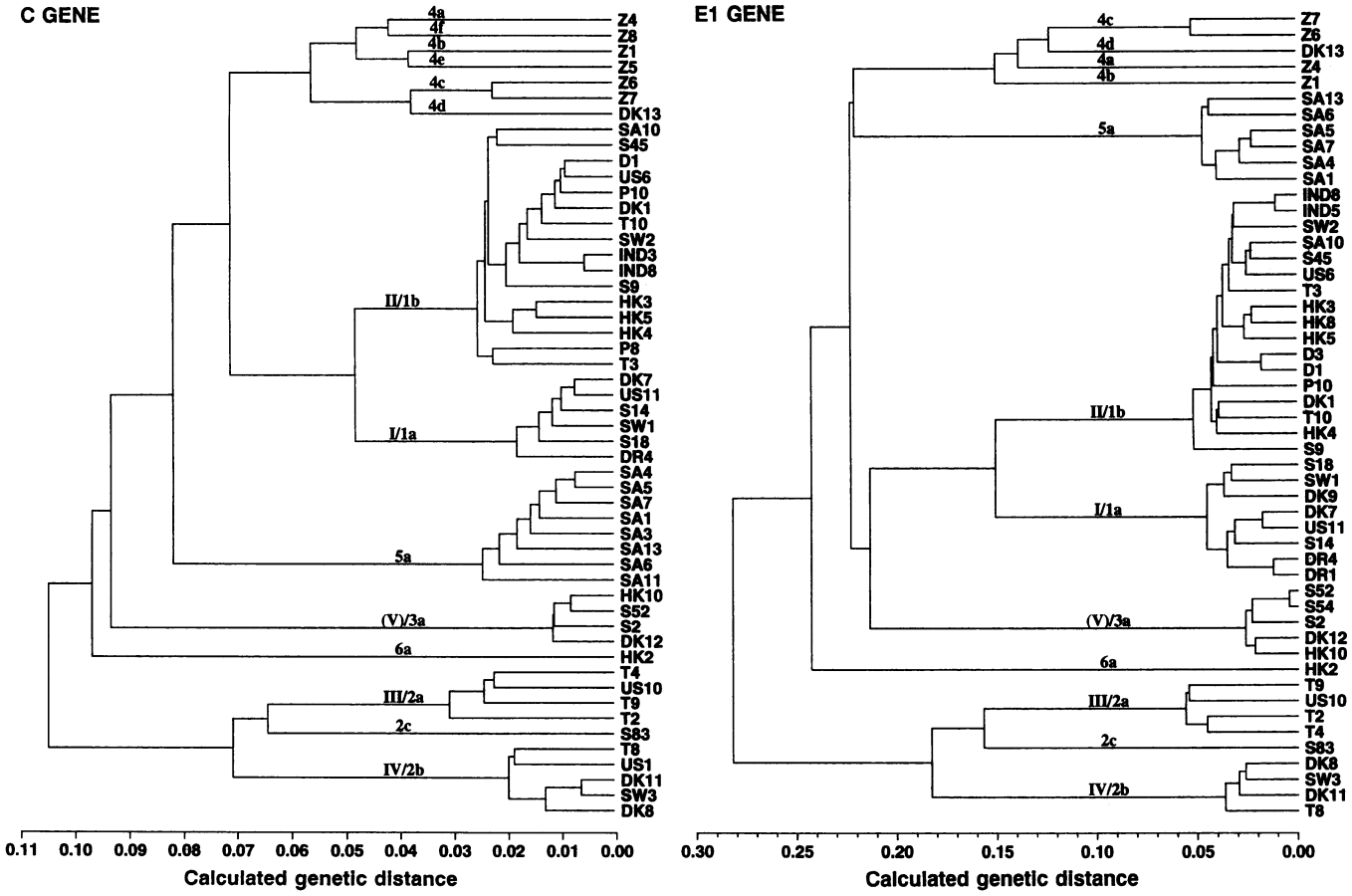

FIG. 1. Phylogenetic trees showing calculated evolutionary relationships of the different HCV isolates based on the C gene sequence of 52 HCV isolates and the E1 gene sequence of 51 HCV isolates. Phylogenetic trees were constructed by the unweighted pair-group method with arithmetic mean (16) using the computer software package GENEWORKS from IntelliGenetics. Lengths of the horizontal lines connecting the sequences, given in absolute values from 0 to 1, are proportional to the estimated genetic distances between the sequences. Genotype designations of HCV isolates are indicated. In 45 HCV isolates, we determined both the C and the E1 gene sequences.

45 HCV isolates that were common to the two studies into at least six major genetic groups (genotypes 1–6) and 12 minor genetic groups (genotypes I/1a, II/1b, III/2a, IV/2b, 2c, V/3a, 4a–4d, 5a, and 6a). It is noteworthy that we observed a major division in genetic distance between HCV isolates of genotype 2 and those of the other genotypes in the phylogenetic analyses of both gene sequences. Furthermore, the divergence of the minor genotypes within genotype 2 exhibits a degree of heterogeneity that is equivalent to that observed among the major genotypes. Analysis of the C gene from isolates Z5 and Z8, which had a unique 5′ NC sequence (6) but from which we could not amplify the E1 gene, revealed that these isolates represented two additional genotypes. We are provisionally assigning designations 4e and 4f to these genotypes that have not been described previously. Although Simmonds et al. (17) have published partial C gene sequences (i.e., nt 29–269) of HCV isolates that appear to be most closely related to our isolates of major genotype 4, final classification of these isolates must await completion of the gene sequence. Unfortunately, a sequence motif within the C gene (i.e., nt 186–221) that has been suggested to be predictive of genotype (8) does not reflect the genotype divisions observed by our analysis of the complete C gene. Overall, we have demonstrated that the genetic relatedness of HCV isolates is equivalent when analyzing the most conserved gene (C) and one of the most variable genes (E1) of the HCV

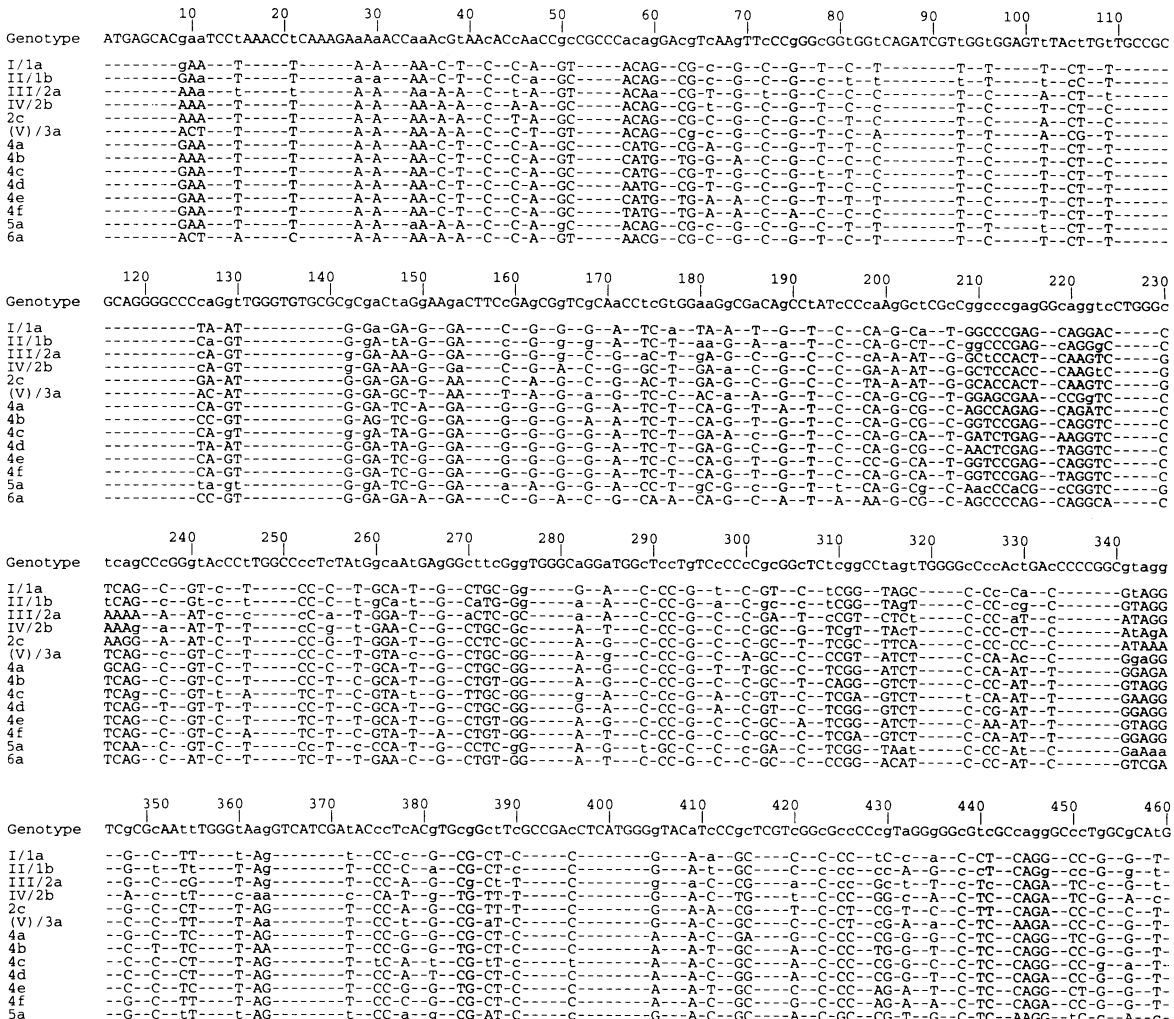

FIG. 2. Alignment of the consensus sequence of the C gene of the different genotypes of HCV. Consensus sequence of the C gene from all 52 HCV isolates studied is shown at the top. Furthermore, a consensus sequence of the C gene was obtained for genotypes I/1a, II/1b, III/2a, IV/2b, 3a, and 5a. The sequence of genotype 4c is represented by isolate Z6. Genotypes 4a, 4b, 4d, 4e, 4f, and 6a each contained only a single isolate. The exact HCV isolates representing the different genotypes can be seen in the phylogenetic tree of the C gene sequences in Fig. 1. Invariant nucleotides within a consensus sequence are capitalized and variable nucleotides are shown in lowercase letters. However, nucleotides that were invariant among all 52 HCV isolates are shown as dashes in the alignment. In the 14 nt positions where no consensus sequence was obtained, we show the nucleotide that differed from that of the other genotypes.

genome, providing strong evidence for the suggested division into major and minor genotypes.

To study further the heterogeneity of the C gene, we obtained the consensus sequence of this gene from the 52 HCV isolates (Fig. 2). We found that a total of 335 (58.5%) of the 573 nucleotides of the C gene were invariant among these HCV isolates. Nucleotides at the first and second codon positions were invariant at 70.7% and 81.7% of these positions, respectively, while nucleotides at the third position were invariant at only 23.0% of such positions. Stretches of 6 or more invariant nt were observed from nt 1–8, 22–27, 85–92, 110–125, 131–141, 334–340, 364–371, 397–404, and 511–516 and may be suitable for anchoring primers for amplification of HCV RNA in cDNA PCR assays. Finally, we documented the genotype-specific sequences within the C gene by aligning the consensus sequences of all 14 genotypes (Fig. 2). Although the full-length sequence of the C gene of isolates representing genotypes I/1a, II/1b, III/2a, IV/2b, and V/3a have been reported by others (2–5, 11), those of 9 of the 14 genotypes (i.e., 2c, 4a–4f, 5a, and 6a) have not been reported previously. Overall, we have mapped universally conserved sequences as well as genotype-specific sequences of the C gene among 14 genotypes of HCV.

Analysis of the Deduced Amino Acid Sequence of the C Gene. To study the heterogeneity of the C protein, we performed a multiple sequence alignment of the predicted amino acids for all 52 HCV isolates (data not shown) and obtained a consensus sequence (Fig. 3). The identities of the predicted 191 aa of the C protein among these HCV isolates were in the range 85.3–100.0%. A total of 132 (69.1%) of the 191 aa of the C protein were invariant. The most prevalent amino acids in the consensus sequence were glycine (13.6%), arginine (12.6%), proline (11.0%), and leucine (9.9%). The most conserved amino acids were tryptophan (5 of 5 aa invariant), aspartic acid (5 of 5 aa invariant), proline (19 of 21 aa invariant), and glycine (23 of 26 aa invariant). Previous analyses indicated that HCV is evolutionarily related to pestiviruses (15). In this regard, it is of interest to note that the C proteins of both viruses have a high content of proline residues (18), which are likely to be important in maintaining the structure of this protein. As is characteristic for a protein that binds to nucleic acid, we found that the C protein has conserved amino acids that are basic and positively charged, and these are capable of neutralizing the negative charge of the HCV RNA encapsidated by this protein (19). Specifically, >16% of the amino acids in the consensus sequence of the C protein of HCV are arginine and lysine that are located primarily in three clusters (i.e., from aa 6–23, 39–74, and 101–121) (20) (Fig. 3). The 10 arginine and lysine residues within aa 39–62 are invariant among all 52 HCV isolates, suggesting that this domain may represent an important RNA-binding site. The capsid proteins of the related flavi- and pestiviruses (15) also have a high content of arginine and lysine (18, 19). Although there are three major hydrophilic regions (i.e., aa 2–23, 39–74, and 101–121) that are conserved in all 52 HCV isolates, the remainder of the C protein is hydrophobic. Interestingly, one such highly conserved hydrophobic domain at aa 24–39 is flanked by proline residues. The hydrophobic domains are likely to be involved in protein–protein and/or protein–RNA interactions during assembly of the nucleocapsid as well as in interaction with the lipoprotein envelope, as has been suggested for flaviviruses (19). Other significant observations are (*i*) a cluster of 5 invariant tryptophan residues at aa 76–107; (*ii*) the lack of an N-linked glycosylation site (NXT/S); (*iii*) two potential nuclear localization signals (i.e., PRRGPR at aa 38–43 and PRGRRQP at aa 58–64) that are present in all 52 HCV isolates (20); and (*iv*) a putative DNA-binding motif SPRG at aa 99–102, found in 51 of the 52 HCV isolates, with SP present in all 52 isolates. Our finding of conserved nuclear localization signals and a DNA-binding motif adds support to the hypothesis that the C protein of HCV might also function as a gene-regulatory protein (20). Furthermore, it has been suggested that the HCV C protein is posttranslationally modified through phosphorylation (20, 21). Interestingly, we found that the C protein of all 52 HCV isolates contained a SP motif that was recently demonstrated to be essential for C protein phosphorylation in hepadnaviruses (22). Our study demonstrates that the C protein has features that are highly conserved among the various genotypes of HCV and that are known to be characteristic of capsid proteins of other related viruses.

To study the heterogeneity of the C protein of different genotypes, we obtained the consensus sequence of the protein for all isolates comprising the 14 HCV genotypes (Fig. 3). We mapped the genotype-specific sequences within the C protein by then aligning these consensus sequences (Fig. 3). It should be noted that phylogenetic analysis of the amino acid sequence of the C proteins was not capable of resolving the minor groups within genotypes 1 and 4 because of the conservation of this protein (data not shown). Overall, we identified only a few type-specific amino acids (Fig. 3). One striking example was that isolates of genotype 4 have an additional methionine at position 20 that is specific for this major genetic group. Finally, we analyzed the conservation of the sequences surrounding the cleavage site between the C and the E1 proteins of the different genotypes, which has been determined to be between aa 191 (alanine) and 192 (tyrosine) in HCV isolates of genotype 1 (1). We previously found that the N-terminal amino acids of E1 were variable even within genotype 1 (7). In this study, we find that the

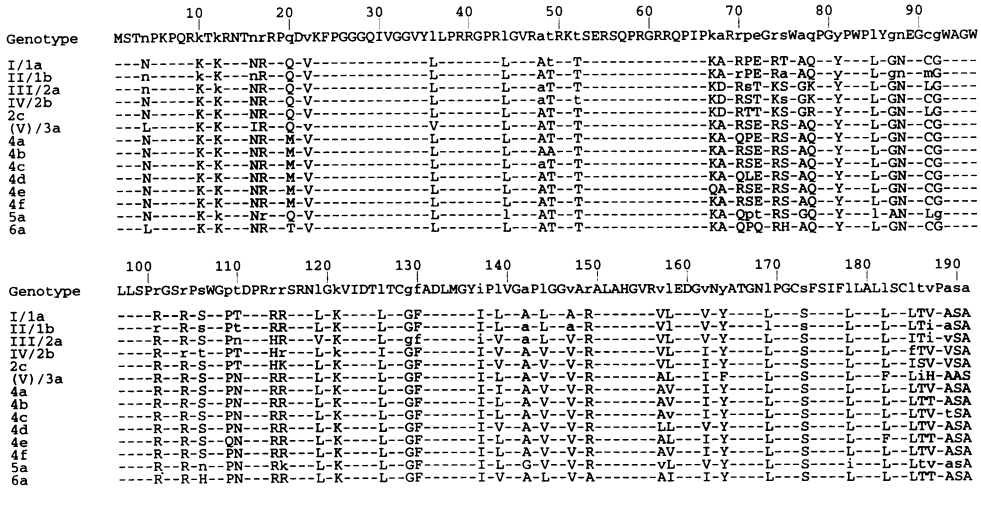

FIG. 3. Alignment of consensus sequence of deduced amino acid sequences of the C gene of the different genotypes of HCV. Consensus sequence of the C protein from all 52 HCV isolates studied is shown at the top. In the 2 aa positions where no consensus sequence was obtained, we show the amino acid that differed from that of other genotypes. See also legend to Fig. 2.

C-terminal sequence of C is SA in all but 1 of the 48 HCV isolates comprising genotypes 1, 2, 4, 5, and 6. However, all 4 HCV isolates of genotype 3 in this study, as well as isolates of genotype 3 published previously (11, 12), contain AS at this position. Thus, studies will be needed to determine the C/E1 cleavage site in genotype 3 isolates. Overall, we have mapped universally conserved sequences, as well as genotype-specific sequences, of the C protein among 14 genotypes of HCV.

Detection of antibodies directed against the HCV core protein is important in diagnosis of HCV infection. The recombinant C22-3 protein, spanning aa 2–120 of the C gene, is a major component of the commercially available second-generation anti-HCV tests. Several studies have indicated that the three major hydrophilic regions of the C protein contain linear immunogenic epitopes (summarized in ref. 23). For example, antibodies against synthetic peptides from aa 1–18, 51–68, and 101–118 were detected in infected patients (23). Our study demonstrates that, while these immunogenic regions are highly conserved, genotype-specific differences are observed at several amino acid positions that may influence the specificity and sensitivity of the serological tests (Fig. 3). One such example is that a single substitution at aa 110 has been demonstrated to affect seroreactivity (23). Despite the high degree of conservation in the immunodominant regions of the C protein among the different genotypes, it is possible that genetic heterogeneity of the C protein could lead to false-negative results in current serological tests.

Methods for Genotype Analysis. Several methods have been used to determine the genotype of HCV isolates without resorting to sequence analysis. These include PCR followed by (*i*) amplification with type-specific primers (24); (*ii*) determination of restriction-length polymorphism (17); and (*iii*) specific hybridization (25). The proposed methods have primarily been based on 5′ NC and C sequences. Our previous studies suggested that 5′ NC-based genotyping systems would be predictive of only the major genetic groups of HCV (6, 7). The most widely used C-based genotype system has been the PCR assay with type-specific primers that was designed for distinguishing HCV isolates of genotypes I/1a, II/1b, III/2a, IV/2b, and V/3a (11, 24). Since this system was developed before identification of genotypes 2c, 4a–4f, 5a, and 6a, there are significant limitations to this typing system. For example, the primers specific for genotype IV/2b (nt 270–251) are as highly conserved within our isolates of genotypes 4c and 6a as within the isolates of genotype IV/2b. Thus, this assay probably cannot distinguish among these genotypes. Another C-based approach involves distinguishing between genotypes 1 and 2 by type-specific antibody responses (26). Synthetic peptides composed of aa 65–81 were found to be genotype-specific for genotypes 1 and 2 in ELISAs. Our analysis of amino acid sequences demonstrated significant variation within isolates of genotypes 1 and 2. Thus, it is likely that these peptides will not identify all isolates of genotypes 1 and 2. Furthermore, the peptide for genotype 1 was highly conserved within isolates of genotypes 3 and 4 (Fig. 3) and might detect antibodies against these genotypes as well. It should be pointed out that most isolates of genotypes 3 and 4 had an identical amino acid sequence at positions 65–81. Overall, the proposed C-based genotyping systems should be revised in light of the C gene sequence data presented here, and a more definitive approach such as sequence analysis of gene regions that are predictive of genotype may be necessary for a definitive determination.

Conclusion. The genetic relatedness of HCV isolates is equivalent when analyzing the most conserved (i.e., C) and the most variable (i.e., E1) genes. The results of this study have implications for the taxonomy of HCV and for the diagnosis, prevention, and therapy of HCV infections.

We thank Ms. T. Tsareva for synthesis of oligonucleotides; Ms. R. Ashworth for assistance with DNA sequencing of reverse-transcribed PCR-amplified HCV RNA; and Mr. T. Heishman for assistance with computer analysis, graphics, and manuscript preparation. We gratefully acknowledge the generosity of the following investigators for providing sera for this study: Drs. H. Alter, V. Arankalle, D.-S. Chen, P. Farci, M. C. Kew, K. Krogsgaard, A. Lok, T. Quinn, F. Renger, M. Sjogren, and A. Widell. Sequencing with the Applied Biosystems automated DNA sequencer (model 373A) was performed at the DNA Analysis Facility, Center for Medical Genetics, The Johns Hopkins University School of Medicine and The John Hopkins Hospital. Computer-assisted nucleic acid analysis was through the GenBank Online Service, PC/GENE, and GeneWorks.

1. Hijikata, M., Kato, N., Ootsuyama, Y., Nakagawa, M. & Shimotohno, K. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 5547–5551.
2. Choo, Q.-L., Richman, K. H., Han, J. H., Berger, K., Lee, C., Dong, C., Gallegos, C., Coit, D., Medina-Selby, A., Barr, P. J., Weiner, A. J., Bradley, D. W., Kuo, G. & Houghton, M. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 2451–2455.
3. Okamoto, H., Kurai, K., Okada, S-I., Yamamoto, K., Lizuka, H., Tanaka, T., Fukuda, S., Tsuda, F. & Mishiro, S. (1992) *Virology* **188**, 331–341.
4. Kato, N., Hijikata, M., Ootsuyama, Y., Nakagawa, M., Ohkoshi, S., Sugimura, T. & Shimotohno, K. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 9524–9528.
5. Okamoto, H., Okada, S., Sugiyama, Y., Kurai, K., Iizuka, H., Machida, A., Miyakawa, Y. & Mayumi, M. (1991) *J. Gen. Virol.* **72**, 2697–2704.
6. Bukh, J., Purcell, R. H. & Miller, R. H. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 4942–4946.
7. Bukh, J., Purcell, R. H. & Miller, R. H. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 8234–8238.
8. Cha, T.-A., Beall, E., Irvine, B., Kolberg, J., Chien, D., Kuo, G. & Urdea, M. S. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 7144–7148.
9. Chan, S.-W., McOmish, F., Holmes, E. C., Dow, B., Peutherer, J. F., Follett, E., Yap, P. L. & Simmonds, P. (1992) *J. Gen. Virol.* **73**, 1131–1141.
10. Mori, S., Kato, N., Yagyu, A., Tanaka, T., Ikeda, Y., Petchclai, B., Chiewsilp, P., Kurimura, T. & Shimotohno, K. (1992) *Biochem. Biophys. Res. Commun.* **183**, 334–342.
11. Okamoto, H., Tokita, H., Sakamoto, M., Horikita, M., Kojima, M., Iizuka, H. & Mishiro, S. (1993) *J. Gen. Virol.* **74**, 2385–2390.
12. Stuyver, L., van Arnhem, W., Wyseur, A., DeLeys, R. & Maertens, G. (1993) *Biochem. Biophys. Res. Commun.* **192**, 635–641.
13. Simmonds, P., Holmes, E. C., Cha, T.-A., Chan, S.-W., McOmish, F., Irvine, B., Beall, E., Yap, P. L., Kolberg, J. & Urdea, M. S. (1993) *J. Gen. Virol.* **74**, 2391–2399.
14. Bukh, J., Purcell, R. H. & Miller, R. H. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 187–191.
15. Miller, R. H. & Purcell, R. H. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 2057–2061.
16. Nei, M. (1987) *Molecular Evolutionary Genetics* (Columbia Univ. Press, New York), pp 287–326.
17. Simmonds, P., McOmish, F., Yap, P. L., Chan, S.-W., Lin, C. K., Dusheiko, G., Saeed, A. A. & Holmes, E. C. (1993) *J. Gen. Virol.* **74**, 661–668.
18. Collett, M. S., Larson, R., Belzer, S. K. & Retzel, E. (1988) *Virology* **165**, 200–208.
19. Rice, C. M., Strauss, E. G. & Strauss, J. H. (1986) in *The Togaviridae and Flaviviridae*, eds. Schlesinger, S. & Schlesinger, M. J. (Plenum, New York, NY), pp. 279–326.
20. Shih, C.-M., Lo, S. J., Miyamura, T., Chen, S.-Y. & Lee, Y.-H. W. (1993) *J. Virol.* **67**, 5823–5832.
21. Lanford, R. E., Notvall, L., Chavez, D., White, R., Frenzel, G., Simonsen, C. & Kim, J. (1993) *Virology* **197**, 225–235.
22. Yu, M. & Summers, J. (1994) *J. Virol.* **68**, 2965–2969.
23. Sällberg, M., Rudén, U., Wahren, B. & Magnius, L. O. (1992) *J. Clin. Microbiol.* **30**, 1989–1994.
24. Okamoto, H., Sugiyama, Y., Okada, S., Kurai, K., Akahane, Y., Sugai, Y., Tanaka, T., Sato, K., Tsuda, F., Miyakawa, Y. & Mayumi, M. (1992) *J. Gen. Virol.* **73**, 673–679.
25. Stuyver, L., Rossau, R., Wyseur, A., Duhamel, M., Vanderborght, B., Van Heuverswyn, H. & Maertens, G. (1993) *J. Gen. Virol.* **74**, 1093–1102.
26. Machida, A., Ohnuma, H., Tsuda, F., Munekata, E., Tanaka, T., Akahane, Y., Okamoto, H. & Mishiro, S. (1992) *Hepatology* **16**, 886–891.