A Comprehensive System for Consistent Numbering of HCV Sequences, Proteins and Epitopes

Carla Kuiken,[1] Christophe Combet,[2] Jens Bukh,[4] Tadasu Shin-I,[5] Gilbert Deleage,[2] Masashi Mizokami,[5] Russell Richardson,[1] Erwin Sablon,[6] Karina Yusim,[1] Jean-Michel Pawlotsky,[3] Peter Simmonds,[7] and the Los Alamos HIV database group[1]

In October 2004, an expert meeting was convened in parallel with the 11th Symposium on Hepatitis C and Related Viruses to discuss how HCV sequence databases could introduce and facilitate a standardized numbering system for HCV nucleotides, proteins and epitopes. Inconsistent and inaccurate numbering of locations in DNA and protein sequences is a problem in the HCV scientific literature. Consistency in numbering is increasingly required for functional and clinical studies of HCV. For example, an unambiguous method for referring to amino acid substitutions at specific positions in NS3 and NS5B coding sequences associated with resistance to specific HCV inhibitors is essential in the investigation of antiviral treatment. This article provides a practical guide to help circumvent these problems in the future, and to bring a common language into discussions in the field. The scope of the current system is limited to the HCV polyprotein and the untranslated regions (UTRs); because of the controversial nature and extreme length variation of the alternate reading frame proteins, numbering for these proteins, if needed, will be decided at a later date.

We propose a numbering system adapted from the Los Alamos HIV database,[1] with elements from the hepatitis B virus numbering system.[2] The system comprises both nucleotides and amino acid sequences and epitopes. It uses the full length genome sequence of isolate H77 (accession number AF009606) as a reference, and includes a method for numbering insertions and deletions relative to this reference sequence. H77 was chosen because it is a commonly used reference strain for many different kinds of functional studies. Furthermore, RNA transcripts from this sequence are of demonstrated infectivity,[3,4] providing evidence that the 5′ and 3′ ends of the sequence are complete. Table 1 lists the boundaries of HCV genomic regions and Fig. 1 provides detailed nucleic acid and amino acid numbering over the complete AF009606 HCV genome sequence.

Protein vs. Polyprotein Numbering

Numbering an amino acid sequence can be absolute, *i.e.*, based on the HCV polyprotein with numbering starting at the first amino acid of the core protein and continuing through the end of NS5B; or it can be relative, *i.e.*, starting over at every protein (core is numbered from 1 to 191, then E1 from 1 to 192). Both of these numbering systems are used in HCV research. While most epitopes are numbered following the polyprotein numbering, drug resistance mutations tend to be numbered using protein numbering, following the practice for HIV-1 and now HBV. All HCV sequence databases incorporate both numbering systems. In order to avoid any confusion between the absolute and the relative numbering, the protein coordinates should be preceded by the protein name when using the relative numbering (*e.g.*, NS3-A156T). Conversion between the relative and absolute numbering can be easily done using either the Locator tool, available at the American HCV database website (http://hcv.lanl.gov/content/hcv-db/LOCATE/locate.html), or the Number tool on the European HCV database website (http://euhcvdb.ibcp.fr/euHCVdb/).

Abbreviations: HCV, hepatitis C virus; UTR, untranslated region; PPT, polypyrimidine tract; CDR, complementarity-determining region.

From the [1]Theoretical Biology and Biophysics group, Los Alamos National Laboratory, Los Alamos, NM; [2]Institut de Biologie et Chimie des Protéines (IBCP UMR5086); CNRS; Univ. Lyon 1; IFR128 BioSciences Lyon-Gerland, Lyon, France; [3]Department of Virology & INSERM U635, Henri Mondor Hospital, University of Paris 12, Créteil, France; [4]Hepatitis Viruses Section, Laboratory of Infectious Diseases, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, and Department of Infectious Diseases, Hvidovre University Hospital, Copenhagen Denmark; [5]Department of Clinical Molecular Informative Medicine, Nagoya City University Graduate School of Medical; Sciences, Kawasumi, Mizuho, Nagoya, Japan; [6]Innogenetics N.V., Ghent, Belgium; and [7]Centre for Infectious Diseases, University of Edinburgh, Summerhall, Edinburgh, United Kingdom.

Received March 24, 2006; accepted August 16, 2006.

The HIV database group consists of: Bette Korber, Werner Abfalterer, Charles Calef, Brian Foley, Robert Funkhouser, Brian Gaschen, Dorothy Lang, Thomas Leitner, James Szinger, and Ming Zhang.

Address reprint requests to: Dr. Carla Kuiken, LANL, MS K710, Los Alamos, NM 87545. E-mail: Kuiken@lanl.gov.

Copyright © 2006 by the American Association for the Study of Liver Diseases.

Published online in Wiley InterScience (www.interscience.wiley.com).
DOI 10.1002/hep.21377

Potential conflict of interest: Nothing to report.

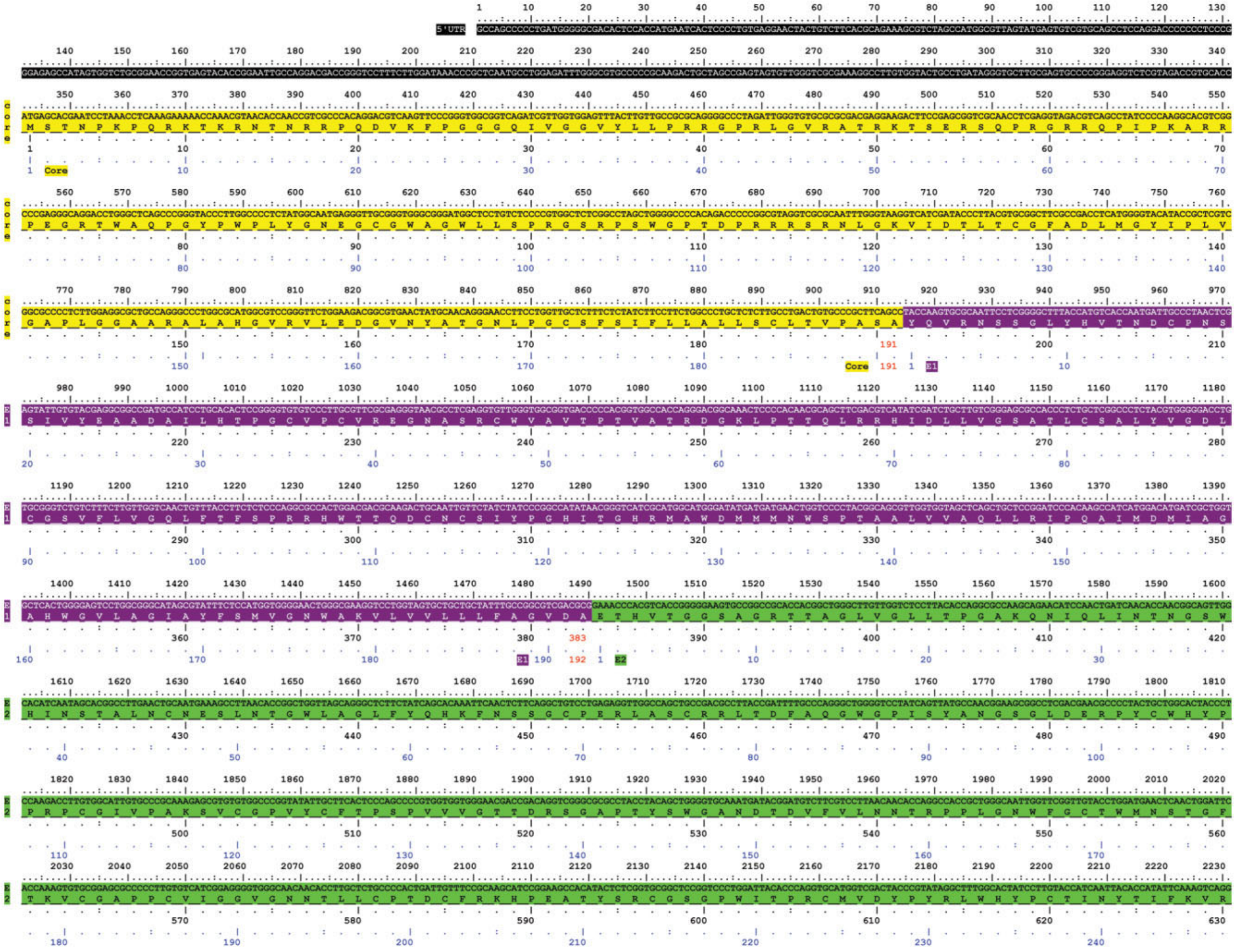

Fig. 1. The complete AF009606 HCV genome sequence.

Numbering of the 5′ UTR

There are two different methods for the numbering of the 5′ UTR that each have their adherents and opponents. One system starts numbering the polyprotein coding sequence at 1, and continues on through the end of the 3′ UTR; the 5′ UTR is numbered in the reverse direction, starting at -1 and continuing on to -341. The other system simply starts at 1 for the first nucleotide of the 5′ UTR, and continues on to the last nucleotide of the 3′ UTR. Both systems have advantages and disadvan-

Table 1. AF009606 Genomic Regions

Region	NA absolute numbering	NA relative numbering	AA absolute numbering	AA relative numbering	Description
5'UTR	1->341	1->341	na	na	5' Untranslated region
Core	342 -> 914	1 -> 573	1 -> 191	1 -> 191	Core protein
E1	915 -> 1490	1 -> 576	192 -> 383	1 -> 192	Envelope glycoprotein 1
E2	1491 -> 2579	1 -> 1089	384 -> 746	1 -> 363	Envelope glycoprotein 2
p7	2580 -> 2768	1 -> 189	747 -> 809	1 -> 63	Putative ion channel
NS2	2769 -> 3419	1 -> 651	810 -> 1026	1 -> 217	Autoprotease
NS3	3420 -> 5312	1 -> 1893	1027 -> 1657	1 -> 631	Serine protease and RNA dependent RNA helicase
NS4A	5313 -> 5474	1 -> 162	1658 -> 1711	1 -> 54	NS3 co-factor
NS4B	5475 -> 6257	1 -> 783	1712 -> 1972	1 -> 261	NS4B protein
NS5A	6258 -> 7601	1 -> 1344	1973 -> 2420	1 -> 448	NS5A phosphoprotein
NS5B	7602 -> 9377	1 -> 1776	2421 -> 3011	1 -> 591	RNA-dependent RNA polymerase
3'UTR	9378 -> 9646	1 -> 269	na	na	3' Untranslated region

Abbreviations: AA, amino acid, NA, nucleic acid.

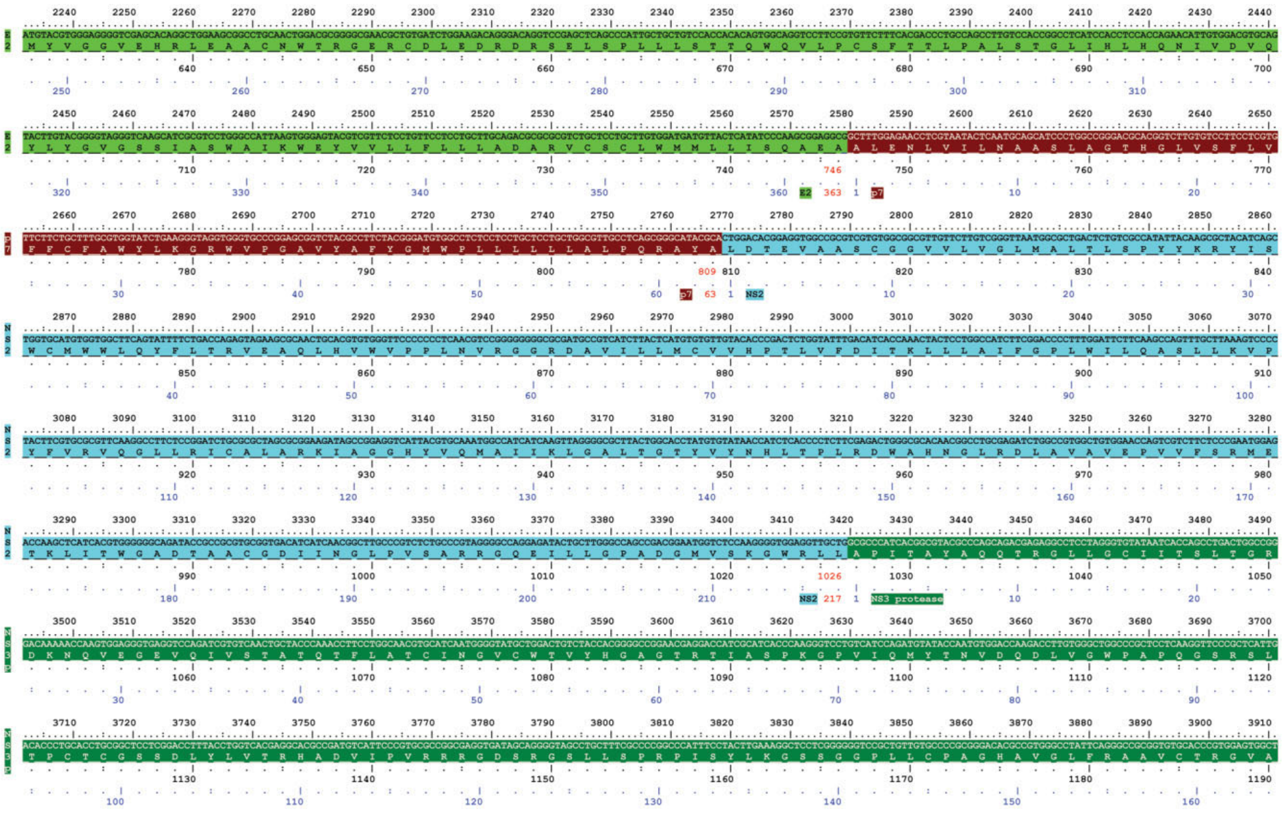

Fig. 1 (Cont'd.)

tages, and one can easily be converted to the other. For reasons of simplicity we have decided to adopt the numbering system that sets the start of the 5′ UTR at 1. Nucleotide numbering continues uninterrupted into the coding sequence. In the case of AF009606, the AUG initiation codon would be numbered as 342-344. This system avoids the practical problem encountered in most sequence editors of numbering a sequence non-consecutively (*i.e.*, in the negative numbering system for the 5′ UTR, the numbering at the 5′ UTR/core junction should proceed . . .−3, −2, −1, +1, +2, +3. . ., whereas sequence editors tend to number the −1 base as zero).

Numbering of the 3′ UTR

Numbering of the 3′ end of the HCV genome has to accommodate a stretch of pyrimidine residues of highly variable length between the 40 alignable nucleotides at the start of the 3′ UTR and the end of the 3′ UTR, including the highly conserved 3′X-tail. Diagrammatically, the 3′ UTR comprises the following sequential stretches (AF009606 numbering):

Coding sequence [. . .] CTC CCC AAC CGA TGA - 9377

3′UTR start (variable region): 9378–AGGTTGGGGTAAA CACTCCGGCCTCTTAG GCCATTTCCTG - 9417

Spacer region: 9418 - <60-100 pyrimidines, mainly T> -9548

3′UTR end (conserved region or 3′X): 9549–GGTGGC TCCATCTTAGCCCTAGTCACGGCTAGCTGT GAAAGGTCCGTGAGCCGCATGACTGCAGAGAGTGC-TGATA CTGGCCTCTCTGCAGATCATGT - 9646

The length of the poly-pyrimidine tract (PPT) changes rapidly over time within an experimentally infected chimpanzee[5] and is known to vary between different clones of the original H77 isolate[3,4] (accessions AF009069-77). The actual numbering of the 3′ end of the 3′ UTR sequence beyond the PPT is therefore of no significance. Furthermore, since the alignment of the PPT is arbitrary, so are any designation of insertions or deletions and any numbering attempt in this region.

An additional complication is caused by the choice of H77 as the reference sequence. The 3′X region of H77 has AAT as the first 3 nucleotides. However, this is not the case in most other isolates in which the 3′X is only 98 nucleotides long. To prevent complications arising from this unusual feature of H77, for numbering purposes, the AAT of H77 should be considered as part of the PPT rather than the 3′X region.

With this exception, the 3′ UTR follows the numbering of the reference sequence AF009606. If the sequence under consideration has a PPT that is not longer than AF009606 (no sequence currently in the HCV database has a longer PPT), the numbering is straightforward. For a sequence with a longer PPT, we propose that the insertion is ignored for the purpose of

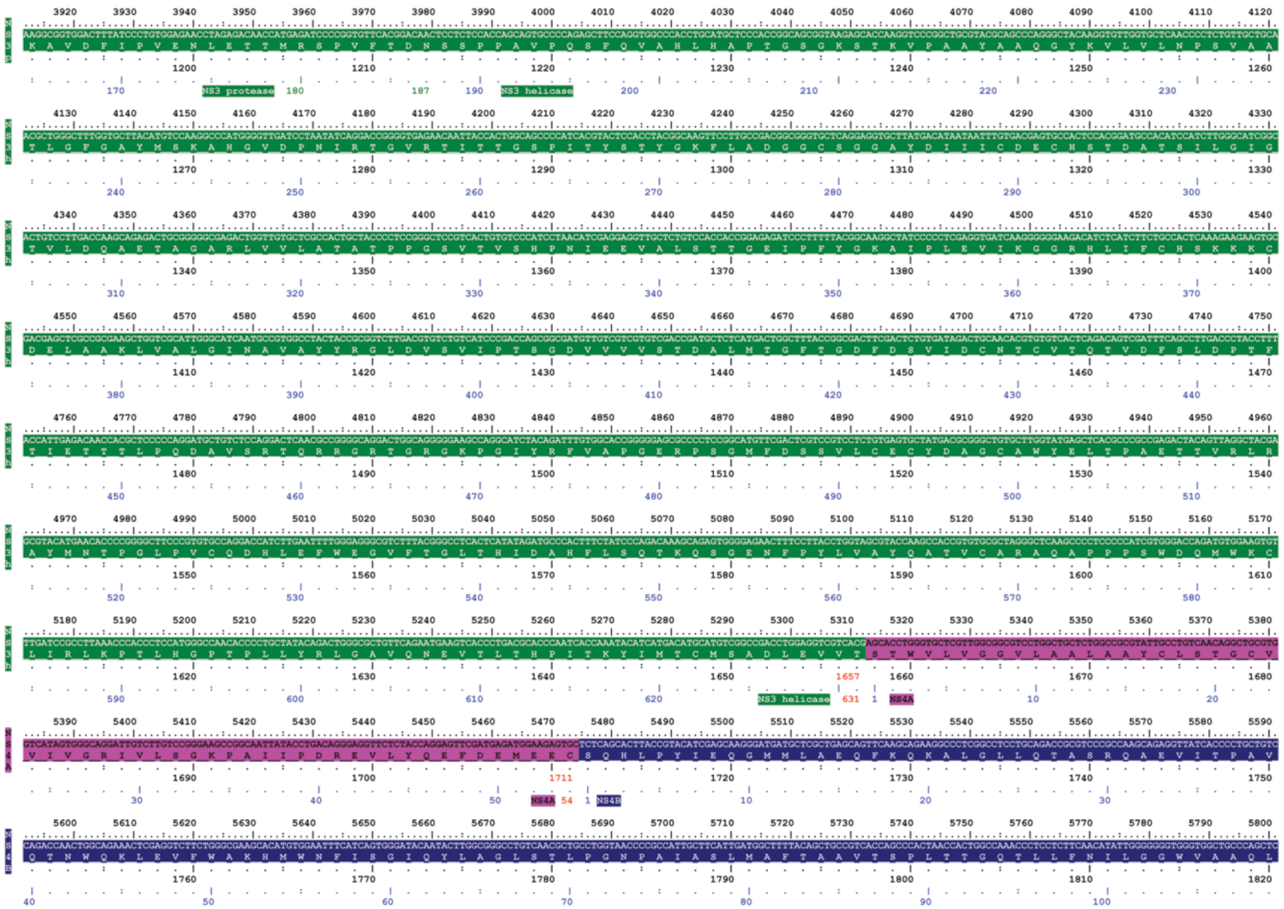

Fig. 1 (Cont'd.)

numbering. This means that the first nucleotide of the continuation of the alignment after the PPT, the region described in [3]) and numbered 9418 in their paper, will be numbered 9549, regardless of whether this is the actual position. Positions after that will again be numbered consecutively.

Numbering of Mutations

Mutations are numbered according to their position, *e.g.*, a NS5A:R217K would be used to indicate that the Arginine in position 217 of the NS5A protein has mutated to a Lysine. The codon involved changes from AGG to AAG (numbering 6906-6908), so the corresponding nucleotide mutation would be denoted G6907A.

Dealing With Insertions and Deletions Relative to the Reference Sequence

Insertions can be incorporated by basing the numbering on an alignment that includes all possible insertions, or by devising a special numbering system for insertions relative to a given numbering system. There are several reasons why the second method is preferable. First, sequences occasionally contain very long insertions, meaning that it would be hard to define an alignment that could accommodate all future insertions, and changing the numbering system in a way that would invalidate all previous numbering would be very disruptive. This problem is circumvented by the method outlined below. Second, insertions in HCV are relatively rare, so that this notation does not have to be used often.

Deletions are much simpler to deal with than insertions, and we outline a simple naming method for unequivocally designating those.

1. ***Insertion in sequence relative to the reference sequence.*** For these cases we propose a residue number/alphabet, where inserted bases or amino acids are indicated by lower case letters following the nucleotide or amino acid position where they occur. For example, three inserted bases in an HCV variant inserted between positions 131 and 132 in the AF009606 reference sequence would be described as 131a, 131b and 131c. (For insertions longer than the length of the alphabet, numbering would proceed 131x, 131y, 131z, 131aa, 131ab, 131ac, . . .131az, 131ba, 131bb, 131bc. . .). A similar scheme has been used for numbering amino acids in the immunoglobulin complementarity-de-

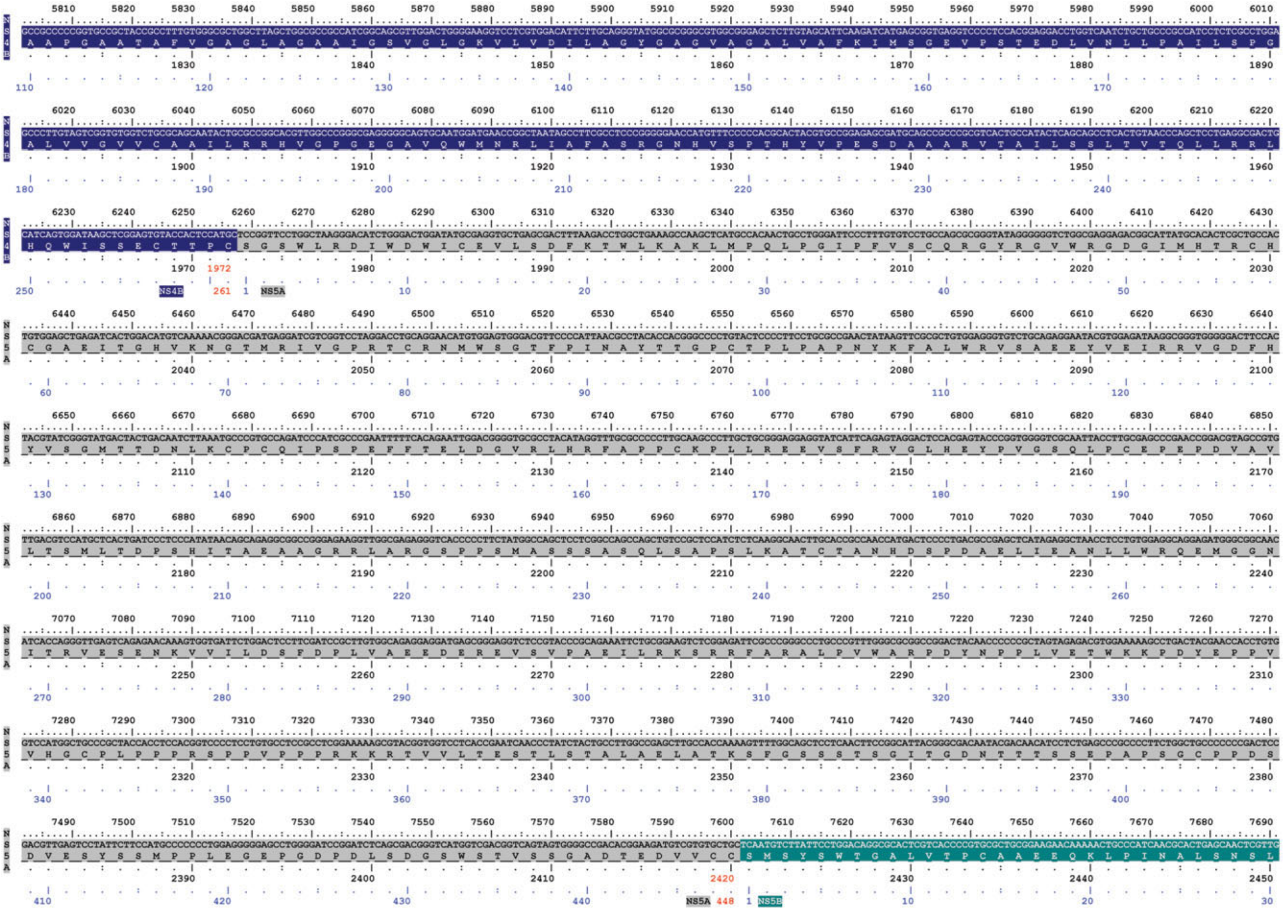

Fig. 1 (Cont'd.)

termining region (CDR) loops (*e.g.*, [6]). Example: in the following NS5A fragment, the location in the variant of the inserted aspartate (D) between AF009606 residues 2412 and 2413 would be referred to as D2412a or NS5A-D440a.

2411 2415 amino acids from start of H77 polyprotein
| |
GA-DTE - AF009606
EADDTE - variant sequence

Table 2 shows the numbering of each variant amino acid in this example.

2. ***Deletion in sequence relative to the reference sequence.*** Mentioning the deletions can be useful in cases where the length of the fragment is important, or when referring to the location of an amino acid that is located after a gap (relative to the reference sequence) in an epitope. Example: in the following NS5A fragment, the variant region would be numbered 2410-2415 (del 2413) or NS5A-438-443 (del 441) to make this explicit.

2410 2415 H77 AA position from start of polyprotein
| |
SGADTE - AF009606
EEA-TE - variant

Table 3 shows the numbering of each variant amino acid in this example.

Mentioning the deletions would be useful in cases where the length of the fragment is important; in this example, if the deletion were not made explicit it could

Table 2. Numbering of Insertions Relative to the H77 Sequence

H77 AF009606	G	A	–	D	T	E
Variant AA	E	A	D	D	T	E
AA Absolute	2411	2412	2412a	2413	2414	2415
AA Relative	439	440	440a	441	442	443
Variant NA	GAA	GCC	GAC	GAC	ACG	GAA
NA Absolute	7572–7574	7575–7577	7577a–7577c	7578–7580	7581–7583	7584–7586
NA Relative	1315–1317	1318–1320	1320a–1320c	1321–1323	1324–1326	1327–1329

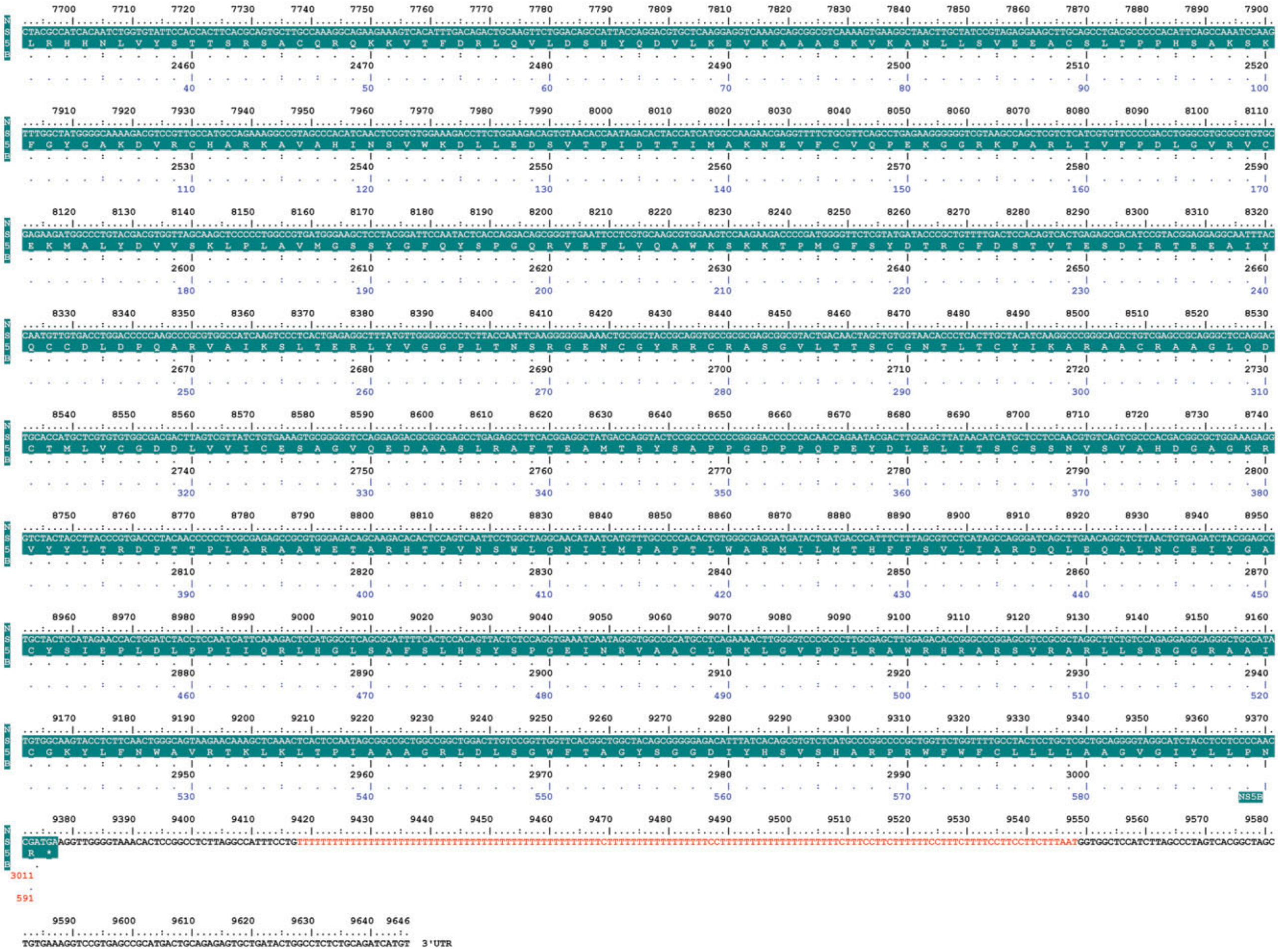

Fig. 1 (Cont'd.)

be assumed that the range 2410-2415 refers to a peptide that is 6 AA long, when it really is only 5 AA.

3. ***Synthetic sequences.*** A separate problem is that of artificially engineered sequences, such as replacement of stretches of the HCV genome with extraneous genetic segments. The numbering system should be able to cope with this situation also, by treating the extraneous segment as a special type of HCV, which can be numbered according to the length of the insert. Finding the stretches that correspond to "real" HCV depends on pattern matching, which requires accurate location of the end and re-start of HCV sequences on either side of the insert to retain correct numbering. Using the same example,

```
2411 2415 amino acids from polyprotein start
| | |
GADT——-E - AF009606
GADTYNTVATLE - variant sequence
=======
insert
```

Table 4 shows the numbering of the insert in this example.

Numbering Positions in Other Genotypes

While the numbering is based on a genotype 1a reference sequence, both the Sequence Locator tool (US database) and

Table 3. Numbering of Deletions Relative to the H77 Sequence

H77 AF009606	S	G	A	D	T	E
Variant AA	E	E	A	–	T	E
AA Absolute	2410	2411	2412	–	2414	2415
AA Relative	438	439	440	–	442	443
Variant NA	GAA	GAA	GCC	–	ACG	GAA
NA Absolute	7569–7571	7572–7574	7575–7577	–	7581–7583	7584–7586
NA Relative	1312–1314	1315–1317	1318–1320	–	1324–1326	1327–1329

Table 4. Numbering of Non-HCV Inserts

H77 AF009606	G	A	D	T	-..-	E
Variant AA	G	A	D	T	Y..L	E
AA Absolute	2411	2412	2413	2414	2414a..j	2415
AA Relative	439	440	441	442	442a..j	443
Variant NA	GGA	GCC	GAC	ACG	TAT. . . CTC	GAA
NA Absolute	7572–7574	7575–7577	7578–7580	7581–7583	7583a–7583ad	7584–7586
NA Relative	1315–1317	1318–1320	1321–1323	1324–1326	1326a–1326ad	1327–1329

the Number tool (European database) are sufficiently flexible to easily accommodate other genotypes. The tools align sequences of all genotypes described to date unambiguously to the reference sequence, so the numbering will be uniform for other genotypes as well as for genotype 1.

Summary and Conclusion

This numbering proposal, using the AF009606 (isolate H77) sequence as a reference, should be able to unequivocally number all possible mutations in HCV, both natural and manmade. The HCV sequence databases[8] and the Los Alamos HCV immunology database[9] (as well as the Los Alamos HIV database) number positions and epitopes according to this system. Moreover, the databases websites provides tools for finding stretches of sequence by their numbers, for assigning start and end coordinates to a sequence, and for converting between the various numbering systems.

Numbering HCV nucleotide sequences is done by analogy to H77. The first step is aligning your sequence to H77. If there is no length variation, the numbering is straightforward; nucleotide numbers run from 1 (start of 5′ UTR) to 9646 (end of 3′ UTR). Insertions relative to H77 are labeled with letters.

Protein numbering works like the nucleotide numbering, but starts at the start of the polyprotein. The sequence databases will support both systems, but use polyprotein numbering as a basis. Absolute numbering moves across the coding regions, relative numbering starts over at every coding region. Relative numbering is almost exclusively used for proteins, polyprotein numbering mainly in immunology, protein numbering in drug resistance research. The Los Alamos immunology database uses polyprotein numbering.

The 5′ UTR numbering starts at 1 and ends at 341; the Core cds starts at 342. The numbering of the 3′ UTR starts at 9378 (after the stop codon), but complications arise due to the variable length of the PPT. The UTR consists of 3 elements: a variable 5′ region, the PPT, and a conserved 3′ region, often called X. The first region is numbered 9378-9410. The PPT consists almost entirely of T's and therefore cannot be meaningfully aligned; it is numbered according to its length in H77, 9411-9545. The X region starts at 9546 (regardless of its actual location, which depends on the length of the PPT) and ends at 9646.

The HCV databases can be found at:
European site: http://euhcvdb.ibcp.fr
Japanese site: http://s2as02.genes.nig.ac.jp/
American site: http://hcv.lanl.gov

Acknowledgment: The European HCV database is funded by HepCVax Cluster (EU FP5 grant QLK2-2002-01329) and VIRGIL Network of Excellence (EU FP6 grant LSHM-CT-2004-503359). The Japanese Hepatitis database is supported by a Grant-in Aid for the Publication of Scientific Research Results (#168111), Grant-in Aid for Scientific Research, Japan Society for the Promotion of Science (JSPS). The American HCV database is funded by the Division of Microbiology and Infectious Diseases, National Institute of Allergy and Infectious Diseases under interagency agreement Y1-A1-1500.

References

1. Korber B, Foley B, Kuiken C, Pillai S, Sodroski J. Numbering positions in HIV relative to HXB2CG. In: Korber BK, Kuiken CL, Foley B, Hahn B, McCutchan F, Mellors JW, Sodroski J, editors. Human Retroviruses and AIDS. Los Alamos, MN: Los Alamos National Laboratory, 1998.
2. Stuyver LJ, Locarnini SA, Lok A, Richman DD, Carman WF, Dienstag JL, Schinazi RF. Nomenclature for antiviral-resistant human hepatitis B virus mutations in the polymerase region. HEPATOLOGY 2001;33:751-757.
3. Kolykhalov AA, Feinstone SM, Rice CM. Identification of a highly conserved sequence element at the 3′ terminus of hepatitis C virus genome RNA. J Virol 1996;70:3363-3371.
4. Yanagi M, St Claire M, Emerson SU, Purcell RH, Bukh J. Transcripts from a single full-length cDNA clone of hepatitis C virus are infectious when directly transfected into the liver of a chimpanzee. Proc Natl Acad Sci U S A 1997;94:8738-8743.
5. Okamoto H, Kojima M, Okada S, Yoshizawa H, Iizuka H, Tanaka T, et al. Genetic drift of hepatitis C virus during an 8.2-year infection in a chimpanzee: variability and stability. Virology 1992;190:894-899.
6. Lucas AH, Moulton KD, Reason DC. Role of kappa II-A2 light chain CDR-3 junctional residues in human antibody binding to the Haemophilus influenzae type b polysaccharide. J Immunol 1998;161:3776-3780.
7. Rybach L, Bucher B, Schwarz G, Walewski JL, Keller TR, Stump DD, Branch AD. Evidence for a new hepatitis C virus antigen encoded in an overlapping reading frame. RNA 2001;7:710-721.
8. Kuiken C, Mizokami M, Deleage G, Yusim K, Penin F, Shin-I T, et al. Hepatitis C databases, principles and utility to researchers. HEPATOLOGY 2006;in press.
9. Yusim K, Richardson R, Tao N, Dalwani A, Agrawal A, Szinger J, et al. Los Alamos hepatitis C immunology database. Appl Bioinformatics2005;4:217-225.