Positively selected amino acid sites in the entire coding region of hepatitis C virus subtype 1b

Yoshiyuki Suzuki, Takashi Gojobori*

Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, 1111 Yata, Mishima-shi, Shizuoka-ken 411-8540, Japan

Received 6 December 2000; received in revised form 10 April 2001; accepted 26 July 2001
Received by G. Bernardi

Abstract

To predict the amino acid sites important for the clearance of hepatitis C virus (HCV) subtype 1b in vivo, positively selected amino acid sites were detected by analyzing the sequence data collected from the international DNA databank. The rate of nonsynonymous substitutions per nonsynonymous site was compared with that of synonymous substitutions per synonymous site for each codon site in the entire coding region. As a result, 13 out of 3010 amino acid sites were found to be positively selected. Among the 13 positively selected amino acid sites, eight were located in the structural proteins and five were in the nonstructural proteins. Moreover, eight were located in B-cell epitopes and two were in T-cell epitopes. These observations suggest that both the antibody and the cytotoxic T lymphocyte are involved in the clearance of HCV subtype 1b in vivo. These positively selected amino acid sites represent candidate vaccination targets for HCV subtype 1b. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Hepatitis C virus; Synonymous substitution; Nonsynonymous substitution; Positive selection; Epitope; Vaccine

1. Introduction

Hepatitis C virus (HCV), the sole member of the genus Hepacivirus in the family Flaviviridae (van Regenmortel et al., 2000), is an enveloped, non-segmented, single-stranded, and positive-sense RNA virus (Choo et al., 1989). The genome of HCV is approximately 9.5 kilobases long, encoding a polyprotein of approximately 3000 amino acids (Kato et al., 1990). The polyprotein is co- and post-translationally cleaved into core protein (C), envelope glycoprotein 1 (E1), E2, p7, nonstructural protein 2 (NS2), NS3, NS4A, NS4B, NS5A, and NS5B in order from its N-terminus by the cellular signalase and the viral proteinases (Grakoui et al., 1993).

The genomic sequences from different HCV isolates are highly divergent (Kato et al., 1989). According to the phylogenetic analysis, HCV has been classified into six clades, in which various numbers of subtypes are included (Robertson et al., 1998). In the polyprotein of HCV, a region including the N-terminal 27–31 amino acid sites in E2 is known to be the most variable and is called hypervariable region 1 (HVR1; Hijikata et al., 1991; Weiner et al., 1991). HVR1 has been proposed as a major target of the immune response (Weiner et al., 1992; Farci et al., 1996; Zibert et al., 1997a). More than 50% of humans infected with HCV establish chronic hepatitis, which may progress to cirrhosis and hepatocellular carcinoma (Alter et al., 1992). Interferon alpha and ribavirin are used for treatment of HCV infection. However, they are not highly effective, especially for HCV subtype 1b (Davis et al., 1998; McHutchison et al., 1998). Therefore, development of effective therapies and vaccines against HCV is an urgent subject worldwide.

For developing effective vaccines against HCV, it is important to identify epitopes involved in the clearance of HCV in vivo, because they may be the candidate vaccination targets. Although it may be difficult to identify those epitopes experimentally because of the lack of an efficient in vitro cell culture system and an in vivo animal model system for proliferation of HCV (Blight et al., 2000), they may be identified as the positively selected amino acid sites. This is because the amino acid mutations at those epitopes may provide selective advantage to the mutants by allowing them to escape from the immune response (Endo et al., 1996; Fitch et al., 1997; Nielsen and Yang, 1998; Suzuki et al., 1998).
Positive selection operating at the amino acid sequence level can be detected by comparing the rate of nonsynonymous substitution per nonsynonymous site \((r_S)\) with that of synonymous substitution per synonymous site \((r_N)\). A higher value of \(r_S > r_N\) is an indicator of positive selection, whereas the inverse is an indicator of negative selection (Hughes and Nei, 1988, 1989). By comparing \(r_N\) and \(r_S\) at single codon sites, it is possible to detect positive selection at single amino acid sites (Fitch et al., 1997; Nielsen and Yang, 1998; Suzuki and Gojobori, 1999; Yamaguchi-Kabata and Gojobori, 2000).

In this study, we identified positively selected amino acid sites in the entire coding region of HCV subtype 1b by using the method of Suzuki and Gojobori (1999) to predict the amino acid sites important for the clearance of HCV subtype 1b in vivo.

2. Materials and methods

2.1. Sequence data

In this study, we focused on HCV subtype 1b because of its clinical importance and a large number of sequence data deposited in the international DNA databank. Note that a relatively large number of sequences are required for detecting positively selected amino acid sites by using the method of Suzuki and Gojobori (1999). Actually, the numbers of sequences for another subtype were too small to detect positively selected amino acid sites in the entire coding region.

For detecting positively selected amino acid sites in the entire coding region of HCV subtype 1b, the simplest way is to analyze only sequence data which contain the entire coding region. However, the number of such sequences was relatively small. Thus, we divided the entire coding region of HCV subtype 1b into eight regions encoding C, E1, E2, NS2, NS3, NS4, NS5A, and NS5B and analyzed each region separately. p7 and E2 were combined as E2, and NS4A and NS4B were combined as NS4, because p7 (63 amino acid sites) and NS4A (54 amino acid sites) were too short to analyze positively selected amino acid sites. The numbers of amino acid sites for C, E1, E2, NS2, NS3, NS4, NS5A, and NS5B were 191, 192, 426, 217, 631, 315, 447, and 591, respectively.

A total of 7262 entries which included the successive terms ‘hepatitis C virus’ in their organism names were collected from the international DNA databank (DDBJ release 40). These entries included all clades and subtypes of HCV. To collect nucleotide sequences for each coding region of HCV subtype 1b, we defined HCV-JS (Accession number: D85516; Tanaka et al., 1995) as a reference sequence for HCV subtype 1b and made 7261 pairwise alignments, each of which consisted of HCV-JS and one of the other sequences, by using the computer program CLUSTAL W (Thompson et al., 1994). We then extracted sequences which did not contain any gaps within each coding region in the pairwise alignment. The subtype 1b sequences were identified by reconstructing a phylogenetic tree for each coding region. The numbers of nucleotide sequences collected for C, E1, E2, NS2, NS3, NS4, NS5A, and NS5B were 129, 135, 59, 74, 68, 70, 69, and 61, respectively. These sequences originated largely from unrelated patients and only a few were from the same patients. In this paper, the amino acid positions are numbered according to HCV-JS.

2.2. Data analysis

A multiple alignment was made for each coding region by using CLUSTAL W. The positively selected amino acid sites were identified by using the method of Suzuki and Gojobori (1999). Briefly, a phylogenetic tree was reconstructed by the neighbor-joining method (Saitou and Nei, 1987) using the number of synonymous substitutions (Nei and Gojobori, 1986). The ancestral sequence was inferred at each node in the phylogenetic tree by using the maximum parsimony method (Hartigan, 1973). Then, the average numbers of synonymous \((s_s)\) and nonsynonymous \((s_N)\) sites and the total numbers of synonymous \((c_s)\) and nonsynonymous \((c_N)\) substitutions throughout the phylogenetic tree were estimated for each codon site. A probability \((P)\) of obtaining the observed or more biased numbers of synonymous and nonsynonymous substitutions was computed for each codon site, assuming a binomial distribution. In the computation, \(s_s/(s_s + s_N)\) and \(s_N/(s_s + s_N)\) were used as the probabilities of the occurrences of synonymous and nonsynonymous substitutions, respectively. The significance level was set at 5%. A significantly larger value of \(c_N\) than \(c_S\) was considered as an indicator of positive selection, whereas the inverse was considered as an indicator of negative selection. The number of synonymous substitutions per synonymous site \((d_S)\) and that of nonsynonymous substitutions per nonsynonymous site \((d_N)\) were estimated by \(c_S/s_S\) and \(c_N/s_N\), respectively.

3. Results and discussion

The results for identifying positively selected amino acid sites in the entire coding region of HCV subtype 1b are summarized in Fig. 1. \(d_S\) exceeded \(d_N\) at most of the amino acid sites (2426/3010, 80.60%), and negative selection was detected at more than half (1560/3010, 51.83%) of all amino acid sites. The HCV polyprotein contains many B-cell and T-cell epitopes. For example, nearly the entire coding region of E1 and E2, and NS3 have been reported as B-cell and T-cell epitopes, respectively (Zibert et al., 1997a, 1999; Tabatabai et al., 1999). However, the amino acid sites involved in the clearance of HCV subtype 1b in vivo may be limited, because most of the sites in the entire
coding region, including E1, E2, and NS3, are negatively selected (Fig. 1).

Indeed, \(d_N\) exceeded \(d_S\) at only 265 amino acid sites (8.80%), and positive selection was detected at only 13 amino acid sites (0.43%) (Fig. 1). Among the 13 positively selected amino acid sites, eight were located in the structural proteins (C, E1, and E2) and five were in the nonstructural proteins (NS2, NS3, NS4, NS5A, and NS5B). Since the structural proteins occupy only 26.9% (809/3010) of the HCV polyprotein, positively selected amino acid sites seemed to be located more densely in the structural proteins than in the nonstructural proteins \((P = 0.009)\).

Table 1 summarizes the functions of positively selected amino acid sites in the entire coding region of HCV subtype 1b. Most of the sites are located in B-cell (Zibert et al., 1997a, 1999; Jolivet-Reynaud et al., 1998; Pereboeva et al., 1998, 2000; Nakano et al., 1999) and T-cell (Tabatabai et al., 1999; Wang and Eckels, 1999) epitopes, suggesting that both antibodies and cytotoxic T lymphocytes (CTLs) are involved in the clearance of HCV subtype 1b in vivo. However, a larger number of positively selected amino acid sites were located in B-cell epitopes than in T-cell epitopes. This is probably because

Table 1

<table>
<thead>
<tr>
<th>Position</th>
<th>Protein</th>
<th>Function</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>75</td>
<td>C</td>
<td>B-cell epitope</td>
<td>Pereboeva et al., 1998</td>
</tr>
<tr>
<td>231</td>
<td>E1</td>
<td>B-cell epitope</td>
<td>Zibert et al., 1999</td>
</tr>
<tr>
<td>235</td>
<td>E1</td>
<td>B-cell epitope</td>
<td>Zibert et al., 1999</td>
</tr>
<tr>
<td>253</td>
<td>E1</td>
<td>B-cell epitope</td>
<td>Zibert et al., 1999</td>
</tr>
<tr>
<td>345</td>
<td>E1</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>387</td>
<td>E2</td>
<td>B-cell epitope</td>
<td>Zibert et al., 1997a</td>
</tr>
<tr>
<td>461</td>
<td>E2</td>
<td>B-cell epitope</td>
<td>Nakano et al., 1999; Zibert et al., 1999</td>
</tr>
<tr>
<td>574</td>
<td>E2</td>
<td>B-cell epitope</td>
<td>Nakano et al., 1999; Zibert et al., 1999</td>
</tr>
<tr>
<td>827</td>
<td>NS2</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>1384</td>
<td>NS3</td>
<td>T-cell epitope</td>
<td>Tabatabai et al., 1999; Wang and Eckels, 1999</td>
</tr>
<tr>
<td>1644</td>
<td>NS3</td>
<td>B-cell epitope</td>
<td>Pereboeva et al., 2000</td>
</tr>
<tr>
<td>2719</td>
<td>NS5B</td>
<td>T-cell epitope</td>
<td>Tabatabai et al., 1999</td>
</tr>
<tr>
<td>2968</td>
<td>NS5B</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

* The amino acid position is numbered according to HCV-JS.

b NA, not available.
the recognition of T-cell epitopes is restricted by the haplotype of human leukocyte antigen, whereas that of B-cell epitopes is not. The sequence data used in this study were collected from the international DNA databank, which included HCV subtype 1b sequences largely from unrelated patients. Therefore, the positive selection may be more efficiently detected in B-cell epitopes than in T-cell epitopes in this study. This is consistent with the above observation that positively selected amino acid sites are more densely located in the structural proteins than in the nonstructural proteins, because B-cell epitopes are mainly located in structural proteins. Since most positively selected amino acid sites were located in B-cell and T-cell epitopes, amino acid positions 345, 827, 2719, and 2968 may also be parts of B-cell and T-cell epitopes.

The positively selected amino acid sites in B-cell and T-cell epitopes may be the vaccination targets against HCV subtype 1b, because these sites should be highly immunogenic and involved in the clearance of HCV subtype 1b in vitro. However, it should be noted that these sites are often highly variable, so that HCV may escape from the immune response by producing antigenic mutants (Nowak et al., 1991; Weiner et al., 1992; Farci et al., 2000). This may be facilitated by a high evolutionary rate of HCV (Ina et al., 1994; Smith et al., 1997; Allain et al., 2000; Suzuki et al., 2000) and shifting immunodominance of the immune response (Nowak et al., 1995). Nevertheless, it has been reported that if the immune response is sufficiently strong in an acute infection, HCV is unable to escape from the immune response even if it is directed against highly variable epitopes (Missale et al., 1996; Zibert et al., 1997a,b; Farci et al., 2000). Moreover, the composite vaccines containing different amino acid residues, particularly predicted future amino acid residues, at positively selected amino acid sites may be useful for preventing proliferation of escape mutants (Bush et al., 1999). Since both antibodies and CTLs seem to be involved in the clearance of HCV subtype 1b, it may be more effective to use both B-cell and T-cell epitopes as the vaccination targets.

In conclusion, we identified 13 positively selected amino acid sites in the entire coding region of HCV subtype 1b. These sites may be candidate vaccination targets against HCV subtype 1b.

Acknowledgements

The authors thank Dr Allison Wyndham at the National Institute of Genetics, Japan, for providing valuable comments on an earlier version of the manuscript. We are grateful to two anonymous reviewers for suggestions to improve the manuscript. This work was supported, in part, by grants from the Ministry of Education, Culture, Sports, Science, and Technology, Japan. Y.S. is supported by the JSPS Research Fellowships for Young Scientists.

References


