African Journal of Biotechnology Vol. 9 (29), pp. 4561-4567, 19 July, 2010
Available online at http://www.academicjournals.org/AJB
ISSN 1684–5315 ©2010 Academic Journals

Full Length Research Paper

Sequence comparison and phylogenetic analysis of core gene of hepatitis C virus from Pakistani population

Yasir Waheed[#]*, Sadia Tahir[#], Tahir Ahmad and Ishtiaq Qadri

National University of Sciences and Technology (NUST) Center of Virology and Immunology, H-12 Sector, Islamabad, Pakistan.

Accepted 21 June, 2010

In Pakistan, more than 10 million people are living with hepatitis C virus (HCV) with high morbidity and mortality. The aims of the present study are to report HCV core gene sequences from Pakistani population and perform their sequence comparison/phylogenetic analysis. The core gene of HCV has been cloned from six different patients and sequences submitted at the National Center of Biotechnology Information (NCBI). Nucleotides and deduced amino acid sequence comparison of six isolates was performed with each other and with two HCV genotype 3a type examples reported from Japan. Phylogenetic tree of HCV core sequences was constructed using CLC software. Nucleotides sequence comparison showed that our sequences have 94 to 96% homology with NZL1 strain and 90 to 93% homology with HCV-K3A/650 strain. Deduced amino acid sequence comparison showed that our sequences have 92 to 98% homology with NZL1 strain and 88 to 94% homology with HCV-K3A/650 strain. Phylogenetic analysis suggests that our sequences are clustered with sequences reported from Japan. This is the first phylogenetic analysis of HCV core gene from Pakistani population. Our sequences and sequences from Japan are grouped into same cluster in the phylogenetic tree. Sequence comparison and phylogenetic analysis showed that our isolates have high homology with Japanese isolates.

Key words: Hepatitis C virus, core, phylogenetic analysis, Pakistan.

INTRODUCTION

Hepatitis C virus (HCV) was discovered in 1989 as the major causative agent of non-A and non-B hepatitis (Rice, 1996). It belongs to Flaviviridae family and is a plus stranded RNA virus (Choo et al., 1999). About 200 million people are infected with HCV worldwide, which covers about 3.3% of the world population (Waheed et al., 2009). Most patients with HCV persistent infection develop chronic hepatitis, fibrosis and even liver cancer (Gao et al., 2009). It was estimated by the World Health Organization in 2004 that the annual deaths due to liver cancer caused by HCV and cirrhosis were 308,000 and 785,000, respectively. HCV has six major genotypes and their distribution depends on the geographic area and transmission routes. In Pakistan, the major genotype of HCV is 3a (Waheed et al., 2009; Turhan et al., 2005).

Pakistan is a developing country of 180 million people with low health and educational standards. Due to non implementation of international standards regarding blood transfusions, reuse of syringes and needles, poor sterilization practice by doctors, dentists and barbers, lack of awareness, the prevalence of HCV is increasing. HCV prevalence is 4.95% in general and 57% in IDU population of Pakistan (Waheed et al., 2009).

The genome of HCV comprises of a linear RNA molecule of 9600 nucleotides, with single open reading frame encoding a polyprotein precursor of 3000 amino acids. The 5' region of the viral RNA encodes for the structural proteins (C, E1 and E2), followed by the nonstructural proteins (NS2, NS3, NS4A, NS4B, NS5A and NS5B).

*Corresponding author. E-mail: yasir_waheed_199@hotmail.com. Tel: +92-300-5338171. Fax: +92-51-9271593.

Abbreviations: HCV, Hepatitis C virus; **NLS,** nuclear localization signals; **ER,** endoplasmic reticulum; **EDTA,** ethylenediaminetetraacetic acid; **cDNA,** complementary DNA; **X-Gal,** bromo-4-chloro-3-indolyl-b-D-galactopyranoside; **IPTG,** isopropyl β-D-1-thiogalactopyranoside; **DTCS,** dye terminator cycle sequencing; **PCR,** polymerase chain reaction.

[#] These authors contributed equally to this work.

Table 1. Demography of patients.

Accession no.	Sex	Age	Genotype	ALT	Viral Titer
GQ180059	Male	24 years	3a	56	5.6×10^6
GQ180060	Female	28 years	3a	16	5.6×10^5
GQ180061	Male	40 years	3a	26	1.5×10^5
GQ180062	Female	32 years	3a	68	6.5×10^7
GQ180063	Female	42 years	3a	55	1.1×10^6
GQ180064	Female	35 years	3a	78	6.7×10^5

Both host and viral proteases cleave the polyprotein into at least 10 different proteins (Major and Feinstone, 1997; Santolini et al., 1994). The HCV core protein is highly basic, RNA-binding protein which is responsible for the formation of viral capsid (Yasui et al., 1998). Full length core sequence is required for the production of E2 glycoprotein carrying N linked glycosylation (Zhu et al., 2002). It contains 191 amino acids with three different domains: an N-terminal hydrophilic domain of 120 amino acids (domain D1), a C-terminal hydrophobic domain of about 50 amino acids (domain D2) and the last 20 amino acids (domain D3) (Santolini et al., 1994). D1 domain contains many positively charged amino acids and is involved in RNA binding and nuclear localization due to the presence of three predicted nuclear localization signals (NLS) (Chang et al., 1994). Domain D2 is responsible for core protein association with outer mitochondria membranes, endoplasmic reticulum (ER) membranes and lipid droplets (Schwer et al., 2004). Domain D3 serve as a signal peptide for the downstream envelope protein E1 (Santolini et al., 1994). Core protein also have role in apoptosis and malignant transformation of cells (Yan et al., 2008; Liu et al., 2002). In this study, we reported HCV core gene sequences from six different patients from Pakistani population. Their sequence analysis and phylogenetic studies were performed by comparing them with reported HCV core sequences, from different genotypes. HCV core gene was selected in this study because most genotype systems are from this region and it is quite easy to amplify this region.

MATERIALS AND METHODS

RNA extraction and polymerase chain reaction (PCR)

Randomly selected HCV positive patients from the genotype 3a were included in this study; their demography is shown in Table 1. 1300 ul of blood sample was taken in ethylenediaminetetraacetic acid (EDTA) vacutainer tubes and centrifuged at 12,000 g for 2 min to get the serum. Viral RNA extraction was done by using Qiagen RNA extraction kit according to the manufacturer protocols.

Specific primers were designed by the sequence comparison of NZL1 strain and HCV-K3A/650 strain of genotype 3a. The sequences of primers were, 5'- CCCGAATTCGCCATGAGCACACTTCCTAA ACCTCAAG – 3' (sense) and 5'–CCCGCGGCCGCTTAACTGGCT GCTGGATGAAT TAAGC–3' (antisense). These primers amplified a 573 bp core region from HCV positive samples.

The RNA extracted was taken as template for the complementary DNA (cDNA) synthesis. The reaction mixture for reverse transcription had a total volume of 20 ul which contained 13 ul of RNA, 1 ul dNTPs (10 mM), 20 units of molony murine leukemia virus reverse transcriptase enzyme (Fermentas), 4 ul M.Mulv buffer and 1 ul specific antisense primer. Cycle conditions for cDNA were as follows: 42°C for 55 min followed by 70°C for 10 min.

The PCR reaction mixture contained 5 ul of cDNA as template, 1 ul of each sense and antisense primer, 2 ul of dNTPs (2 mM), 2.5 ul of Dream Taq buffer, 13 ul of nuclease free water and 1.5 unit of DreamTaq Enzyme (Fermentas). The cycle conditions were as follows: 94°C for 3 min followed by 35 cycles of 94°C for 45 s, 62°C for 45 s, 72°C for 60 s and a final extension at 72°C for 7 min. Reactions was held at 4°C. Amplified PCR products were analyzed by electrophoresis on 1.2% agarose gel.

Cloning

PCR product was separated on 1% TAE gel and purified by using Qiagen gel extraction kit according to the manufacturer's protocols. PCR product was then cloned with InsTAclone PCR Cloning kit (Fermentas) according to the manufacturer's protocol. In brief, 10 ul of PCR product was mixed with 3 ul of TA vector, 10 ul of water, 6 ul of 5 x buffers and 1 ul of ligase enzyme. The ligation mix was incubated at 4°C for 16 h and then transformed into BL 10 competent cells by heat shock method. 40 ul of 5-bromo-4-chloro-3-indolyl-b-D-galactopyranoside (X-Gal) and 40 ul of isopropyl β-D-1-thiogalactopyranoside (IPTG) were spread on the agar plate containing 1% ampicillin; the transformed cells were spread on it and incubated at 37°C overnight.

Sequencing

Clones were subjected to sequencing by using Beckman coulter CEQ 8000. The sequencing reaction contained 5 ul of template DNA, 6 ul of water, 1 ul of core specific sense or antisense primer and 8 ul of dye terminator cycle sequencing (DTCS) mix. The thermo cycler conditions for sequencing reaction were 96°C for 20 s, 50°C for 20 s, 60°C for 4 min for 30 cycles followed by final hold at 4°C. 5 ul of stop solution containing 2 ul of 3 M sodium acetate, 2 ul of 100 mM disodium EDTA and 1 ul of 20 mg/ml of glycogen was added to each tube. The sequencing reaction containing stop solution was then washed with 100% followed by 70% ethanol and vacuum dried. The pellet was resuspended in 40 ul of sample loading solution, transferred to the wells of sample plate and placed in the sequencer.

Sequence comparison and phylogenetic analysis

Three clones from each patient were taken; their sequencing was

Table 2. Percentage nucleotide identity of six core isolates with isolate number Pk-ncvi/1 to Pk-ncvi/6 with each other and 3a reference strains having isolate name NZL1 and HCV-K3A/650.

HCV Isolates	Nucleotides Identity (%)					
	Pk-ncvi/1	Pk-ncvi/2	Pk-ncvi/3	Pk-ncvi/4	Pk-ncvi/5	Pk-ncvi/6
NZL1	96	95	96	95	94	94
HCV-K3A/650	92	93	91	93	90	90
Pk-ncvi/1	100	95	93	93	92	93
Pk-ncvi/2	95	100	93	97	91	91
Pk-ncvi/3	93	93	100	92	91	92
Pk-ncvi/4	93	97	92	100	91	91
Pk-ncvi/5	92	91	91	91	100	91
Pk-ncvi/6	93	91	92	91	91	100

Table 3. Percentage amino acids identity of six core isolates Pk-ncvi/1 to Pk-ncvi/6 compared with each other and 3a reference strains with isolate name NZL1 and HCV-K3A/650.

HCV Isolates	Amino acids Identity (%)					
	Pk-ncvi/1	Pk-ncvi/2	Pk-ncvi/3	Pk-ncvi/4	Pk-ncvi/5	Pk-ncvi/6
NZL1	98	98	94	97	92	92
HCV-K3A/650	94	94	91	93	88	89
Pk-ncvi/1	100	98	93	96	91	91
Pk-ncvi/2	98	100	93	97	91	91
Pk-ncvi/3	93	93	100	93	89	87
Pk-ncvi/4	96	97	93	100	90	90
Pk-ncvi/5	91	91	89	90	100	86
Pk-ncvi/6	91	91	87	90	86	100

done from both sense and antisense primer. These sequences were aligned in CLC workbench software (www.clcbio.com) to draw a consensus sequence and the consensus sequence was submitted to NCBI. Pair wise nucleotides and deduced amino acid sequence comparison was done by using Clustal W software (Thompson et al., 1994). Phylogenetic analysis of our six isolates with sixty seven core sequences from different genotypes was performed by using CLC workbench software (www.clcbio.com).

RESULTS

Cloning and sequencing

HCV core region was amplified by using specific sense and antisense primers, a single band of 573 base pairs was obtained and was cloned in TA vector (Fermentas). Positive clones were identified on the basis of blue/white colony selection. 10-15 white colonies were selected from each patient, clone confirmation was done by colony PCR and mini prep followed by restriction digestion. Three random colonies from each patient were subjected to sequencing and their sequences were aligned in CLC workbench (www.clcbio.com) to draw the consensus sequence. The sequence similarity rates between the clones generated from one patient sample was 99%. Consensus core gene sequences from six different patients with isolate number Pk-ncvi/1 to Pk-ncvi/6 were submitted to NCBI under accession numbers QG180059 to QG180064.

Sequence analysis

Pair wise nucleotides and deduced amino acid sequence comparison of our six isolates Pk-ncvi/1 to Pk-ncvi/6 were performed with each other and two reference strains of HCV 3a genotype with isolate name NZL1 (Sakamoto et al., 2004) and HCV-K3A/650 (Yamada et al., 1994) as shown in Tables 2 and 3. Highest nucleotides identity (97%) was observed between Pk-ncvi/2 and Pk-ncvi/4 while lowest nucleotides identity (90%) was observed between Pk-ncvi/5 and Pk-ncvi/6 with HCV-K3A/650 (Yamada et al., 1994). Highest amino acids identity (98%) was observed between Pk-ncvi/1 and Pk-ncvi/2 with each other and with NZL1 (Sakamoto et al., 2004), while lowest amino acids identity (86%) was observed between Pk-ncvi/5 and Pk-ncvi/6.

We found two clusters of basic residues in N-terminus of deduced amino acid sequences of core HCV dominated

by arginine, first cluster starts at position 9 and ends at position 18. In this cluster, 4 out of 10 amino acids are arginine. However in isolate Pk-ncvi/3, arginine at position 9 and 18 is substituted to glycine which is non polar in nature, and in Pk-ncvi/5 at position 17, arginine is replaced by tyrosine. In contrast to this, in isolate Pk-ncvi/4, an additional arginine is present at position 7. Second cluster of arginine is present at position 39 - 70 where eleven out of thirty two residues are arginine. In our isolates, leucine at position 139 and valine at position 140 are highly conserved; however position 144 is occupied by valine instead of leucine except in Pk-ncvi/6. Serine at position 173, phenylalanine 177 and leucine 179 is highly conserved in all the sequences while leucine 182 is substituted to phenylalanine in most of the sequences.

Phylogenetic analysis

Six different core gene sequences of HCV reported in this study along with sixty seven core sequences obtained from NCBI with different genotypes were used for the construction of phylogenetic tree as shown in Figure 1.

Hepatitis C virus genotypes are mostly grouped into single cluster in the tree. There are twenty HCV 3a sequences, grouped into Cluster I and Cluster II, out of which six are our newly reported sequences, twelve sequences are from the Centre of Excellence in Molecular Biology (CEMB) Pakistan and two sequences are from Japan. Cluster I contains our six sequences with isolate number Pk-ncvi/1 to Pk-ncvi/6 along with four other sequences, out of which PKIS-1 and PK3a-C1 are from Pakistani origin while NZL1 (Sakamoto et al., 2004) and HCV-K3A/650 (Yamada et al., 1994) are from Japan. Cluster II contains ten different isolates from Pakistani origin. Amino acid sequence comparison of Cluster I and Cluster II is shown in Figure 2.

DISCUSSION

In this study, we reported six core gene sequences of HCV isolated from patients of Pakistani population along with their sequence comparison and phylogenetic analysis. This is the first phylogenetic analysis of any HCV gene from Pakistani population. It was reported that two clusters of basic residues in N-terminus of HCV core are important for encapsidation and substitution of as few as four basic residues to alanine in either cluster of basic residues or removing the cluster, which significantly affects the assembly (Klein et al., 2005).

In Pk-ncvi/5 at position 17, arginine is replaced by tyrosine; this substitution might have some effect on encapsidation. In isolate Pk-ncvi/4, an additional arginine at position 7 enhances the overall strength of the basic residues. Second cluster of arginine is present at position 39 - 70 where eleven out of the thirty two residues are arginine. These residues might also have some important function, since these are conserved in reference as well as in our reported sequences. Cleavage by signal peptide peptidase requires a signal sequence and three hydrophobic residues: Leucine 139, Valine 140 and Leucine 144 (Okamoto et al., 2004). In our isolates, leucine at position 139 and valine at position 140 are highly conserved, however position 144 is occupied by valine instead of leucine except in Pk-ncvi/6. Thus it seems that valine is more common at this position than leucine as reported by Okamoto et al. (2004). Serine 173 (Chang et al., 1994), phenylalanine 177, leucine 179 (Okamoto et al., 2004) and leucine 182 (Hussy et al., 1996) have been reported as potential cleavage sites for signal peptide peptidase; these are conserved in our reported sequences except in leucine 182.

Pair wise nucleotide and deduced amino acid sequence comparison of our six sequences with two reported full length 3a strains from Japanese origin was performed. Nucleotide sequence comparison showed that our sequences have 94 to 96% homology with NZL1 (Sakamoto et al., 2004) strain and 90 to 93% homology with HCV-K3A/650 (Yamada et al., 1994) strain. Deduced amino acid sequence comparison showed that our sequences have 92 to 98% homology with NZL1 (Sakamoto et al., 2004) strain and 88 to 94% homology with HCV-K3A/650 (Yamada et al., 1994) strain. Nucleotide and deduced amino acid sequence comparison showed that our sequences have high homology with 3a reported sequences from Japan.

Phylogenetic tree was constructed by using six core sequences reported in this study and sixty seven core sequences from different genotypes. Phylogenetic analysis showed that 3a sequences are grouped in two clusters. Our six core sequences with four other sequences, two from Japan and two from Pakistan, are grouped in Cluster I. Cluster II contains ten different 3a sequences from Pakistani origin. Cluster I suggest that our isolates and Japanese isolates have high homology and grouped into same cluster.

Amino acid sequence comparison of Cluster I and Cluster II was performed. At position 72, glutamic acid was present in all the members of Cluster I, while in Cluster II it was replaced by aspartic acid. Position 182 was occupied by phenylalanine in most of the isolates from Cluster I while in Cluster II, it was replaced by leucine. Amino acid sequence comparison of Cluster I showed that our six isolates and two isolates form Japanese origin have some common ancestral origin. Cluster II contains a serine rich region and three amino acids addition at the C-terminus of core sequence. These substitution at C-terminus do not have any effect on encapsidation (Lorenzo et al., 2001; Kunkel et al., 2001) but may have effect on signal peptide for the downstream envelope protein E1 (Chang et al., 1994). Further study is required to check the effect of various amino acid substitutions on core encapsidation, RNA binding, nuclear

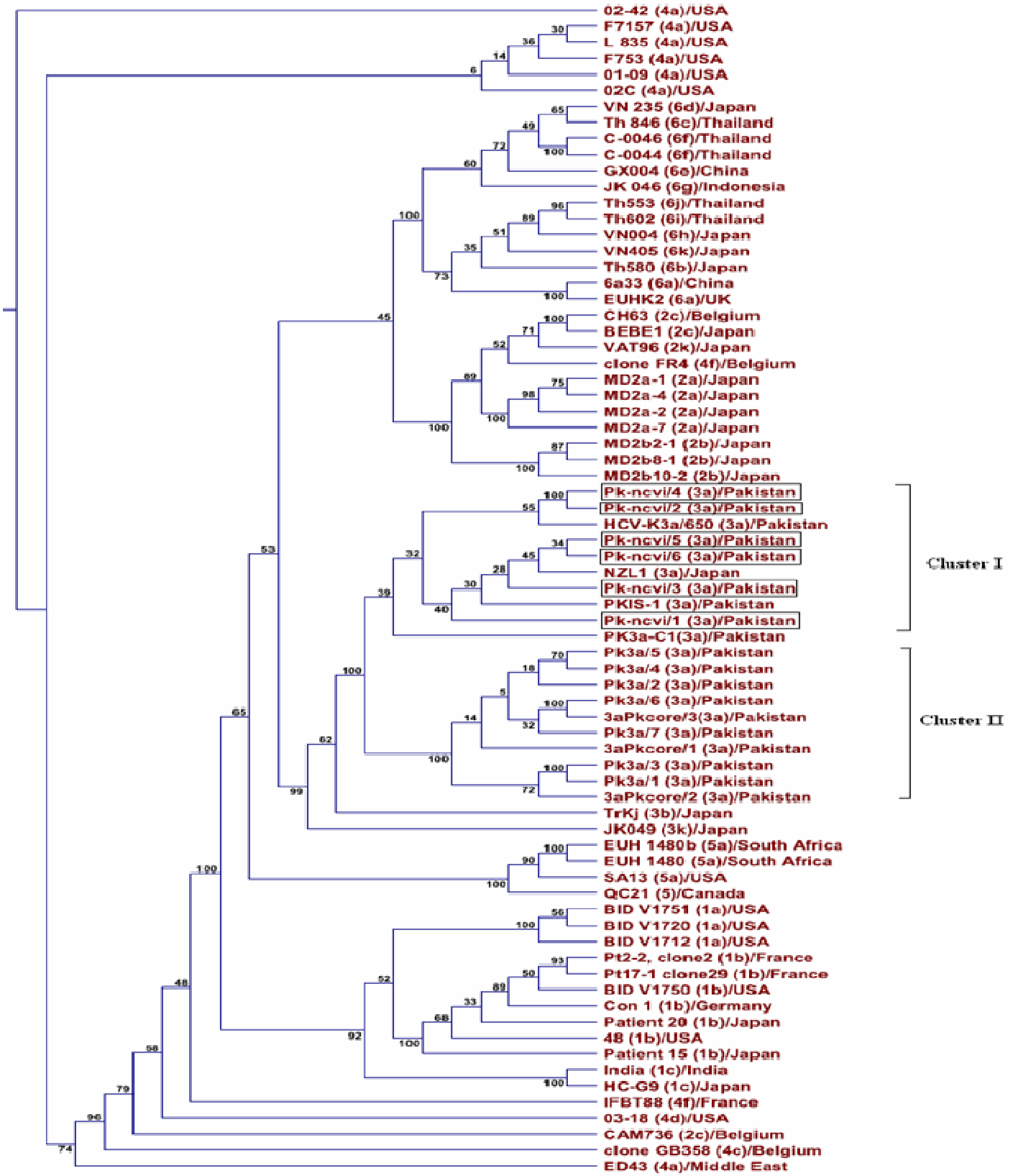

Figure 1. Phylogenetic tree of core gene sequences of HCV. Tree was generated by Neighbor joining algorithm. Boot strap values are shown on the branches. Tree shows the phylogenetic relationship of six newly reported sequences, marked in boxes, with 67 other core gene sequences. The isolate (genotype)/country of the sequences are shown in figure. Tree was constructed by using CLC workbench software. Genotypes 3a is grouped in cluster I and cluster II.

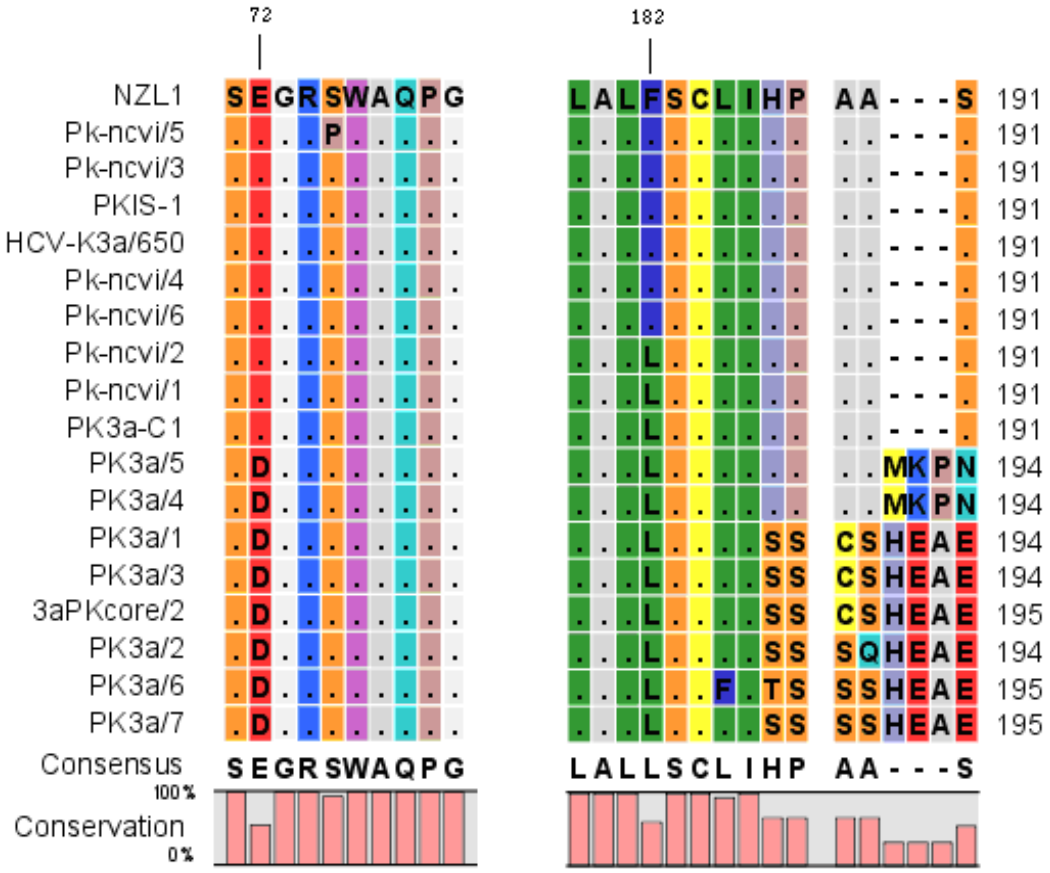

Figure 2. Amino acid sequence alignment of Cluster I and Cluster II of phylogenetic tree. First ten sequences are from Cluster I, remaining eight sequences are from Cluster II, showing amino acid substitution in two Clusters. Position 1 is the first amino acid of deduced amino acid sequence of HCV core.

localization, association with mitochondrial and endoplasmic reticulum membranes.

ACKNOWLEDGEMENTS

This work was supported by Pak-US Science and Technology cooperative program entitled “HCV management in Pakistan” and Higher Education Commission of Pakistan indigenous scholarship program.

REFERENCES

Chang SC, Yen JH, Kang HY, Jang MH, Chang MF (1994). Nuclear localization signals in the core protein of hepatitis C virus. Biophys. Res. Commun. 205: 1284-1290.

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

Gao QJ, Liu DW, Zhang SY, Jia M, Wang LM, Wu LH, Wang SY, Tong LX (2009). Polymorphism of some cytokines and chronic hepatitis B and C virus infection. World J. Gastroenterol. 15(44): 5610-5619

Hussy P, Langen H, Mous J, Jacobsen H (1996). Hepatitis c virus core protein: carboxy terminal boundries of two processed species suggest cleavage by a signal peptide peptidase. Virology, 224: 93-104.

Klein KC, Dellos S, Lingappa JR (2005). Identification of residues in the hepatitis C virus core protein that are critical for capsid assembly in a cell free system. J. Virol. 79: 6814-6826.

Kunkel M, Lorinczi M, Rijnbrand R, Lemon SM, Watowich SJ (2001). Self assembly of nucleocapsid like particles from recombinant hepatitis c virus core protein. J. Virol. 75: 2119-2129.

Liu XF, Zou SQ, Qiu FZ (2002). Construction of HCV-core gene vector and its expression in cholangiocarcinoma. World J. Gastroenterol. 8(1): 135-138

Lorenzo LJ, Carrera SD, Falcon V, Acosta-Rivero N, Gonzalez E, Rosa MC, Menendez I, Morales J (2001). Assembly of truncated HCV core antigen into virus like particles in *E coli*. Biochem. Biophys. Res. Commun., 281: 962-965.

Major ME, Feinstone SM (1997). The Mol. Virol. Hepat. C. Hepatol. 25: 1527-1538.

Okamoto K, Moriishi K, Miyamura T, Matsuura Y (2004). Intermembrane proteolysis and endoplasmic reticulum retention of hepatitis c virus core protein. J. Virol. 78: 6370-6380.

Rice CM (1996). Flaviviridae: the viruses and their replication, In Fields BN, Knipe DM, Howley PM (ed.), Fields virology. Lippincott-Raven Publishers, Philadelphia, Pa. pp. 931-960.

Sakamoto M, Akahane Y, Tsuda F,Tanaka T, Woodfield DG, Okamoto H (1994). Entire nucleotide sequence and characterization of a hepatitis C virus genotype V/3a. J. Gen. Virol. 75: 1761-1768.

Santolini E, Migliaccio G, Monica NL (1994) Biosynthesis and biochemical properties of the hepatitis C virus core protein. J. Virol.

68: 3631-3641.
Schwer B, Ren S, Pietschmann T, Kartenbeck J, Kaehlcke K, Bartenschlager R, Yen TS, Ott M (2004). Targeting of hepatitis C virus core protein to mitochondria through a novel C-terminal localization motif. J. Virol. 78: 7958-7968.
Thompson JD, Higgins DG, Gibson TJ (1994). Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, postion-spiecific gap penalties and weight matrix choice. Nucleic Acids Res. 22: 4673-4680.
Turhan V, Ardic N, Eyigun CP et al. (2005). Investigation of genotype distribution of hepatitis C virus among Turkish population in Turkey and various European countries. Chin. Med. J. 118: 1392-1394.
Waheed Y, Shafi T, Safi SZ, Qadri I (2009). Hepatitis C virus in Pakistan: A systematic review of prevalence, genotypes and risk factors. World J. Gasteroenterol. 15: 5647-5653.
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 with different genotypes. J. Gen. Virol. 75: 3279-3284.
Yan XB, Mei L, Feng X, Wan MR, Chen Z, Pavio N, Brechot C (2008). Hepatitis C virus core proteins derived from different quasispecies of genotype 1b inhibit the growth of Chang liver cells. World J. Gastroenterol. 14: 2877-2881
Yasui K, Wakita T, Kohara KT Funahasi SI, Ichikawa M. Kajita T, Moradpour D, Wands JR, Kohara M (1998). The native form and maturation process of hepatitis C virus core protein. J. Virol. 72: 6048-6055.
Zhu LX, Liu J, Li YC, Kong YY, Staib C, Sutter G, Wang Y, Li GD (2002). Full-length core sequence dependent complex-type glycosylation of hepatitis C virus glycoprotein. World J. Gastroenterol. 8: 499-504.