Romanian Biotechnological Letters *Vol. 18, No.1, 2013*

ORIGINAL PAPER

Global diversity of HCV: *In-silico* analysis based on core region

Received for publication, April 15, 2012
Accepted, November 15, 2012

SOBIA KANWAL[1], ANWARULLAH[2] AND TARIQ MAHMOOD[3]*
[1]*Department of Animal Sciences, Faculty of Biological Sciences, Quaid-i-Azam University, Islamabad-45320, Pakistan*
[2]*Department of Biochemistry, Faculty of Biological Sciences, Quaid-i-Azam University, Islamabad-45320, Pakistan*
[3]*Department of Plant Sciences, Faculty of Biological Sciences, Quaid-i-Azam University, Islamabad-45320, Pakistan*
**Corresponding Author's email: tmahmood@qau.edu.pk; tmahmood.qau@gmail.com*
Telephone: +92 51 9064 3144, Fax: +92 51 9064 3004

Abstract

Hepatitis C virus is a devastating virus, known to mankind since 1989, but circulating in the human blood for centuries. Viral prevalence and incidences rates vary from country to country. This study is basically a computational analysis based on core gene of HCV conducted to understand the pattern of evolution of HCV around the globe. The nucleotide and protein sequences were retrieved from HCV database and were analyzed by using ClustalW for alignment of sequences and MEGA 4 was used for construction of phylogenetic tree to see the evolutionary pattern of HCV in five different continents of world. In-silico analysis of HCV *core* gene shows that it has undergone various diversifications in *core* region worldwide showing different evolutionary rates measured by the Neighbour Joining method. The number of chains varies in the core structure without affecting the overall phenotypic expression.

Keywords: Core region, HCV, phylogenetic analysis, *in silico* analysis

Introduction

The investigation of origin and spread of hepatitis C virus (HCV) in human population is important in future perspectives for managing its spread worldwide. HCV is a virus, distressing more than 170 million people worldwide (HOUGHTON [1]). It is an enveloped, single-stranded RNA molecule and the only known member of the *hepacivirus* genus in the family *Flaviviridae*, with genome size of approximately 9.6 kb that contains single Open Reading Frame (ORF) encoding a polyprotein of about 3000 amino acids. The structural and non structural proteins instruct by ORF include the core, capsid, E1, E2, RNA polymerase, NS4B and NS5A (CHOO & al. [2]). Biologically, HCV core protein has been implicated in cellular proliferation as it regulates virus-induced transformation and pathogenesis (JIN& al. [3]) and has RNA-binding activity (SANTOLINI & al., HWANG & al. [4, 5]). Further it has the ability to form homo-multimers as well (MATSUMOTO & al., NOLANDT & al. [6, 7]).

Due to high genetic variability HCV isolates are grouped into six genotypes based on their genetic diversity (ROBERTSON & al. [8]) and all six isolates have <72% homology with respect to main branch in phylogenetic tree. Among these genotypes some strains are present in particular regions while some are found worldwide (VERBEECK & al. [9]).

The phylogenetic analysis has been carried out widely to categorize the origin, evolutionary pattern of HCV (BUKH & al., CHAMBERLIN & al., SMITH & al. [10, 11, 12]) global epidemiology and epidemic history (NDJOMOU & al. [13]). The computational analysis holds up the hypothesis that immune selection was a significant driving force in divergence of HCV genotypes (SMITH & al. [12]). The analysis of viral ancestry sampled from different regions worldwide is helpful in marking out the migration of the virus. For

viral infections of different kinds, the spreading of the parasite and its host cannot be simply traced; therefore phylogenies may possibly be a mode to check migratory pathways of the virus (WALLACE & al., HOLMES [14, 15]).

The computational phylogenetic analysis of nucleotide and protein sequences was performed to check the evolutionary pattren among HCV strains in different regions and the pattern of evolution has shown that the HCV strains are might be under strong positive selection. Moreover, the structural analysis was done to check the conservation of amino acids chains in core region of HCV, contributing in shaping the appropriate drug therapy in future studies.

Materials and method

HCV resource at Los Alamos provides access to multiple databases that contain annotated sequences. The HCV sequence database have collection and annotation of sequence data and make available them to the public that contains easily accessible search interface and a large number of sequence analysis tools (KUIKEN & al. [16]). HCV sequence database was searched for HCV "core region" reported in Asia, Africa, North America, South America and Europe.

The FASTA formatted files for nucleotide and protein sequences were prepared for multiple sequence alignment using CLUSTALW algorithm and MEGA 4 (KUMAR& al. [17]) was used to obtain phylogenetic tree. Out of the total entries retrieved for each region, 10 sequences were randomly picked for comparative analysis of nucleotide and protein sequences (Table.1).

For structural analysis of core region PDB (Protein Data Bank) was searched for "HCV core". Out of 15 structures generated by search first 5 structures (3KQH, 3KQK, 3KQL, 3KQN and 3KQU) were selected for analysis.

Table.1. Status of HCV core gene reported in Los Alamos HCV sequence database up till October 2010.

Gene Name	Region	Entries
HCV core	Asia	714
HCV core	Africa	34
HCV core	South America	10
HCV core	North America	1165
HCV core	South America	1031
Total		2954

Nucleotide sequence analysis:

Forty six nucleotide sequences were selected and aligned using CLUSTALW (parameters reading DNA Pairwise Parameters; Gap opening penalty: 15, Gap extension penalty: 6.66, Multiple Parameters; Gap opening penalty: 15, Gap extension penalty: 6.66, Transition weight: 0.5, Delay divergent cutoff: 30%, DNA weight matrix: ClustalW (1.6), negative matrix OFF). To obtain phylogenetic tree alignment was analyzed in MEGA 4 tool (parameters reading; Tree inference Method: neighbor joining, Include sites Gaps/ missing data: complete deletion, Codon positions: 1^{st} + 2^{nd} + 3^{rd} + noncoding, substitution model: Maximum nucleotide likelihood, Substitutions to include: transitions + transversions).

Protein sequence analysis:

Forty six protein sequences were selected and aligned using CLUSTALW (parameters reading DNA Pairwise Parameters; Gap opening penalty: 10, Gap extension penalty: 0.1, Multiple Parameters; Gap opening penalty: 10, Gap extension penalty: 0.2, protein weight matrix: BLOSUM, Residue specific penalties: ON, Hydrophilic penalties: ON, Gap separation distance: 4, End gap separation: OFF, negative matrix OFF, Delay divergent

cutoff: 30 %,). To obtain phylogenetic tree alignment was analyzed in MEGA 4 tool (parameters reading; Tree inference Method: neighbor joining, Include sites: Gaps/ missing data: complete deletion, substitution model: Amino p-distance, Substitutions to include: All).

The viral protein structures of core were of "Hydrolase" type as reported in PDB indicating that the number of amino acid chains residues are conserved but chains number vary among all structures except for chain A which is conserved in all structures, 3KQH, 3KQK, 3KQL, 3KQN and 3KQU (Table. 2).

Table.2. Structural analysis illustrates the number of amino acid chains reported in Protein Data Bank.

3KQH		**3KQK**		**3KQL**		**3KQN**		**3KQU**	
Chain A	436	Chain A	436	Chain A	437	Chain A	437	Chain A	437
Chain B	436	Chain B	436	Chain B	437	Chain E	6	Chain B	437
Chain C	6	Chain C	6	Chain E	6			Chain C	437
Chain D	6	Chain D	6	Chain F	6			ChainD	437
								ChainE	437
								Chain F	437
								Chain M	19
								Chain N	19

Results

Multiple sequence alignment of HCV core protein and nucleotide sequences was done to check the evolutionary pattern of this region among five different geographical regions. The analysis has revealed that an uncharacteristic pattern that did not follow a specific model of resemblance, is present among the HCV core region strains at both nucleotide and protein level.

Nucleotide sequence analysis:

The phylogenetic analysis of HCV core nucleotide sequence reveals irregular pattern of evolution. More diversity has been observed in the actively mutating viral strains that had evolved earlier and are circulating in the human population for longer time periods during the course of evolution. A unique South American strain in clade1 has indicated an ancient evolutionary history similarly a single Asian strain in same clade has shown static mutation behavior (Figure.1). The tree topology of all different geographical regions have shown active rate of mutation in core region of HCV at nucleotide level like shown by European strains (Clade1, Cluster I), North American and South American strains (Clade1, Cluster III). Some migration pattern has been observed among African, Asian, European and South American strains (Clade1, Cluster II). It was observed that some African strains have been originated in Africa (forming out group) and undergone diversification before distribution to other regions in clade2 (Figure.1). Further, sequences from Asia (Clade1, Cluster I, II & IV and Clade2, Cluster I) appeared at distinct positions in a tree showing high level of diversity. The nucleotide substitution rate per site per year in the core region of HCV has found to be 0.02.

Protein sequence analysis:

The phylogenetic analysis based on Multiple Sequence Alignment of protein sequences reported from five different geographical regions depicts an anomalous behavior because of different protein expression pattern. It showed 0.01 amino acid substitution rate per site per year (Figure.2). High order branching has been observed in European (Clade1, Cluster I), some African (Clade1, Cluster V), North American and South American strains (clade1, cluster IV) illustrate active mutation rate in these regions. However some strains have shown highly divergent behavior by forming clusters with other geographical regions in which Asian

strains (Clade1, Cluster II and Clade2, Cluster I) are showing high diversity. A single strain of Africa and Asia has shown static mutation in clade1 (Figure.2). The clade1, cluster VI has shown active mutation in African and South American strains and demonstrating that migration might take place among these regions. Unique African strain is forming out group similarly as in phylogenetic tree of nucleotide sequences (Clade2, Cluster I) (Figure.2).

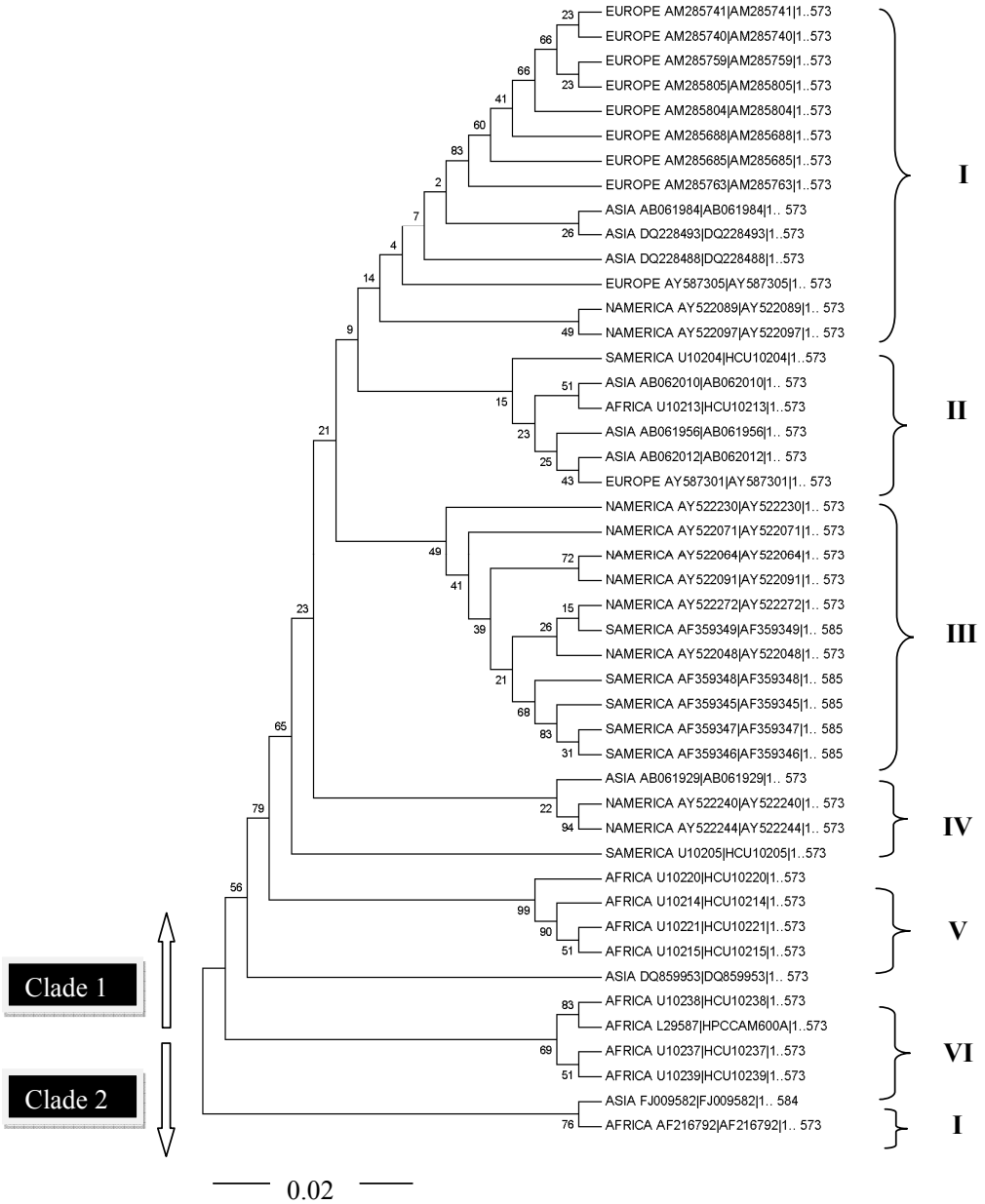

Figure 1. Phylogenetic tree obtained from MEGA 4 showing 46 nucleotide sequences of HCV core.

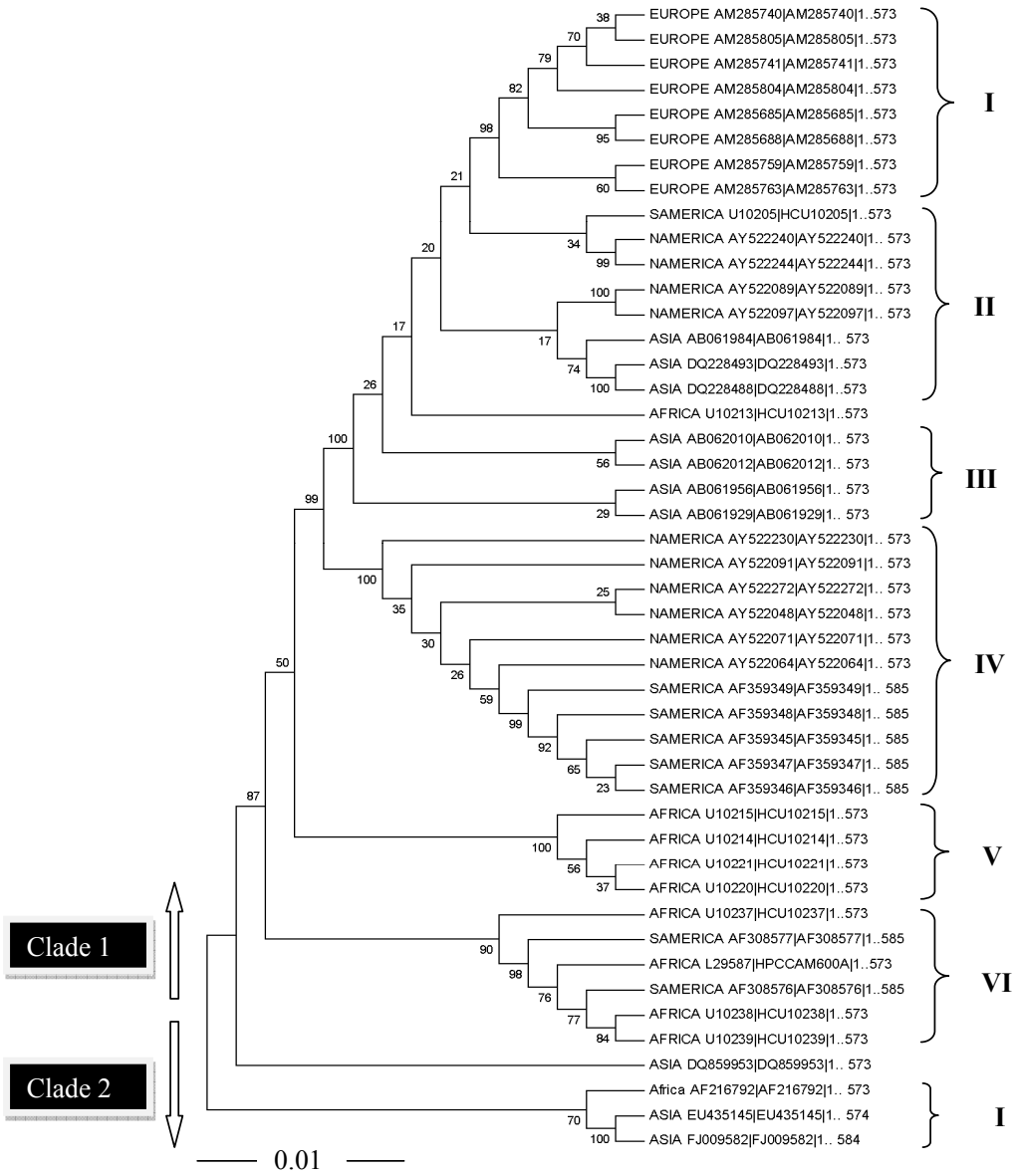

Figure 2. Phylogenetic tree obtained from MEGA 4 showing 46 protein sequences of HCV core.

Discussion

HCV evolution is a highly active and it exploits all known ways of genetic variation including recombination and mutation, to ensure its survival (BOSTAN & al. [18]). It has been reported earlier that HCV genome evolves rapidly at different rates similar to other RNA viruses (OGATA & al. [19]). The high mutation rate of the HCV might be one of the points of determinant action of the natural selection and thereby cooperated in inducing the divergence of viral species.

Asian strains showed the similarity with European strains, it might be due to migration events as a study describes that there is a sizeable community of South Asians like Asian labor migrants settled in European countries Currently, approximately 2 million South Asians are living in Africa; some came in late nineteen and early twentieth century (OONK [20]. Similarly, a study showed that South American countries are populated with European, Asian and Africans (COLLIER & al. [21]).

The main objective of the current study was to computationally analyze HCV core region and to check the evolutionary pattern in different regions of the world. The phylogenetic analysis depicts that viral genome undergone various significant changes with time at different rates in which core region is considered to be more diverse (BOSTAN AND MAHMOOD [22]). It was observed that Asian strains have shown high mobility of HCV strains among different regions especially to Europe and Africa. A limited migration pattern has been identified among strains of Europe, North America, South America and Africa that have shown high diversity in their respective geographical regions as reported previously that in areas of endemicity a highly divergent pattern is observed among the strains suggesting long infection duration (OLIVER & al. [23]). The cluster II in clade 1 of nucleotide sequences (Figure.1) have shown close similarity with European, Asian and African strains might be due to migration of Europeans, Asians and Africans to a same region where HCV strains followed same pattern of divergence as discussed that migration has been made among Europeans, Africans and Asians to South America (COLLIER & al. [21]). Moreover, analysis has shown that strain is originated from Africa where it had gone through mutation as reported that common ancestor belongs to Africa where it undergone diversification and spread to other regions of the world (NDJOMOU & al. [13]). It has been seen that core region is undergoing diversification at high rate in European, South and North American countries. Some strains showing no branching giving an idea about no change that have occurred in them for years might be as they are under high negative selection pressure due to some environmental factors. It was depicted from the data that there is a high substitution rate per site per year seen at nucleotide level as compared to amino acid level showing that protein has low rate of mutation that can change the structure of protein up to greater extent. It is already known that base substitution rate of RNA viruses ranges from 10^{-1} to 10^{-4} per genome per site per year so virus proteins are subject to change in a short evolutionary time (OGATA & al. [20]) and nucleotide substitution plays a critical role in HCV sequence divergence.

Moreover, the structural analysis illustrate that chain A has critical importance in pathogencity of viral core region and whether chains number remain same or not it does not affect the overall function of core protein but it had high sensitivity and specificity for prediction of sustained virological response. However, it has been reported earlier that amino acid substitution pattern of the core region and genetic variation can affect the treatment efficiency (NORIO & al. [24]).

In-silico analysis of HCV core gene has shown that it has undergone various diversifications in different regions indicating inconsistent pattern of evolution both at nucleotide and protein level.

Conclusion

It can be concluded from the present study that there are differences in the evolutionary pattern of HCV core region in different epidemics worldwide showing different evolutionary behaviour during different phases of an individual epidemic, all depending on the specific growth rate. As far as its structural analysis is concerned the number of chains varies that doesn't affect the overall phenotypic expression of core gene. Further, the structural characterization of HCV genetic components should be known for effective drug therapy.

References

1. M. HOUGHTON, The long and winding road leading to the identification of the hepatitis C virus. *J Hepatol*., 51, 939, 948 (2009).
2. Q. L. CHOO, K.H. Richman, J. Han, K. Berger, C. Lee, C. Dong, C. Gallegos, D. Coit, R. Medina-Selby, P. J. Barr, Genetic organization and diversity of the hepatitis C virus. *Proc Natl Acad Sci USA*., 88, 2451, 2455 (1991).
3. JIN DY, WANG H L, ZHOU Y, CHUN ACS, KIBLER KV, HOU YD, KUNG H FU, JEANG KT, Hepatitis C virus core protein-induced loss of LZIP function correlates with cellular transformation. *EMBO J*., 19, 729 – 740 (2000)
4. E. SANTOLINI, G. MIGLIACCIO, N. LA MONICA, Biosynthesis and biochemical properties of the hepatitis C virus core protein. *J Virol*., 68, 3631, 3641(1994).
5. S. B.HWANG, S. Y. LO, J. H. OU, M. M. C. LAI, Detection of cellular proteins and viral core protein interacting with the 5' untranslated region of hepatitis C virus RNA. *J Biomed Sci*., 2, 227,236 (1995).
6. M. MATSUMOTO, S. B. HWANG, K. S. JENG, N. ZHU, M. M. C. LAI, Homotypic interaction and multimerization of hepatitis C virus core protein. *J Virol.,* 218, 43,51 (1996).
7. O. NOLANDT, V. KEM, H. MULLER, E. PFAFF, L. THEILMANN, R. WELKER, H. G. KRAUSSLICH, Analysis of hepatitis C virus core protein interaction domains. *J Gen Viro.,l* 78, 1331, 1340 (1997).
8. B. ROBERTSON, G. MYERS, C. HOWARD, T. BRETTIN, J. BUKH, B. GASCHEN, T. GOJOBORI, G. MAERTENS, M. MIZOKAMI, O. NAINAN, S. NETESOV, K. NISHIOKA, I. SHIN, P. SIMMONDS, D. B. SMITH, L. STUYVER, A. J. WEINER, Classification, nomenclature, and database development for hepatitis C virus (HCV) and related viruses: proposals for standardization. *Arch Virol*., 143, 2493., 2503 (1998) .
9. J. VERBEECK, P. MAES, P. LEMEY, O. G. PYBUS, E. WOLLANTS, E. SONG, F. NEVENS, J. FEVERY, W. DELPORT, S. VAN DER MERWE, M. VAN RANSTM, Investigating the origin and spread of hepatitis C virus genotype 5a. *J Virol*., 80, 4220, 4226 (2006).
10. J. BUKH, R. H. MILLER, M. C. KEW, R. H. PURCELL, Hepatitis C virus RNA in southern African blacks with hepatocellular carcinoma. *Proc Natl Acad Sci., USA*., 90, 1848,1851 (1993).
11. R. W. CHAMBERLAIN, N. ADAMS, A. A. SAEED, P. SIMMONDS, R. M. ELLIOTT, Complete nucleotide sequence of a type 4 hepatitis C virus variant, the predominant genotype in the Middle East. *J Gen Virol*., 78, 1341, 1347 (1997).
12. D. B. SMITH, S. PATHIRANA, F. DAVIDSON, E. LAWLOR, J. POWER, P. L. YAP, P. SIMMONDS, The origin of hepatitis C virus genotypes. *J Gen Virol*., 78, 321, 328 (1997).
13. J NDJOMOU, OG PYBUS, B MATZ, Phylogenetic analysis of hepatitis C virus isolates indicates a unique pattern of endemic infection in Cameroon. *J Gen Virol.,* 84: 2333– 2341 (2003)
14. T. A. WALLACE, R. L. PRUEITT, M. YI, T. M. HOWE, J. W. GILLESPIE, H. G. YFANTIS, R. M. STEPHENS, N. E. CAPORASO, C. A. LOFFREDO, S. AMBS, Tumor immunobiological differences in prostate cancer between African–American and European–American men. *Cancer Res*., 68, 927, 936 (2008).
15. E. C. Holmes, Evolutionary history and phylogeography of human viruses. *Annu. Rev Microbiol*., 62, 307,328 (2008).
16. C KUIKEN, K YUSIM, , L BOYKIN, R RICHARDSON, The Los Alamos hepatitis C sequence database. *Nucleic Acids Res.,* 36: 512–516 (2004).
17. S. KUMAR, M. NEI, J. DUDLEY, K. TAMURA, For evolutionary analysis of DNA and protein sequences. *Brief Bioinform*., 9, 299, 306 (2008).
18. N. BOSTAN, M. M. MUSTAFA, W. SAFDAR, Q. JAVED, T. MAHMOOD, Phylogenetics of HCV: Recent advances. *Afr J Biotechnol*., 9, 5792, 5799 (2010).
19. N. OGATA, H. J. ALTERT, R. H. MILLER, R. H. PURCELL, Nucleotide sequence and mutation rate of the H strain of hepatitis C virus. *Proc Natl Acad Sci USA*., 88, 3392, 3396 (1991).
20. OONK G, Global Indian Diasporas: exploring trajectories of migration and theory, 1st ed.University press, Armsterdam (2007)
21. COLLIER, SIMON, THOMAS E. SKIDMORE, HAROLD BLAKEMORE, The Cambridge Encyclopedia of Latin America and the Caribbean (2nd ed). Cambridge, Cambridge University Press, (1994).
22. N. BOSTAN, T. MAHMOOD, An Overview about Hepatitis C: A Devastating Virus. *Crit Rev Microbiol*., 36, 91,133 (2010).
23. G. P OLIVER , B ELEANOR, T RACHEL, L PHILIPPE, VM PETER, R BOUACHAN, S BOUNKONG, P RATTANAPHONE, S ISABELLE, S H ISLA. L LING, N. N PAUL & K PAUL, Genetic History of Hepatitis C Virus in East Asia. *J. Virol*.,83, 2,1071-1082 (2009).
24. N AKUTA, F SUZUKI, M HIRAKAWA, Y KAWAMURA, H YATSUJI, H SEZAKI, Y SUZUKI, T HOSAKA, M KOBAYASHI, M KOBAYASHI, , S SAITOH, Y ARASE Amino Acid Substitution in HCV Core Region and Genetic Variation near IL28B Gene Predict Viral Response to Telaprevir with Peginterferon and Ribavirin. *Hepatology*., 52: 421–429 (2010).