Hepatitis C virus F protein sequence reveals a lack of functional constraints and a variable pattern of amino acid substitution

Juan Cristina,[1] Fernando Lopez,[1] Gonzalo Moratorio,[1] Lilia López,[1,2] Silvia Vasquez,[3] Laura García-Aguirre[1] and Ausberto Chunga[4]

Correspondence
Juan Cristina
cristina@cin1.cin.edu.uy

[1]Laboratorio de Virología Molecular, Centro de Investigaciones Nucleares, Facultad de Ciencias, Iguá 4225, 11400 Montevideo, Uruguay

[2]Cátedra de Hemoterapia, Facultad de Medicina, Av. Italia s/n, Montevideo, Uruguay

[3]Instituto de Investigaciones Clínicas, Facultad de Medicina 'San Fernando', Universidad Nacional Mayor de San Marcos, Parque de la Medicina, Avenida Grau Cuadra 13 s/n, Lima 01, Peru

[4]Servicio de Inmunología, Hospital Nacional Edgardo Rebagliati Martins HNERN, Domingo Cueto s/n, Jesús María, Lima 11, Peru

Hepatitis C virus (HCV) is an important human pathogen that affects 170 million people worldwide. The HCV genome is an RNA molecule that is approximately 9·6 kb in length and encodes a polyprotein that is cleaved proteolytically to generate at least 10 mature viral proteins. Recently, a new HCV protein named F has been described, which is synthesized as a result of a ribosomal frameshift. Little is known about the biological properties of this protein, but the possibility that the F protein may participate in HCV morphology or replication has been raised. In this work, the presence of functional constraints in the F protein was investigated. It was found that the rate of amino acid substitutions along the F protein was significantly higher than the rate of synonymous substitutions, and comparisons involving genes that represented independent phylogenetic lineages yielded very different divergence/conservation patterns. The distribution of stop codons in the F protein across all HCV genotypes was also investigated; genotypes 2 and 3 were found to have more stop codons than genotype 1. The results of this work suggest strongly that the pattern of divergence in the F protein is not affected by functional constraints.

Received 9 August 2004
Accepted 27 September 2004

INTRODUCTION

Hepatitis C virus (HCV) is the major causative agent of post-transfusion hepatitis and parenterally transmitted, non-A, non-B hepatitis throughout the world (Alter & Seeff, 2000). HCV is an enveloped RNA virus that is classified in the family *Flaviviridae*. HCV has high genomic variability and at least six different genotypes and an increasing number of subtypes have been reported (Simmonds, 1999). The HCV genome is approximately 9·6 kb in length and encodes a polyprotein that is cleaved proteolytically to generate at least 10 mature viral protein products (Reed & Rice, 2000). Recently, a new protein named F has been described; it is expressed as a result of a ribosomal frameshift within the capsid-encoding sequence, a mechanism unique among members of the family *Flaviviridae* (Xu *et al.*, 2001; Choi *et al.*, 2003). This protein was localized in the cytoplasm of infected cells, with a notable perinuclear localization (Roussel *et al.*, 2003), and was found to be associated with the endoplasmic reticulum (Xu *et al.*, 2003). This subcellular localization of HCV F protein is similar to that of the HCV core and NS5A proteins, raising the hypothesis that the F protein may participate in HCV morphogenesis or replication (Xu *et al.*, 2003). In addition, sera from patients who were positive for HCV genotype 1a or 1b were shown to react differently to synthetic peptides encoded by the F reading frame, and these findings have suggested genotype-dependent specific features for the F protein (Boulant *et al.*, 2003). In order to contribute to elucidating these matters, we performed an analysis of genetic variability and amino acid substitution rates for the HCV F protein.

The GenBank/EMBL/DDBJ accession numbers for the sequences reported in this work are AJ582128–AJ582131 and AJ781117–AJ781124.

METHODS

Serum samples. Serum samples were obtained from 20 patients with chronic hepatic disease from Hospital Nacional Eduardo Rebagliati Martins (Lima, Peru) and Hospital de Clinicas (Montevideo,

Uruguay). The Peruvian patients were screened by using an enzyme immunoassay (Innogenetics) and a confirmatory line immunoassay test (Innogenetics), according to the manufacturer's instructions. The Uruguayan patients were screened by using an enzyme immunoassay (Abbott) according to the manufacturer's instructions.

RNA extraction and cDNA synthesis and amplification. HCV RNA was extracted from serum samples (100 µl) by using a QIAamp viral RNA kit (Qiagen) according to the manufacturer's instructions. Extracted RNA was eluted from the columns with 50 µl RNase-free water, and cDNA synthesis and PCR amplification of the core region were carried out as described by Bukh *et al.* (1994). To avoid false-positive results, the recommendations of Kwok & Higuchi (1989) were strictly adhered to. Amplicons were purified by using a QIAquick PCR purification kit (Qiagen) according to the manufacturer's instructions.

Sequencing. The primers used for amplification were also used for sequencing the PCR fragments. All PCR fragments were sequenced in both directions to avoid discrepancies. The sequencing reaction was carried out by using a BigDye DNA sequencing kit on a 373 DNA sequencer apparatus (both from Perkin Elmer).

Sequence analysis. The amino acid sequences of the core protein, as well as the F protein (obtained from the core gene in the F reading frame), were aligned by using the CLUSTAL W program (Thompson *et al.*, 1994).

Substitution-rate analysis. Substitution rates along the HCV F protein were measured by using a sliding window. Pairwise nucleotide distances (synonymous and non-synonymous) within each window were estimated by the method of Comeron (1995), as implemented in the computer program K-estimator. For those windows where the method was inapplicable (due to the negative argument of the logarithm), we used the Jukes–Cantor method (Jukes & Cantor, 1969). The window size used was 30 codons, shifting three codons at a time. The ratios of non-synonymous (dn) to synonymous (ds) substitutions for the core and F proteins were calculated by using data obtained from the computer program SNAP, as implemented by Korber (2000).

RESULTS

Amino acid substitution rates in the F and core regions across HCV genotypes

In order to gain insight into the pattern of amino acid substitutions in the F protein, the deduced HCV F amino acid sequences, obtained from the core gene sequences from the South American patients, were aligned with those from 12 other strains that were representative of all six HCV types isolated elsewhere for which total sequences have been obtained. The origin of the sequences and the strains used are listed in Table 1. Once aligned, we compared the dn/ds ratios obtained for the F and core proteins from all pairwise comparisons among all strains involved in these studies. Examples of the results of these studies are shown in Table 2. Unexpectedly, very high dn/ds ratios were found for the F protein in comparison with the values found for the core protein. The mean dn/ds ratio for all pairwise comparisons for the F protein was 2·5, whereas the mean ratio for the core protein was 0·10.

Table 1. Origins of HCV strains

Genotype	Name	Geographical location	GenBank accession no.
1a	H77	USA	AF009606
1	J1	Japan	D10749
1b	JK1	Japan	X61596
1	HD1	Germany	U45476
1b	K1R2	Japan	D50482
2	J6	Indonesia	D00944
2	J8	Indonesia	D10988
3a	NZL1	Japan	D17763
3a	K3a	Japan	D28917
3	V-D	Germany	X76918
4a	ED43	Egypt	Y11604
5a	EUH	UK	Y13184
6a	euhk	Hong Kong	Y12083
1	PE8	Peru	AJ582128
1	PE96	Peru	AJ582129
1	PE108	Peru	AJ582130
RF2_1a/1b	PE22	Peru	AJ582131
1	PE118	Peru	AJ781123
1	PE22-2	Peru	AJ781119
1	PE127	Peru	AJ781117
1	PE153	Peru	AJ781124
1	PE15	Peru	AJ781118
1	URG10	Uruguay	AJ781122
1	URL1	Uruguay	AJ781120
1	UR5419	Uruguay	AJ781121

Within-gene covariation between synonymous and non-synonymous substitutions

In order to estimate substitution rates along all regions of the F protein, we used a sliding-window analysis to estimate variation in the rates of synonymous and non-synonymous substitutions within the F protein.

As shown in Fig. 1, the rates of non-synonymous substitutions were significantly higher than those of synonymous substitutions for all pairwise comparisons, in agreement with previous results (Table 2). Interestingly, the profiles of synonymous and non-synonymous distances exhibited low covariation. This means that those regions of the F protein that are more divergent at the amino acid level are not more divergent at the synonymous level (see Fig. 1).

For the purpose of testing whether the observed pattern of divergence was governed by a deterministic force, such as natural selection, it was necessary to analyse processes of divergence between phylogenetically independent lineages. Therefore, as shown in Fig. 1, we obtained the profiles of synonymous and non-synonymous distances for different HCV genotypes and subtypes. The differences between the pairs of profiles were evident for all examples shown (see Fig. 1). This indicated that the pattern of conservation/divergence along the F protein was not due to deterministic

Table 2. Amino acid substitution rates for the F and core proteins across all HCV genotypes

Name (genotype)		F protein			Core protein		
		ds	dn	dn/ds	ds	dn	dn/ds
J6 (2)	H77 (1a)	0·06	0·24	4·00	0·75	0·06	0·08
J6 (2)	JK1 (1b)	0·07	0·22	3·14	0·65	0·07	0·11
J6 (2)	ED43 (4a)	0·07	0·21	3·00	0·62	0·06	0·10
J6 (2)	EUH (5a)	0·09	0·23	2·56	0·69	0·07	0·10
J6 (2)	euhk (6a)	0·09	0·27	3·00	0·83	0·09	0·11
J6 (2)	NZL1 (3a)	0·07	0·24	3·43	0·74	0·07	0·09
J6 (2)	V-D (3)	0·08	0·23	2·88	0·68	0·07	0·10
J6 (2)	K3a (3a)	0·1	0·28	2·80	0·8	0·1	0·13
H77 (1a)	JK1 (1b)	0·04	0·1	2·50	0·28	0·02	0·07
H77 (1a)	ED43 (4a)	0·05	0·16	3·20	0·56	0·03	0·05
H77 (1a)	EUH (5a)	0·11	0·18	1·64	0·48	0·07	0·15
H77 (1a)	euhk (6a)	0·06	0·22	3·67	0·71	0·05	0·07
H77 (1a)	NZL1 (3a)	0·07	0·16	2·29	0·47	0·04	0·09
H77 (1a)	V-D (3)	0·07	0·17	2·43	0·5	0·05	0·10
H77 (1a)	K3a (3a)	0·09	0·21	2·33	0·61	0·06	0·10
JK1 (1b)	ED43 (4a)	0·08	0·18	2·25	0·59	0·04	0·07
JK1 (1b)	EUH (5a)	0·12	0·18	1·50	0·47	0·08	0·17
JK1 (1b)	euhk (6a)	0·09	0·2	2·22	0·57	0·07	0·12
JK1 (1b)	NZL1 (3a)	0·08	0·2	2·50	0·57	0·06	0·11
JK1 (1b)	V-D (3)	0·08	0·21	2·63	0·61	0·07	0·11
JK1 (1b)	K3a (3a)	0·1	0·24	2·40	0·63	0·09	0·14
ED43 (4a)	EUH (5a)	0·12	0·19	1·58	0·49	0·08	0·16
ED43 (4a)	euhk (6a)	0·1	0·22	2·20	0·83	0·05	0·06
ED43 (4a)	NZL1 (3a)	0·05	0·21	4·20	0·77	0·03	0·04
ED43 (4a)	V-D (3)	0·05	0·22	4·40	0·78	0·04	0·05
ED43 (4a)	K3a (3a)	0·08	0·26	3·25	0·89	0·07	0·08
EUH (5a)	euhk (6a)	0·16	0·23	1·44	0·64	0·1	0·16
EUH (5a)	NZL1 (3a)	0·11	0·2	1·82	0·51	0·09	0·18
EUH (5a)	V-D (3)	0·11	0·21	1·91	0·5	0·09	0·18
EUH (5a)	K3a (3a)	0·13	0·24	1·85	0·6	0·11	0·18
euhk (6a)	NZL1 (3a)	0·09	0·2	2·22	0·63	0·06	0·10
euhk (6a)	V-D (3)	0·09	0·22	2·44	0·72	0·06	0·08
euhk (6a)	K3a (3a)	0·09	0·24	2·67	0·68	0·08	0·12
NZL1 (3a)	V-D (3)	0	0·04	NA	0·1	0	0·00
NZL1 (3a)	K3a (3a)	0·01	0·07	7·00	0·13	0·03	0·23
V-D (3)	K3a (3a)	0·01	0·1	10·00	0·19	0·03	0·16

NA, Not applicable.

forces and that the intragenic distribution of synonymous and non-synonymous substitutions was random in the HCV F protein.

Distribution of stop codons in the F protein across all HCV genotypes

In order to gain insight into the functionality of the F protein, we studied the distribution of stop codons across all HCV genotypes and subtypes available in the HCV databases. As shown in Table 3, some genotypes, particularly 2 and 3, had more stop codons in their F proteins than subtypes 1a and 1b. This means that the overall structure of the F protein varies greatly among the different genotypes, which may have important consequences in relation to the functionality of the protein in different genotypes and subtypes.

DISCUSSION

HCV F protein is a newly discovered HCV gene product that is expressed by a translational ribosomal frameshift in the core gene. Although ribosomal frameshifting for gene expression has been demonstrated for RNA viruses of several different families, including retroviruses (Jacks & Varmus, 1985), coronaviruses (Brierley *et al.*, 1989) and

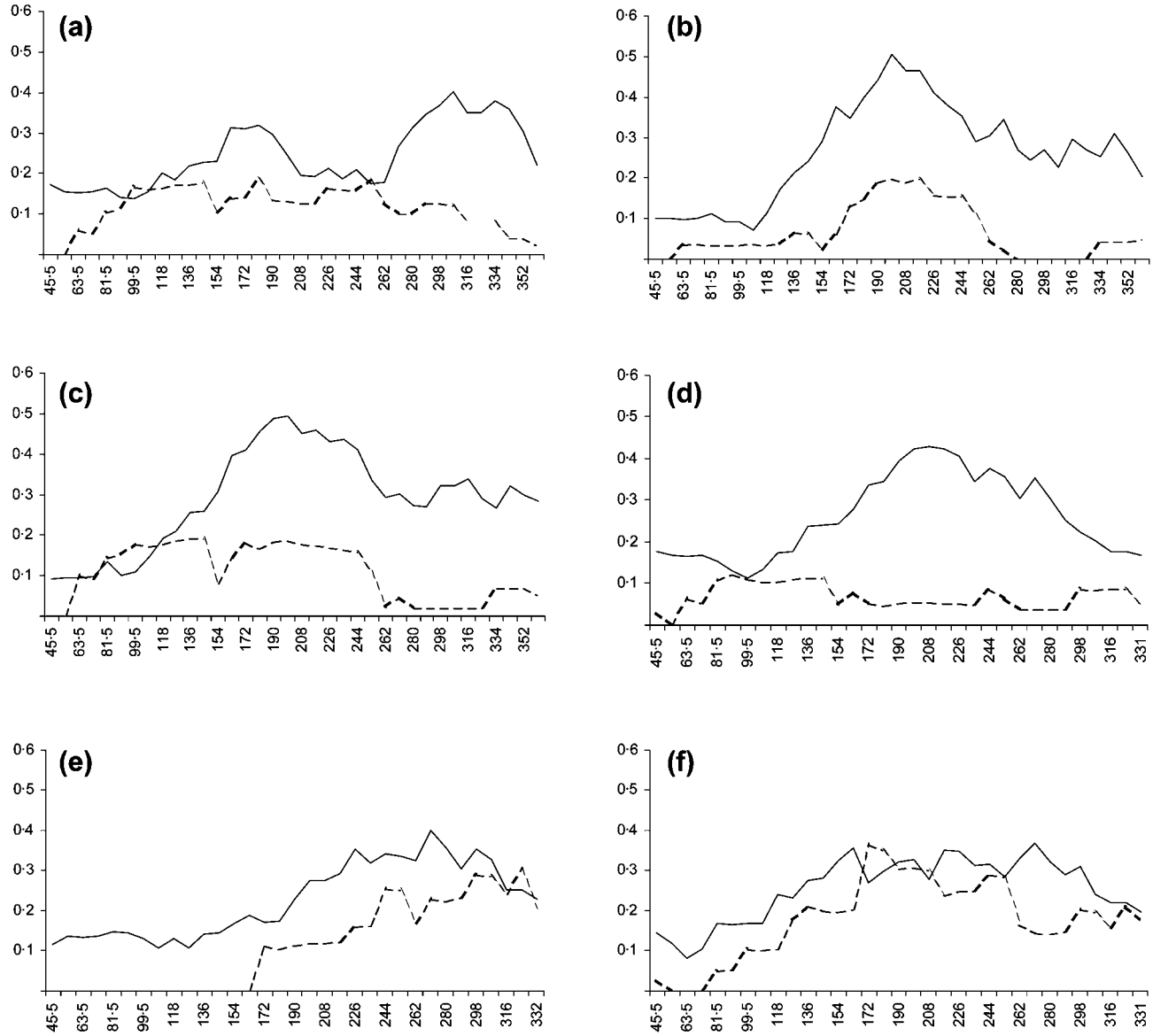

Fig. 1. Profiles of synonymous and non-synonymous distances in the HCV F protein. Numbers on the *y* axis denote distance. Numbers on the *x* axis show the codon positions in the mid-point of the window. Synonymous substitutions are shown by a broken line and non-synonymous substitutions by a solid line. The following comparisons are shown: (a) strains JK1 (genotype 1b) and V-D (genotype 3); (b) strains H77 (genotype 1a) and J8 (genotype 2); (c) strains J8 (genotype 2) and V-D (genotype 3); (d) strains V-D (genotype 3) and ED43 (genotype 4); (e) strains ED43 (genotype 4) and EUH (genotype 5); and (f) strains EUH (genotype 5) and euhk (genotype 6).

astroviruses (Marczinke *et al.*, 1994), HCV is the first example within the family *Flaviviridae* to use this mechanism to express a gene that is embedded totally in another coding sequence. Little is known about the biological properties of the HCV F protein (Xu *et al.*, 2003). It is a labile protein with a half-life of less than 10 min in Huh7 hepatoma cells and *in vitro* (Xu *et al.*, 2003).

Strikingly, we found very high dn/ds ratios for all pairwise comparisons across all HCV genotypes for the F protein (i.e. dn/ds > 1; see Table 2). One possible explanation for the F protein having such a large dn/ds ratio is that its non-synonymous variation is simply the result of variation in the overlapping core gene, which is in a different reading frame. Nevertheless, these results showed a different pattern of amino acid substitutions for the F protein than for the the core protein and other regions of the HCV genome.

We found a high degree of genetic variability and amino acid substitution rates along the F protein (Fig. 1). This is in contrast to the results commonly found in other viral

Table 3. Distribution of stop codons in the F protein across all HCV genotypes

Genotype	No. sequences examined	No. sequences with stop codons	Stop codon position (aa)
1a	29	29	161
1b	96	46	161
		49	143
2a	32	32	125
2b	30	2	138
		4	143
		2	150
		22	161
2c	15	15	125
2e	4	4	154
2f	3	2	154
		1	189
2g	1	1	9
3a	66	9	28
		24	148
		33	143
3b	12	12	125
3e	1	1	143
3f	1	1	125
3h	4	4	21
3k	5	3	9
		2	139
4a	4	2	21
		3	125
4c	4	4	125
4d	2	2	125
4e	3	3	125
4f	1	1	131
4r	1	1	21
5a	13	13	125
6k	5	5	21
6a	8	8	125
6b	3	3	125
6g	4	4	139

systems, such as human immunodeficiency virus (Zanotto *et al.*, 1999) and hepatitis A virus (Costa-Mattioli *et al.*, 2003), and even with other HCV proteins, such as the core protein (not shown). This suggests that deterministic forces are not acting to conserve a particular domain or region.

The results of this work are in agreement with previous reports indicating that the F protein displays no clear sequence homologies to other proteins of known function, except that it is highly basic (Xu *et al.*, 2001). Interestingly, the F protein does not appear to be essential for viral RNA replication, as its absence did not abolish the replication of an HCV RNA replicon in Huh7 hepatoma cells (Lohmann *et al.*, 1999; Blight *et al.*, 2000).

HCV chimeras have been constructed and shown to be infectious (Yagani *et al.*, 1998) and the HCV F protein has been expressed (Roussel *et al.*, 2003). Taking this into account, specific experiments can be designed to determine whether HCV F proteins from different HCV genotypes show differences in specific functions, such as morphogenesis, replication and ligand interactions. These will provide a definitive picture of the role of the F protein in the biology of HCV.

ACKNOWLEDGEMENTS

We acknowledge support from ICGEB, PAHO and RELAB through Project CRP.LA/URU03-032. We thank Dr Fabián Alvarez-Valin, from Sección Biomatemáticas, Facultad de Ciencias, Montevideo, Uruguay, for helpful discussions. We thank anonymous reviewers from previous versions of this manuscript for helpful suggestions.

REFERENCES

Alter, H. J. & Seeff, L. B. (2000). Recovery, persistence, and sequelae in hepatitis C virus infection: a perspective on long-term outcome. *Semin Liver Dis* **20**, 17–35.

Blight, K. J., Kolykhalov, A. A. & Rice, C. M. (2000). Efficient initiation of HCV RNA replication in cell culture. *Science* **290**, 1972–1974.

Boulant, S., Becchi, M., Penin, F. & Lavergne, J.-P. (2003). Unusual multiple recoding events leading to alternative forms of hepatitis C virus core protein from genotype 1b. *J Biol Chem* **278**, 45785–45792.

Brierley, I., Digard, P. & Inglis, S. C. (1989). Characterization of an efficient coronavirus ribosomal frameshifting signal: requirement for an RNA pseudoknot. *Cell* **57**, 537–547.

Bukh, J., Purcell, R. H. & Miller, R. H. (1994). Sequence analysis of the core gene of 14 hepatitis C virus genotypes. *Proc Natl Acad Sci U S A* **91**, 8239–8243.

Choi, J., Xu, Z. & Ou, J. (2003). Triple decoding of hepatitis C virus RNA by programmed translational frameshifting. *Mol Cell Biol* **23**, 1489–1497.

Comeron, J. M. (1995). A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. *J Mol Evol* **41**, 1152–1159.

Costa-Mattioli, M., Ferré, V., Casane, D. & 7 other authors (2003). Evidence of recombination in natural populations of hepatitis A virus. *Virology* **311**, 51–59.

Jacks, T. & Varmus, H. E. (1985). Expression of the Rous sarcoma virus *pol* gene by ribosomal frameshifting. *Science* **230**, 1237–1242.

Jukes, T. H. & Cantor, C. R. (1969). Evolution of protein molecules. In *Mammalian Protein Metabolism*, pp. 21–132. Edited by H. N. Munro. New York: Academic Press.

Korber, B. (2000). HIV sequence signatures and similarities. In *Computational and Evolutionary Analysis of HIV Molecular Sequences*, pp. 55–72. Edited by A. G. Rodrigo & G. H. Learn, Jr. Dordrecht: Kluwer.

Kwok, S. & Higuchi, R. (1989). Avoiding false positives with PCR. *Nature* **339**, 237–238.

Lohmann, V., Körner, F., Koch, J.-O., Herian, U., Theilmann, L. & Bartenschlager, R. (1999). Replication of subgenomic hepatitis C virus RNAs in a hepatoma cell line. *Science* **285**, 110–113.

Marczinke, B., Bloys, A. J., Brown, T. D. K., Willcocks, M. M., Carter, M. J. & Brierley, I. (1994). The human astrovirus RNA-dependent RNA polymerase coding region is expressed by ribosomal frameshifting. *J Virol* **68**, 5588–5595.

Reed, K. E. & Rice, C. M. (2000). Overview of hepatitis C virus genome structure, polyprotein processing, and protein properties. *Curr Top Microbiol Immunol* **242**, 55–84.

Roussel, J., Pillez, A., Montpellier, C., Duverlie, G., Cahour, A., Dubuisson, J. & Wychowski, C. (2003). Characterization of the expression of the hepatitis C virus F protein. *J Gen Virol* **84**, 1751–1759.

Simmonds, P. (1999). Viral heterogeneity of the hepatitis C virus. *J Hepatol* **31** (Suppl. 1), 54–60.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673–4680.

Xu, Z., Choi, J., Yen, T. S. B., Lu, W., Strohecker, A., Govindarajan, S., Chien, D., Selby, M. J. & Ou, J. (2001). Synthesis of a novel hepatitis C virus protein by ribosomal frameshift. *EMBO J* **20**, 3840–3848.

Xu, Z., Choi, J., Lu, W. & Ou, J. (2003). Hepatitis C virus F protein is a short-lived protein associated with the endoplasmic reticulum. *J Virol* **77**, 1578–1583.

Yagani, M., St Claire, M., Shapiro, M., Emerson, S. U., Purcell, R. H. & Bukh, J. (1998). Transcripts of a chimeric cDNA clone of hepatitis C virus genotype 1b are infectious *in vivo*. *Virology* **244**, 161–172.

Zanotto, P. M. de A., Kallas, E. G., de Souza, R. F. & Holmes, E. C. (1999). Genealogical evidence for positive selection in the *nef* gene of HIV-1. *Genetics* **153**, 1077–1089.