J Mol Evol (2002) 54:62–70
DOI: 10.1007/s00239-001-0018-9

JOURNAL OF MOLECULAR EVOLUTION

Hepatitis C Virus Evolutionary Patterns Studied Through Analysis of Full-Genome Sequences

Marco Salemi, Anne-Mieke Vandamme

Rega Institute for Medical Research, KULeuven, Minderbroedersstraat 10, B-3000 Leuven, Belgium

Received: 13 January 2001 / Accepted: 21 June 2001

Abstract. The evolutionary patterns of hepatitis C virus (HCV), including the best-fitting nucleotide substitution model and the molecular clock hypothesis, were investigated by analyzing full-genome sequences available in the HCV database. The likelihood ratio test allowed us to discriminate among different evolutionary hypotheses. The phylogeny of the six major HCV types was accurately inferred, and the final tree was rooted by reconstructing the hypothetical HCV common ancestor with the maximum likelihood method. The presence of phylogenetic noise and the relative nucleotide substitution rates in the different HCV genes were also examined. These results offer a general guideline for the future of HCV phylogenetic analysis and also provide important insights on HCV origin and evolution.

Key words: Hepatitis C virus — Evolutionary models — Star-like phylogeny — Likelihood ratio test — Γ distribution — Molecular clock

Introduction

Hepatitis C virus (HCV) is a positive strand RNA virus of approximately 9.4 kb belonging to the *Flaviridae* family (Choo et al. 1989). HCV is an important pathogen worldwide and is a major cause of chronic hepatitis, cirrhosis, and hepatocellular carcinoma (Nishioka et al. 1991; Farci et al. 1991; Saito et al. 1990). Six major genotypes, described as clade or type 1 to 6, have been identified so far, and many of these types are further subdivided into distinct but more closely related subtypes (Simmonds et al. 1994a, b). Recently, the creation of a HCV database (http://s2as02.genes.nig.ac.jp) has provided easy access to almost 100 full-genome sequences and thousands of partial sequences which constitute a valuable resource for the investigation of HCV evolution at the molecular level.

Extensive phylogenetic analyses have been performed to clarify the origin and the evolution of HCV (Bukh et al. 1993; Chamberlain et al. 1997; de Lamballerie et al. 1997; Mellor et al. 1995; Ina et al. 1994; Simmonds et al. 1994a, b, 1995, 1996; Smith et al. 1997). However, some aspects of the evolutionary dynamics of the virus still remain elusive. What is the nucleotide substitution model best describing the evolution of the HCV clades? A good model of nucleotide substitution is essential to infer correct phylogenetic relationships among viral variants and can be used to answer different epidemiological questions, such as the origin and global spread of different genetic types and subtypes (Salemi et al. 2000a). It is known that HCV is a fast-evolving virus with an average evolutionary rate comparable to that of HIV-1 (Ina et al. 1994; Smith et al. 1997; Suzuki et al. 2000). Smith et al. (1997) discussed the origin of HCV, suggesting that the subtypes diverged around 300 years ago and that the divergence of the different types should have occurred more than 500 years ago. Using a cohort of recipients with a known data of infection from the same source,

Correspondence to: Marco Salemi; *email:* marco.salemi@uz.kuleuven.ac.be

Table 1. Full-genome sequences selected from the HCV database for the study

Accession No.	Strain name	Subtype[a]	Country of origin	Reference
AF009606	H77	1a	USA	Kolykhalov et al. (1997)
D10749	J1	1a	Japan	Okamoto et al. (1992a)
M62321	HCV-1	1a	USA	Choo et al. (1991)
M67463	HCV-H	1a	USA	Inchauspe et al. (1991)
AJ132996	AD78	1b	Germany	Rispeter et al. (1997)
D90208	HCV-J	1b	Japan	Kato et al. (1990)
X61596	JK1	1b	Japan	Honda et al. (1992)
U01214	L2	1b	Korea	Choo et al. (1991)
D14853	G9	1c	Indonesia	Okamoto et al. (1994)
D00944	J6	2a	Japan	Okamoto et al. (1992a)
D10988	HCV-J8	2b	Japan	Okamoto et al. (1992b)
D50409	BEBE-1	2c	Italy	Nakao et al. (1996)
AB031663	VAT96	2c	Moldova	Samokhvalov et al. (2000)
AF046866	CB	3a	Australia	Unpublished
D28917	K3a	3a	USA	Yamada et al. (1994)
X76918	CENS1	3a	Germany	Unpublished
D49374	HCV-Tr	3b	Japan	Chayama et al. (1994)
D63821	JK-049	3 (10a)	Indonesia	Tokita et al. (1996)
Y11604	ED43	4a	Egypt	Chamberlain et al. (1997)
AF064490	SA13	5a	?	Bukh et al. (1998)
Y13184	EUH1480	5a	Scotland	Chamberlain et al. (1997)
D63822	JK-046	6 (11a)	Indonesia	Tokita et al. (1996)
D84263	VN235	6 (7b)	Vietnam	Tokita et al. (1998)
D84264	VN405	6 (8b)	Vietnam	Tokita et al. (1998)
D84265	VN404	6 (9a)	Vietnam	Tokita et al. (1998)
Y12083	EUHK2	6a	Hong Kong	Adams et al. (1997)
D84262	TH580	6b	Thailand	Tokita et al. (1998)

[a] One strain belonging to clade 3 and four strains belonging to clade 6 received a different classification in previous studies. For clarity reasons we include the old nomenclature in parentheses together with the proposed standard classification (Robertson et al. 1998).

they calculated an evolutionary rate of $1.0 \cdot 10^{-3}$ nucleotide substitution per site per year for the E1 gene and $1.7 \cdot 10^{-3}$ for the NS5B gene. An evolutionary rate of about $7.51 \cdot 10^{-3}$ was calculated for the nonsynonymous substitution in the core gene (Ina et al. 1994). We previously used the likelihood ratio test to verify that HCV quasi-species within a single infected patient evolve clock-like over years and that closely related viral strains infecting different patients usually show no or little significant difference in their rate of evolution (Allain et al. 2000). The mean evolutionary rate in the E1/E2 region within the different patients was of the same order of magnitude of previous estimates: $0.63 \cdot 10^{-3}$ nucleotide substitution per site per year. These data seem to imply a recent origin of HCV, but whether or not the molecular clock assumption holds for the different HCV lineages has not been statistically tested.

There are other open questions relevant to HCV molecular evolution. For example, What is the amount of rate heterogeneity across sites? and What are the relative rates of evolution of the different viral genes? And did the different types arise from a common ancestor in a star-like burst? or Are they the result of a tree-like evolution? In the present study we address these questions by analyzing a representative data set of HCV full-genome sequences empolying maximum likelihood-based techniques.

Materials and Methods

Compilation of Sequence Data. At the time of writing, 93 full-genome sequences were available in the HCV database (http://s2as02.genes.nig.ac.jp) And 27 more have been published recently but not yet included (Nagayama et al. 2000). However, most of these sequences are from HCV clade 1 subtype b. It is impossible to analyze so many full-genome sequences with the computational intensive maximum likelihood (ML) methods that we use in our study. Besides, including several closely related 1b strains would not add any substantial information. Finally, some of the sequences in the database are chimeric or have been isolated from chimpanzees several years after experimental infection. Since it is our goal to study the molecular evolution of HCV in its natural host, these sequences have been excluded. We compiled a data set of 27 ful-genome sequences: all 23 strains belonging to the six major HCV clades available other than 1b (with the restriction outlined above) and 4 HCV 1b strains chosen as representative of this subtype (see Table 1). The sequences were aligned using the Clustal algorithm implemented in DAMBE (Xia 2000a). The HCV genome consists of a 5′ untranslated region (UTR) followed by a single open reading frame coding for a polyprotein precursor giving rise to the core, E1, E2, p7, NS2, NS3, NS4A, NS4B, NS5A, and NS5B viral proteins. The coding nucleotide sequences were aligned against their amino acid sequences to avoid the introduction of insertions or deletions within a codon (Xia 2000a), and all the gaps were removed from the final alignment. The alignment is available from the authors upon request.

Analysis of the Phylogenetic Signal. The presence of phylogenetic signal in the data set was investigated with the Hillis and Huelsenbeck (1992) method based on the skewness of the tree length distribution.

Given a data set, the tree length under a maximum parsimony criterion for all possible topologies (or a random sample of them) is computed. If there is no phylogenetic signal in the data, the distribution tends to be symmetric. If there is a phylogenetic signal, the distribution tends to be (left) skewed (Hillis and Huelsenbeck 1992). Since the number of unrooted possible trees grows exponentially with the number of taxa, we estimated the tree length distribution of 1 million random trees for the 27 HCV strains using the option "Evaluation Random Trees . . ." of PAUP*4.0d65, written by David L. Swofford.

Another way to visualize the presence of phylogenetic signal/noise in a particular data set of aligned sequences is to perform a likelihood mapping analysis investigating groups of four sequences randomly chosen, called quartets (Strimmer and von Haeseler 1997). For a quartet, just three unrooted tree topologies are possible. The likelihood of each topology can be estimated with the ML method and the three likelihoods can be reported as a dot in an equilateral triangle. For N sequences $\binom{N}{4}$ possible quartets exist and the distribution of the dots in the triangle can give an overall impression of the tree-likeness of the data. When the N sequences are not clustered, the order of the sequences is not relevant and the question, which of the possible tree topologies is supported by any cluster, is meaningless. Thus we can distinguish two main areas in the equilateral triangle (Strimmer and von Haeseler 1997; Nieselt-Struve 1998): the three corners, representing fully resolved tree topologies, i.e., the presence of a tree-like phylogenetic signal in the data; and the center and the three areas on the sides, which represent a star-like and a net-like phylogeny, respectively, and indicate the presence of phylogenetic noise. Thus, the percentage of dots belonging to each area gives an idea about the mode of evolution in the data set under investigation. Likelihood mapping analyses were performed with the program TREE-PUZZLE (Strimmer and von Haeseler 1997) on the full genome, the UTR, and the different gene regions of the HCV genome, separately. For each analysis all 17,750 possible quartets for the 27 HCV strains were evaluated.

HCV Consensus Tree. A phylogeny reconstruction with the wrong nucleotide or protein substitution model may lead to the wrong tree topology. However, to test evolutionary hypotheses with the likelihood ratio test, the phylogenetic relationships among the taxa in the data set should be known with reasonable accuracy (Huelsenbeck and Rannala 1997). To solve this sort of vicious circle we used the following strategy: first, we obtained HCV phylogenetic trees with different tree-construction algorithms, employing full-genome sequences and first + second or third codon positions (1st + 2nd or 3rd cdp) only; next, we computed a strict consensus tree (our "reasonably accurate" phylogeny) that was used to test different nucleotide substitution models; finally, we computed a new ML tree with the selected model. The starting trees were obtained with the neighbor-joining (NJ) method, the Fitch and Margoliash (Fitch) method, and the ML method. NJ and ML trees were calculated according to the Tamura and Nei (1993) model with Γ-distributed rates across sites (Yang 1996) implemented in PAUP*4.0b written by David L. Swofford. The Fitch tree was obtained with DAMBE, using the same substitution model. The NJ and Fitch trees were statistically evaluated using 1000 bootstrap samples, whereas p values were calculated for the ML tree (Felsenstein 1981). Though all the trees showed a similar topology, thye were not identical, especially with respect to the branching order of the major HCV clades (see Results). A strict consensus tree was generated using the option "Compute Consensus . . ." in the Trees menu of PAUP*4.0d65, written by David Swofford. The tree is reported in Fig. 1, and it was then used to test different nucleotide substitution models by means of the likelihood ratio test.

Models of Nucleotide Substitution and Likelihood Ratio Test. Using the HCV consensus tree, several parametric models were evaluated according to the likelihood ratio test (Huelsenbeck and Rannala 1997): JC69 (Jukes and Cantor 1969), K80 (Kimura 1980), HKY85 (Hasegawa et al. 1985), TN93 (Tamura and Nei 1993), and REV (Yang 1993). Eight discrete categories of Γ-distributed rates among sites were assumed. The program BASEML implemented in PAML 3.0 software package (Yang 2000) was used for the calculations. Finally, a more sophisticated model described below was used to investigate in detail the substitution rate heterogeneity among sites in coding regions of the HCV genome.

The molecular clock hypothesis was evaluated on the final tree (see next section) with the likelihood ratio test (Huelsenbeck and Rannala 1997) and the best-fitting nucleotide substitution model. The clock was tested for all cdps and for the 1st + 2nd, 1st, 2nd, and 3rd cdp separately.

Inferring and Rooting the HCV ML Tree. A new ML tree was inferred with the best-fitting nucleotide substitution model. For this final calculation we used the branch-and-bound algorithm (with the consensus tree as a backbone constraint) implemented in PAUP*4.0b to get the optimal ML tree. The tree was rooted with the hypothetical common ancestor of the HCV sequences, the central node in the consensus tree shown in Fig. 1, that was reconstructed employing the ML algorithm implemented in BASEML (Yang and Roberts 1995).

Patterns of Nucleotide Substitution Among HCV Genes. It is known that substitution rates are different in different genes (Li 1997). The same holds for HCV. For example, a hypervariable region has been identified at the 5′ end of the E2 gene (Weiner et al. 1991), and it has been observed that the UTR and the coding region for the core protein are much more conserved than other parts of the genome (Simmonds 1995). To investigate in further detail the heterogeneity of substitution rates across genes, we grouped the sites in the HCV genome into 11 classes, 1 representing the sites belonging to the UTR and the other 10 for the different proteins coded by the open reading frame of the polyprotein precursor. The relative substitution rate of each class (using the UTR as the reference class with a substitution rate equal to 1) was estimated with the ML approach implemented in BASEML. We employed the branch-and-bound ML tree obtained as described above and a modified general time-reversible nucleotide substitution model assuming Γ-distributed rates across sites (REV+Γ model) that allows for different parameters (including Γ-distributed rates across sites and transition transversion ratio) to investigate the heterogeneity of substitution rates in each class of sites (Salemi et al. 2000a).

Results

Analsis of the Phylogenetic Signal. The presence of phylogenetic information at the different cdp was assessed with the Hillis and Huelsenbeck test (1992) (see Materials and Methods). The skewness values of the tree length distribution were −0.692 for 1st cdp, −0.728 for 2nd cdp, and −0.744 for 3rd cdp. The three values are statistically significant ($p < 0.01$) according to tabulated values of the skewness test statistic from Hillis and Huelsenbeck (1992). They suggest that not only the 1st and 2nd but also the 3rd cdp, which exhibits the leftmost skewed distribution, remains phylogenetically informative.

The phylogenetic noise in the different HCV genes was analyzed with likelihood mapping (see Materials and Methods). As shown in Table 2, the presence of noise is negligible (3.3%) when the full genome is considered. Among the single genes, UTR, NS4A, and p7 exhibit the highest noise content (about 30%). The phy-

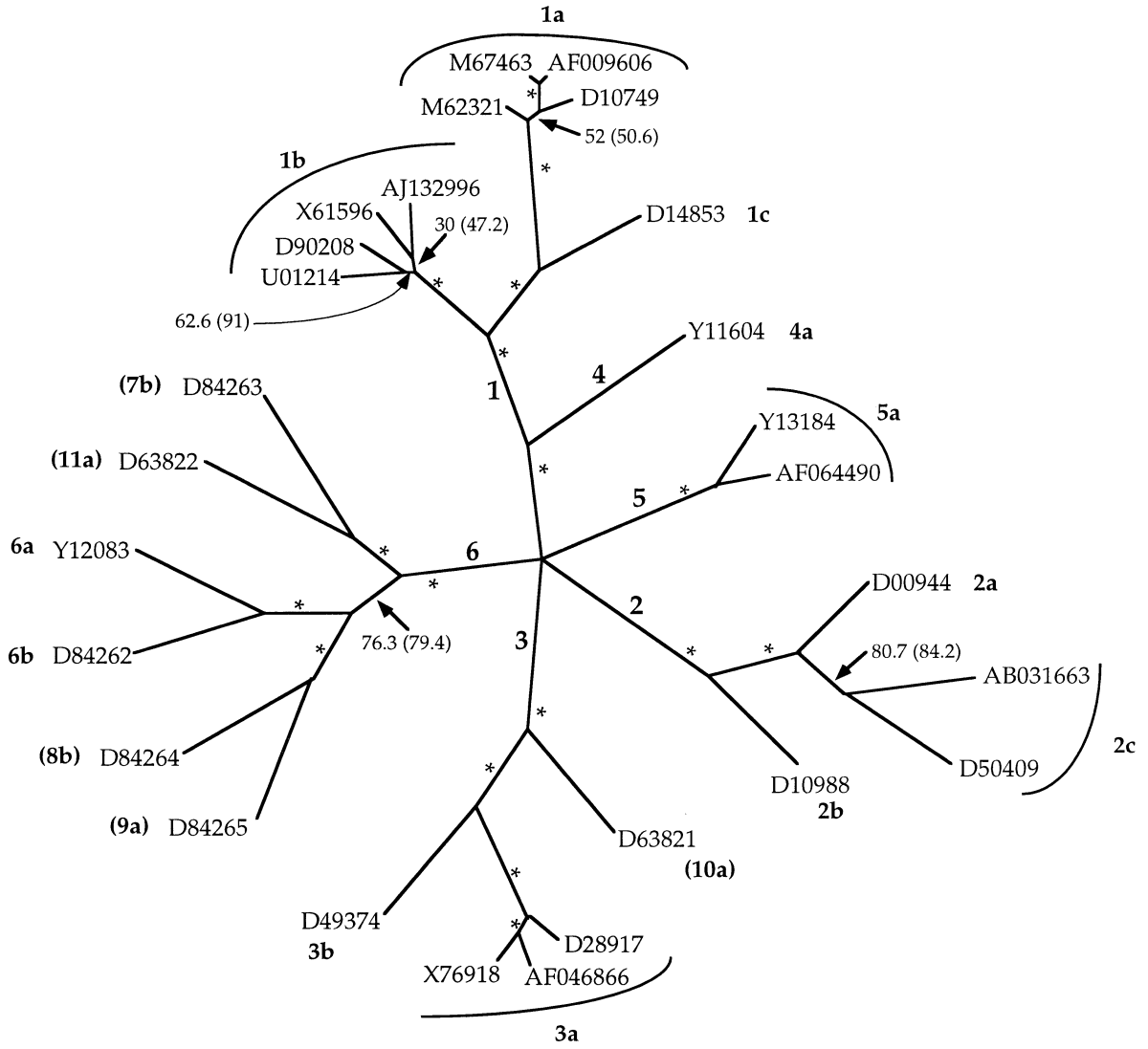

Fig. 1. Unrooted strict consensus tree of 27 full-genome HCV strains, based on the phylogenetic trees obtained with the neighbor-joining (NJ), the Fitch and Margoliash (Fitch), and the maximum likelihood (ML) method algorithms described under Materials and Methods. The six major lineages of HCV are indicated by the *boldface numbers* along the branches; subtypes of the viruses are also given (when subtypes are classified according to the old nomenclature, the classification is reported in parentheses). 1000 bootstrap samples were used for the NJ and Fitch trees. The *asterisks* represent clusters supported by >90% bootstraps (for both the NJ and the Fitch trees). For less-supported clades, bootstraps for the NJ tree and the Fitch tree (the latter in parentheses) are shown. All the branches of the consensus tree have a p value of <0.001 in the ML analysis.

logenetic noise is lower in the remaining gene regions, ranging between 9.6% (NS3) and 15.2% (E1).

Nucleotide Substitution Model. In Fig. 1 the HCV consensus tree obtained as described under Materials and Methods is given. Not surprisingly the different clades of HCV are supported with very high bootstrap and p values by all tree-building methods used (see Fig. 1), in agreement with previously published phylogenetic analyses (Simmonds et al. 1994a, b). Clade 1 and the strain Y11604 representative of clade 4 clustered together in all phylogenetic analyses with high support (see Fig. 1). On the other hand, the branching order of the remaining clades was different depending on the algorithm employed to infer the tree. Thus the star-like phylogeny shown in the consensus tree in Fig. 1 represents a soft polytomy reflecting the uncertainty in the phylogenetic relationships among these clades.

Likelihoods and numbers of parameters estimated on the HCV consensus tree under different nucleotide substitution models are reported in Table 3. Since all three cdp are informative from the phylogenetic point of view (see previous section), we used the full-genome alignment including all cdp for the analysis. Each model assumes Γ-distributed rates across sites (eight discrete categories), to take into account different nucleotide substitution rates which can occur at different cpd and in different genes. The most complex REV+Γ model shows

Table 2. Likelihood mapping anaysis of the 17,750 possible quartets (cluster of four sequences) of the 27 full-genome HCV strains

Gene region	No. sites[a]	Constant sites (%)	Net-like/star-like phylogeny (phylogenetic noise)[b]	Tree-like phylogeny (phylogenetic signal)[b]
Full genome	9201	36.9	3.3	96.7
UTR	246[c]	82.5	29.6	70.4
CORE	573	52.5	13.2	86.8
E1	576	25	15.2	84.8
E2	1059	33.6	11.2	88.8
p7	189	18	28.6	71.4
NS2	651	26.1	12.1	87.9
NS3	1893	41.2	9.6	90.4
NS4A	162	35.2	29	71
NS4B	783	36.4	12.1	87.9
NS5A	1296	29.2	9.9	90.1
NS5B	1773	38.8	5.1	94.9

[a] The number of sites after removing insertions and deletions from the aligned sequences.
[b] Values indicate the percentage of quartets that are compatible with a net-like/star-like or a tree-like phylogeny. The likelihood mapping analysis was performed using the program TREE-PUZZLE (Strimmer and von Haseler 1997), as described under Materials and Methods.
[c] Only 246 nucleotides of the UTR region were used since the full-length UTR was not available for all strains.

Table 3. Likelihood values of different nucleotide substitution models for the full-length-genome HCV consensus tree[a]

Site	np[b]	$-ln$Likelihood[c]	T_i/T_v	α
K80	51	139,713 ($p < 0.000$)	2.0	0.369
HKY85	51	129,264 ($p < 10^{-162}$)	2.18	0.357
TN93	52	129,207 ($p < 10^{-138}$)	1.89	0.356
REV	55	128,887	2.20	0.354

[a] All models assume Γ-distributed rates across sites, with the shape parameter α estimated via maximum likelihood.
[b] Number of parameters estimated by each model.
[c] The p value in parentheses refers to the comparison with the REV model using the likelihood ratio test (see Materials and Methods).

the highest likelihood, which is always significantly better than the likelihood of less complex models. The estimated number of transitions is about twice the number of transversions and the shape parameter α of the Γ distribution is always <1, indicating strong rate heterogeneity (see Table 3). It is worth noting that all the models give similar estimates for both the expected transition transversion ratio and α.

The molecular clock hypothesis does not hold for the different HCV clades when the whole genome is considered. The assumption of a clock-like evolution for the tree in Fig. 2 (see below) is always rejected, both for the 3rd cdp ($-ln\text{Likelihood}_{\text{clock}} = 70{,}593$, $-ln\text{Likelihood}_{\text{no clock}} = 70{,}557$; $p < 10^{-6}$) and for the 1st + 2nd cdp ($-ln\text{Likelihood}_{\text{clock}} = 70{,}593$, $-ln\text{Likelihood}_{\text{no clock}} = 70{,}557$; $p < 10^{-48}$). The molecular clock was also rejected when the different HCV gene regions were analyzed separately (data not shown).

Phylogeny of the HCV Types. Figure 2 shows the branch-and-bound maximum-likelihood tree of the 27 full-genome HCV sequences rooted with their inferred common ancestor and obtained with the REV+Γ model. The "hard" polytomy at the origin of the tree indicates that three major lineages arose from the common ancestor. The first lineage split into the lineage leading to clade 5 and clade 3, while the second lineage split into the lineage leading to clade 2 and the one leading to clade 1 and 4. Looking at the branch lengths of the tree it looks clear that these first splitting events occurred very early after the origin of the common ancestor. The third lineage, eventually giving rise to the different subtypes of clade 6, occurred much earlier than the first two, suggesting an older origin of clade 6.

Relative Nucleotide Substitution Rates Across Genes. The rate parameters estimated for each gene of the HCV genome are reported in Table 4. As expected the UTR is the slowest-evolving region (substitution rate equal to 1 by default) while, on average, the polyprotein precursor evolves about seven times faster. The core protein is also very stable, evolving only three times faster than the UTR. The fastest-evolving genes are E2 and p7, with a relative rate of 9.06 and 8.68, respectively. There is a strong rate heterogeneity across sites in each gene ($\alpha < 1$), but α values vary considerably. Compared to the average rate heterogeneity of the entire polyprotein ($\alpha = 0.37$), the UTR shows the strongest rate heterogeneity, $\alpha = 0.11$, while the p7 region shows the weakest, $\alpha = 0.69$.

Discussion

The main goal of the present study is to address some of the questions concerning the evolutionary patterns of HCV. For this reason, we analyzed 27 full-genome sequences representative of the six major phylogenetic clades of HCV. In their proposal for classification and nomenclature of the virus, Robertson et al. (1998) gave the following recommendations: (i) use the core, E1, or NS5B region for the phylogenetic reconstruction; and (ii) perform a NJ tree with the Kimura two-parameter model supported by bootstrap analysis. However, we have shown that the core and the E1 regions both experience a relatively high noise content (see Table 2 and Results). As a consequence, they should not be considered the most reliable regions for investigating the phylogeny of HCV. On the other hand, NS5B indeed appears to have the highest phylogenetic signal among the different genes (almost 95% of quartets in the tree-like area of

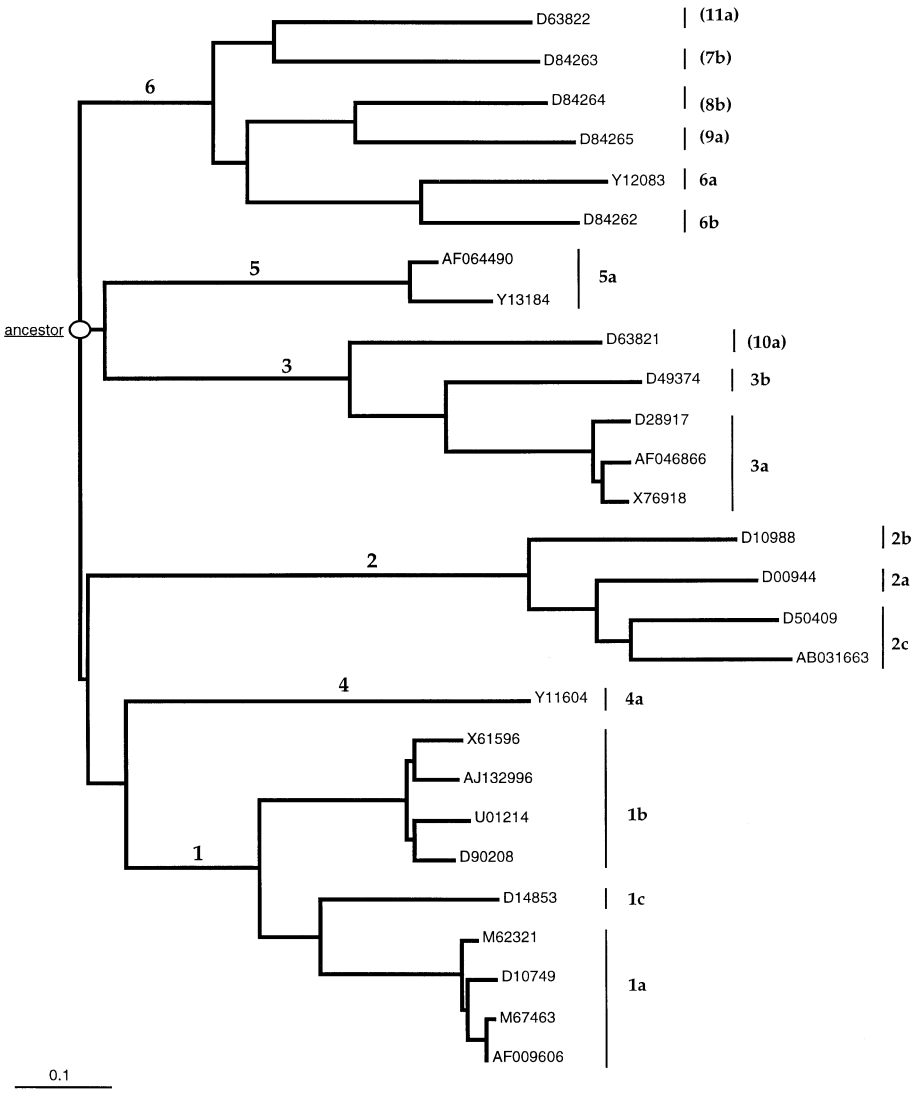

Fig. 2. Maximum likelihood tree of 27 full-genome HCV strains, obtained using the REV model with Γ-distributed rates across sites ($\alpha = 0.35$) and the branch-and-bound algorithm, which assures finding the "optimal" maximum likelihood tree. The six major lineages of HCV are indicated by the *boldface numbers* along the branches; subtypes of the viruses are also given (when subtypes are classified according to the old nomenclature, the classification is reported in parentheses). Horizontal branch lengths are drawn to scale, with the *bar* indicating 0.1 nucleotide replacement per site.

likelihood mapping; see Table 2), confirming that it can be employed reliably for the phylogenetic analysis when the full-genome sequence is not available. Another important finding is that the REV+Γ model fits the HCV data set better than simpler models, such as the Jukes and Cantor (1969) and the Kimura (1980) two-parameter model, with highly significant *p* values (see Table 3 and Results). Our data show, not surprisingly perhaps, that there is a strong rate heterogeneity across sites in the different HCV gene regions. For example, the α shape parameter of the Γ distribution estimated for the NS5B gene is 0.4. Using the wrong model to estimate genetic distances and not allowing for rate heterogeneity across sites usually leads to an underestimation of the distances and can have a negative impact on inferring the correct phylogenetic reconstruction (Sharp et al. 2000; Yang 1993, 1996). There can be several advantages in the use of ML rather than distance-based methods, such as NJ, when investigating the phylogeny of a set of taxa (Page and Holmes 1998). However, when an elevated number

Table 4. Relative nucleotide substitution rates for different genes of the HCV genome[a]

Gene region	Relative rate	T_i/T_v	α
UTR	1	4.13	0.11
CORE	2.90	2.33	0.27
E1	8.26	2.02	0.51
E2	9.06	2.40	0.36
p7	8.68	1.94	0.69
NS2	8.16	2.15	0.52
NS3	6.32	2.88	0.29
NS4A	5.98	2.73	0.39
NS4B	6.29	2.23	0.36
NS5A	7.47	2.03	0.46
NS5B	4.39	1.93	0.40
Polyprotein	7.35	2.21	0.37

[a] Relative nucleotide substitution rates were estimated with the REV model with Γ-distributed rates across sites implemented in PAML3.0 (Yang 1997) and employing the HCV tree in Fig. 2. The model allows different substitution rates, different nucleotide frequencies, and different α parameters of the Γ-distribution of each gene. The UTR region was chosen as the reference (substitution rate = 1).

of strains is considered and the use of ML is not feasible because of its computational complexity, we recommend estimating nucleotide distances among HCV strains with a model allowing for rate heterogeneity and setting the α parameter according to Table 4 for the different gene regions. The amount of rate heterogeneity across sites in the HCV gene regions gives insights into evolutionary constrains. An α value of <1 indicates that most of the sites in the sequence evolve very slowly, or are practically invariable, while a few of them, the so-called mutational hot spots, can evolve very rapidly. The lower the value of α, the more constrained the sequence. Within the HCV genes, core and NS3 show a rather high level of selective constrains and few mutational hot spots, with estimated α values of 0.27 and 0.29, respectively, the lowest among the coding regions. The core region codes for the nucleocapsid protein C of the virus (Houghton 1996). NS3 is a serine protease responsible for cleavage at the NS3/NS4A junction and at the downstream NS4A/NS4B, NS4B/NS5A, and NS5A/NS5B junctions (Houghton 1996). The NS3 protein also shows NTPase activity and is thought to be involved in helicase activity (Houghton 1996). Our results suggest that the UTR, core, and NS3 genes are the target of a strong purifying selection and their evolutionary patterns may deserve further investigation.

The molecular clock hypothesis clearly does not hold for HCV even when the three cdp are analyzed separately (see Results). The highly significant rejection of the molecular clock leaves few hopes, and any dating based on nonsynonymous as well as synonymous rates is likely to be misleading. Since a recent study demonstrated a clock-like evolution for closely related viral strains infecting different patients (Allain et al. 2000), the present result seems to imply that the rejection of the molecular clock may be due to unequal evolutionary rates among different subtypes and/or different clades of HCV. In particular, the branch lengths for HCV types 2 and 5 in the tree in Fig. 2 show that these clades have been evolving much faster and much slower, respectively, than the average of the other major genotypes. What is clear from the tree is that at least three lineages of HCV arose at the same time, followed by a fourth lineage, in a star-like burst giving origin to the different types and subtypes known today. This finding may be in agreement with the observation by Holmes et al. (1995) that HCV infected a roughly constant number of humans in an endemic state, and then, at a certain point in a relatively recent past, there was a dramatic increase in the viral population size and the virus became transmitted in an epidemic (i.e., exponentially growing) state. The predominant mode of HCV transmission is through transfer of blood or blood products, while sexual transmission seems to constitute a minor pathway of infection (Bresters et al., 1993; Mauser-Bunschoten et al. 1995). Though blood transfusions became widespread only following WWII, procedures utilizing human tissues, such as administration of yellow fever vaccines containing human serum, measles immune serum, or vaccinia virus prepared from glycerinated human lymph, started long before (Findlay and MacCallum 1937; Sawyer et al. 1944; Smith et al. 1997). In addition, hepatitis was observed at the beginning of the 20th century as a consequence of the use of unsterilized needles and syringes after injection with arsphenamine in the treatment of syphilis (Mortimer 1995). All these elements could indicate that the rapid star-like emergence of the different HCV lineages may have been fostered by the lack of sterile conditions in vaccination campaigns and the major sociocultural changes of the last two centuries. In this regard, it is certainly suggestive that the star-like evolution of HCV appears like the one of HIV-1 group M subtypes, whose origin has recently been dated to around 1920–1930 (Korber et al. 2000; Salemi et al. 2000b). Similar circumstances may be responsible for the emergence and the rapid genetic diversification of both viruses. Further studies and a detailed knowledge of the evolutionary rates in the different HCV lineages could be important to address the origin and evolution of this virus.

Acknowledgments. We are grateful to Dr. Martha Lewis for a critical reading and her help in revising the English form of the manuscript. This study was supported in part by Grant 3009894N from the Belgian Foundation for Scientific Research. M.S. is supported by the Fonds voor Wetenschappelijk Onderzoek, Vlanderen, in the category Krediet aan Navorsesrs.

References

Adams NJ, Chamberlain RW, Taylor LA, Davidson F, Lin CK, Elliott RM, Simmonds P (1998) Complete coding sequence of hepatitis C virus genotype 6a. Biochem Biophys Res Commun 234:393–396

Allain JP, Dong Y, Vandamme A-M, Moulton V, Salemi M (2000) Evolutionary rate and genetic drift of hepatitis C virus are not correlated with the immune response: Studies of infected donor-recipient clusters. J Virol 74:2541–2549

Bresters D, Mauser-Bunschoten EP, Reesink HW, et al. (1993) Sexual transmission of hepatitis C virus. Lancet 342:210–211

Bukh J, Purcell RH, Miller RH (1993) At least 12 genotypes of hepatitis C virus predicted by sequence analysis of the putative E1 gene of isolates collected worldwide. Proc Natl Acad Sci USA 90:8234–8238

Bukh J, Apgar CL, Engle R, Govindarajan S, Hegerich PA, Tellier R, Wong DC, Elkins R, Kew MC (1998) Experimental infection of chimpanzees with hepatitis C virus of 5a: Genetic analysis of the virus and generation of a standardized challenge pool. J Infect Dis 178:1193–1198

Chamberlain RW, Adams N, Saeed AA, Simmonds P, Elliot RM (1997) Complete nucleotide sequence of a type 4 hepatitis C virus variant, the predominant genotype in the Middle East. J Gen Virol 78:1341–1347

Chayama K, Tsubota A, Koida I, Arase Y, Saitoh S, Ikeda K, Kumada H (1994) Nucleotide sequence of hepatitis C virus (type 3b) isolated from a Japanese patient with chronic hepatitis C. J Gen Virol 75:3623–3628

Choo QL, Kuo G, Weiner AJ, Overby LR, Bradley DW, Houghton M (1989) Isolation of a cDNA clone derived from blood-borne non-A, non-B viral hepatitis genome. Science 244:359–362

Choo QL, Richman KH, Han JH, et al. (1991) Genetic organization and diversity of the hepatitis C virus. Proc Natl Acad Sci USA 88: 2451–2455

de Lamballerie X, Charrel RM, Attoui H, De Micco P (1997) Classification of hepatitis C virus variants in six major genotypes based in analysis of the envelope 1 and nonstructural 5B genome regions and complete polyprotein sequence. J Gen Virol 78:45–51

Farci P, Alter AJ, Wong D, Miller RH, Shin JW, Jett B, Purcell RH (1991) A long term study of hepatitis C virus replication in a non-A, non-B hepatitis. N Engl J Med 325:98–104

Felsenstein J (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. J Mol Evol 17:368–376

Felsenstein J, Churchill GA (1996) A hidden Markov model approach to variation among sites in rate of evolution. Mol Biol Evol 13:93–104

Findlay GM, MacCallum FO (1937) Note on acute hepatitis and yellow fever immunization. Trans Roy Soc Trop Med Hyg 31:297–308

Hasegawa M, Kishino H, Yano T (1985) Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22: 160–174

Hillis DM, Huelsenbeck J (1992) Signal, noise, and reliability in molecular phylogenetic analysis. J Hered 83:189–195

Holmes EC, Nee S, Rambaut A, Garnett GP, Harvey PH (1995) Revealing the history of infectious disease epidemics through phylogenetic trees. Phil Trans Roy Soc Lond B 349:33–40

Honda M, Kaneko S, Unoura M, Kobayashi K, Murakami S (1992) Sequence comparisons for a hepatitis C virus genome RNA isolated from a patient with liver cirrhosis. Gene 120:17–31

Houghton M (1996) Hepatitis C virus. In: BN Fields, DM Knipe, PM Howley, et al. (eds) Virology. Lippincott, Raven Press, Philadelphia, pp 1035–1051

Huelsenbeck J, Rannala B (1997) Phylogenetic methods come of age: Testing hypotheses in an evolutionary context. Science 276:227–232

Ina Y, Mizokami M, Ohba K, Gojobori T (1994) Reduction of synonymous substitutions in the core protein gene of hepatitis C virus. J Mol Evol 38:50–56

Inchauspe G, Zebedee S, Lee DH, Sugitani M, Nasoff M, Price AM (1991) Genomic structure of the human prototype strain H of hepatitis C virus: Comparison with American and Japanese isolates. Proc Natl Acad Sci USA 88:10292–10296

Jukes TH, Cantor CR (1969) In: HN Munro (ed) Mammalian protein methabolism. Academic Press, New York, pp 21–123

Kato N, Hijikata M, Ootsuyama Y, Nakagawa M, Ohkoshi S, Sugimura T, Shimotohno K (1990) Molecular cloning of the human hepatitis C virus genome from Japanese patients with non-A, non-B hepatitis. Proc Natl Acad Sci USA 87:9524–9528

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120

Kolykhalov AA, Agapov EV, Blight KJ, Mihalik K, Feinstone SM, Rice CM (1997) Transmission of hepatitis C by intrahepatic inoculation with transcribed RNA. Science 277:570–574

Korber B, Muldoon M, Theiler J, Gao F, Gupta R, Lapedes A, Hahn BH, Wolinsky S, Bhattacharya T (2000) Timing the ancestor of the HIV-1 pandemic strains. Science 288:1789–1796

Li WH (1997) Molecular evolution. Sinauer Associates, Sunderland, MA

Mauser-Bunschoten EP, Bresters D, Reesink HW (1995) Transmission of hepatitis C virus to spouses. Ann Intern Med 122:154–155

Mellor J, Holmes EC, Jarvis LM, Yap PL, Simmonds P, International HCV Collaborative Study Group (1995) Investigation of the pattern of hepatitis C virus sequence diversity in different geographical regions: Implications for virus classification. J Gen Virol 76:2493–2507

Mortimer PP (1995) Arsphenamine jaundice and the recognition of instrument-borne virus infection. Genitourinary Med 71:109–119

Nagayama K, Kurosaki M, Enomoto N, Miyasaka Y, Marumo F, Sato C (2000) Characteristics of hepatitis C genome associated with disease progression. Hepatology 31:745–750

Nakao H, Okamoto H, Tokita H, Inoue T, Iizuka H, Pozzato G, Mishiro S (1996) Full-length genomic sequence of a hepatitis C virus genotype 2c isolate (BEBE1) and the 2c-specific PCR primers. Arch Virol 141:701–704

Nieselt-Struve K (1998) Combining likelihood-mapping and statistical geometry to a new sequence analysis tool. In: MK Uyenoyama, A von Haeseler (eds) Proceedings of the Trinational Workshop on Molecular Evolution, University of Munich, Munich, Germany, June 5–7, 1997

Nishioka K, Watanabe J, Furuta S, et al. (1991) A high prevalence of antibody to the hepatitis C virus in patients with hepatocellular carcinoma in Japan. Cancer 67:429–433

Okamoto H, Okada S, Sugiyama Y, Kurai K, Iizuka H, Machida A, Miyakawa Y, Mayumi M (1991) Nucleotide sequence of the genomic RNA of hepatitis C virus isolated from a human carrier: Comparison with reported isolates for conserved and divergent regions. J Gen Virol 72:2697–2704

Okamoto H, Kanai N, Mishiro S (1992a) Full-length nucleotide sequence of a Japanese hepatitis C virus isolate (HC-J1) with high homology to USA isolates. Nucleic Acids Res 20:6410

Okamoto H, Kurai K, Okada S, Yamamoto K, Lizuka H, Tanaka T, Fukuda S, Tsuda F, Mishiro S (1992b) Full-length sequence of a hepatitis C virus genome having poor homology to reported isolates: Comparative study of four distinct genotypes. Virology 188: 331–341

Okamoto H, Kojima M, Sakamoto M, Iizuka H, Hadiwandowo S, Suwignyo S, Miyakawa Y, Mayumi M (1994) The entire nucleotide sequence and classification of a hepatitis C virus isolate of a novel genotype from an Indonesian patient with chronic liver disease. J Gen Virol 75:629–635

Page RDM, Holmes EC (1998) Molecular evolution: A phylogenetic approach. Blackwell Science, Oxford, UK

Rispeter K, Lu M, Lechner S, Zibert A, Roggendorf M (1997) Cloning and characterization of a complete open reading frame of the hepatitis C virus genome in only two cDNA fragments. J Gen Virol 78:2751–2759

Robertson B, Myers G, Howard C, et al. (1998) Classification, nomenclature, and database development for hepatitic C virus (HCV) and related viruses: Proposals for standardization. Arch Virol 143: 2493–2503

Saito I, Miyamura T, Ohbayashi A, et al. (1990) Hepatitis C virus infection is associated with development of hepatocellular carcinoma. Proc Natl Acad Sci USA 87:6547–6549

Salemi M, Desmyter J, Vandamme A-M (2000a) Tempo and mode of human and simian T-lymphotropic viruses (HTLV/STLVs) evolution revealed by analyses of full genome sequences. Mol Biol Evol 17:374–386

Salemi M, Strimmer K, Hall WW, Duffy M, Delaporte E, Mboup S, Peeters M, Vandamme A-M (2000b) Dating the common ancestor of SIVcpz and HIV-1 group M and the origin of HIV-1 subtypes using a new method to uncover clock-like molecular evolution. FASEB J 10.1096/fj.00-0449fje

Samokhvalov EI, Hijikata M, Gylka RI, Lvov DK, Mishiro S (2000) Full-genome nucleotide sequence of a hepatitis C virus variant (isolate name VAT96) representing a new subtype within the genotype 2. Virus Genes 20:183–187

Sawyer WA, Meyer KF, Eaton MD, Bauer JH, Putnam P, Schwentker FF (1944) Jaundice in army personnel in western region of United States and its relation to vaccination against yellow fever. Am J Hyg 39:337–430

Sharp PM, Bailes E, Gao F, Beer BE, Hirsch VM, Hahn BH (2000) Origins and evolution of AIDS viruses: Estimating the time-scale. Biochem Soc Trans 28:275–282

Simmonds P (1995) Variability of hepatitis C virus. Hepatology 21: 570–583

Simmonds P, Alberti A, Alter HJ, et al. (1994a) A proposed system for the nomenclature of hepatitis C viral genotypes. Hepatology 19: 1321–1324

Simmonds P, Smith DB, McOmish F, Yap PL, Kolberg J, Urdea MS, Holmes EC (1994b) Identification of genotypes of hepatitis C virus by sequence comparisons in the core, E1 and NS-5 regions. J Gen Virol 75:1053–1061

Simmonds P, Mellor J, Sakuldamrongpanich T, Nuchaprayoon C, Tanprasert S, Holmes EC, Smith DB (1996) Evolutionary analysis of variants of hepatitis C virus found in South East Asia: Comparison with classification based upon sequence similarity. J Gen Virol 77:3013–3024

Smith DB, Pathirana S, Davidson F, Lawlor E, Power J, Yap PL Simmonds P (1997) The origin of hepatitis C virus genotypes. J Gen Virol 78:321–328

Strimmer K, von Haeseler A (1997) Likelihood mapping: A simple method to visualize phylogenetic content of a sequence alignment. Proc Natl Acad Sci USA 94:6814–6819

Suzuki Y, Yamaguchi-Kabata Y, Gojobori T (2000) Nucleotide substitution rates of HIV-1. AIDS Rev 2:39–47

Tamura K, Nei M (1993) Estimation of the number of nucleotide substitution in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10:512–526

Tokita H, Okamoto H, Iizuka H, Kishimoto J, Tsuda F, Lesmana LA, Miyakawa Y, Mayumi M (1996) Hepatitis C virus variants from Jakarta, Indonesia classifiable into novel genotypes in the second (2e and 2f), tenth (10a) and eleventh (11a) genetic groups. J Gen Virol 77:293–301

Tokita H, Okamoto H, Iizuka H, Kishimoto J, Tsuda F, Miyakawa Y, Mayumi M (1998) The entire nucleotide sequences of three hepatitis C virus isolates in genetic groups 7–9 and comparison with those in the other eight genetic groups. J Gen Virol 79:1847–1857

Weiner AJ, Brauer MJ, Rosenblatt J, Richman KH, Tung J, Crawford K, Bonino F, Saracco G, Choo QL, Houghton M, Han JH (1991) Variable and hypervariable domains are found in regions of HCV corresponding to the Flavivirus envelope and NS1 proteins and the Pestiviruses envelope glycoproteins. Virology 180:842–848

Yamada N, Tanihara K, Mizokami M, Ohba K, Takada A, Tsutsumi M, Date T (1994) Full-length sequence of the genome of hepatitis C virus type 3a: comparative study different genotypes. J Gen Virol 75:3279–3284

Yang Z (1993) Maximum likelihood estimation of the phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol 10:1396–1401

Yang Z (1996) Among-site rate variation and its impact on phylogenetic analysis. Tree 11:367–372

Yang Z (2000) Phylogenetic analysis by maximum-likelihood (PAML), version 3.0. University College London, London

Yang Z, Roberts D (1995) On the use of nucleic acid sequences to infer early branchings in the tree of life. Mol Biol Evol 12:451–548

Xia X (2000a) DAMBE 4.0 (software package for data analysis in molecular biology and evolution). Department of Ecology and Biodiversity, University of Hong Kong, Hong Kong

Xia X (2000b) Data analysis in molecular biology and evolution. Kluwer Academic, Boston