Geographical and chronological origin and evolution of Hepatitis C Virus.

Amjesh R[1], Achuthsankar S Nair[2], V.S.Sugunan[3],

[1,2]Centre for Bioinformatics University of Kerala, Karyavattom, Trivandrum, Kerala, India, 69581.

[2]Department of Zoology, His Highness The Maharaja's University College, Trivandrum, Kerala, India, 695034.

[1]amjesh@gmail.com [2]sankar.achuth@gmail.com [3]sugunanvs@yahoo.com

Hepatitis C Virus (HCV) infection is a major health problem that leads to cirrhosis and hepatocellular carcinoma. World over, more than 270-300 million people are estimated to be infected with the virus. HCV is a positive sense single stranded RNA virus and replicates within the cytoplasm of the hepatocyte using its own RNA dependent RNA polymerase (RdRp)[1]. RdRp does not have proof reading capacity[2], and hence generates mutants of the virus, resulting in a chronic infection, which ultimately ends in hepatocellular carcinoma[3]. Such mutations have given rise to several genotypes, subtypes, strains and variants with significant difference in disease outcomes. The mutation rate varies among genotypes, subtypes, strains or even in different sites of the genome[4]. Yet, the extent of heterogeneity is usually moderate, so that estimates of the time of divergence can be computed. The evolution of variants seems to be influenced by the genetic make-up and the immune response of the host and has geographical significance. Here we used phylogenetic analysis and Computational molecular dating techniques to conclude that the ancestral genotype is 7a[5] and that it originated in Canada 363 years ago. Molecular dating was based on the fact that the rate of mutation across all evolutionary lineages is constant over time[6]. Surprisingly, our analyses show that genotype 1d isolated from Canada[5] is the most recent with an evolutionary date of just 33 years. It is evident that HCV is still an emerging virus and demographical parameters seem to have a very strong influence in its evolution. We believe that this emphasises the need for developing drugs that are customised to act against strains that evolve and become geographically endemic.

Materials and methods.

Partial coding sequence of NS5B genes were retrieved from the HCV database.(77) Total number of 65 sequenceses was selected. One sequence from each Genotype and Subtype. The sampling date, sampling country and gene identification numbers are also noted. To calculate the evolutionary distance multiple sequence alignment were done with this 65 sequences using ClustalW (7) programme implemented in MEGA (8). The evolutionary distance between the genes were estimated using Kimura 2 parameter substitution model . Based on the distance model

neighbour-joining tree was constructed by tree building programme in MEGA. The distance from the most recent common ancestor to each taxa were estimated. The phylogenetic tree were constructed using the dnapars program embeded in the PHYLIP program (9). To confirm the reliability of the phylogenetic tree 1000 bootstrapping resampling test were performed using seqboot programme.It produces a collection of trees rather than a point estimate of an optimal tree, since such a tree with no measure of reliability may not be particularly helpful. When faced with such a collection, it is common practice to construct a consensus tree. The consensus tree was produced from out tree file of Dnapars using Consensus program. The tree was drawn by the program Drawgram. The hypothetical ancestral sequences of the each node of the phylogenetic tree were estimated by Dnaml program implemented in Phylip. The tree files produced by Phylip programs are viewed by Fig Tree v1.2.3.(10) Then the distances from the ancestral sequence to each strain were estimated by the Kimura 2 nucleotide substitution model of MEGA. The mean distance is then estimated from distance values obtained from MEGA Neighbour Joining tree and from ancestral sequence produced by Dnaml. The molecular date is estimated by a simple division of genetic distance by calibration rate (nucleotide substitution per site per year). In the previous study nucleotide substitution rate of HCV was estimated as $0.67*10^{-3}$ per site per year (11).

Sampling Country	Sampling Date	Accession No.	Gene Index No.	Genotype
Benin	2001	AF037244	gi|3170059	2d
Cameroon	1995	L38361	gi|1066643	1e
Cameroon	1998	AY257087	gi|30525610	1h
Cameroon	1998	AY257091	gi|30525618	1l
Cameroon	2003	AY265435	gi|30385487	4e
Cameroon	1995	L29596	gi|476675	4f
Cameroon	2004	AY743211	gi|54632752	4k
Cameroon	1998	AY265429	gi|30385475	4p
Cameroon	1998	AY265430	gi|30385477	4t
Canada	2007	EF115984	gi|134038120	1c
Canada	2007	EF115989	gi|134038130	1d
Canada	2007	AY434129	gi|38147572	1j
Canada	2007	AY434113	gi|38147545	1k
Canada	2007	EF116024	gi|134038200	2e
Canada	2007	AY754634	gi|54610706	2m
Canada	2007	EF116059	gi|134038270	2r

Canada	2000	AF279121	gi|9230780	3b
Canada	2007	EF116087	gi|134038326	3g
Canada	2000	AF279120	gi|9230778	3h
Canada	2007	AY434138	gi|38147587	3i
Canada	2007	EF116138	gi|134038428	4b
Canada	2007	EF116139	gi|134038430	4l
Canada	2007	AY434126	gi|38147567	4q
Canada	2007	EF116196	gi|134038544	6e
Canada	2007	EF116156	gi|134038464	6h
Canada	2007	EF116159	gi|134038470	6l
Canada	2007	AY894524	gi|60477635	6o
Canada	2007	EF116153	gi|134038458	6r
Canada	2007	EF116169	gi|134038490	6s
Canada	2007	AY434115	gi|38147548	7a
China	2002	AY834974	gi|56123633	2f
China	2002	AY834938	gi|56123561	6k
China	2002	AY834939	gi|56123563	6n
Egypt	2002	EF694452	gi|158146862	1g
Egypt	1999	AB103457	gi|40714114	4a
Egypt	2002	EF694517	gi|158146992	4m
Egypt	2002	|EF694422	gi|158146805	4o
France	1999	AF515988	gi|29365804	1b
France	1996	L48495	gi|1237395	1i
France	1999	AF515981	gi|29365790	2c
France	1997	AF515968	gi|29365764	2i
France	2006	DQ220919	gi|82704304	2j
France	2005	AJ291258	gi|11322297	4d
France	2005	AJ291249	gi|11322279	4h
France	2004	AY743101	gi|54632532	4n
Gabon	1995	L29614	gi|476686	4c
Gabon	1995	L29618	gi|476688	4g
Guinea	2001	AF037235	gi|3170041	1m
Japan	2008	D10648	gi|221674	2a
Laos	2004	AY735101	gi|52547281	6q

Martinique	2004	AY257465	gi|30720399	2l
Myanmar	2007	AB103135	gi|47826476	6m
Pakistan	2009	AB444475	gi|225380383	3k
South Africa	2001	DQ164544	gi|76576168	5a
Taiwan	1993	DQ666241	gi|110430931	2b
Taiwan	2005	DQ663603	gi|111082412	3a
Thailand	1999	AB027610	gi|6136892	6c
Thailand	2006	DQ640386	gi|109676985	6f
Thailand	2006	DQ640367	gi|109676947	6i
Thailand	1999	AB027608	gi|6136888	6j
Uganda	2006	AY577585	gi|48995479	4r
US	1984	AF268586	gi|13344980	1a
Uzbekistan	2002	AB081066	gi|22122154	2k
Vietnam	2006	DQ155517	gi|73765290	6d
Vietnam	2006	DQ155504	gi|73765264	6p

Table 1. Details of sequence used for the study.

Result and Discussion

Nucleotide sequences of NS5B genes of HCV are obtained from the HCV sequence database. Sampling country, sampling year, accession number, gene index number and genotype of the sequences were tabulated (Table 1). The evolutionary distance of these sequences were calculated using Kimura two parameter's model in the MEGA 4 program. A Neighbor-joining tree was constructed using this distance data (Figure 1) and evolutionary distance from the strain to their most recent common ancestor was recorded from this tree. A phylogenetically reliable tree is constructed using the Dnapars program in the PHYLIP (Figure 2). Hypothetical ancestral sequences of the each node are produced using Dnaml program of PHYLIP(Figure 3). Evolutionary distance calculation of these strains to their most recent common ancestor sequence produced by Dnaml was conducted by Kimura two parameter's model implemented in the MEGA 4 and this distance is also tabulated along with the distance from N-J tree. The mean distance is evaluated and the evolutionary date is calculated using the nucleotide substitution rate of HCV (Table 2). It is found that the genotype 7a (Accession no. AY434115) originated approximately 363 years before. Genotype 1d (Accession no. EF115989) is the newly emerged one and their evolutionary date calibrated as 33 years.

Genotype	**Accession**	**NJ**	**MEGA**	**Mean**	**Divergence**

	No.	Distance	Distance	value	Time
2d	AF037244	0.057	0.036	0.047	70
1e	L38361	0.068	0.086	0.077	115
1h	AY257087	0.093	0.096	0.095	142
1l	AY257091	0.056	0.026	0.041	61
4e	AY265435	0.031	0.026	0.028	42
4f	L29596	0.071	0.042	0.056	84
4k	AY743211	0.044	0.021	0.032	47
4p	AY265429	0.027	0.026	0.026	39
4t	AY265430	0.064	0.053	0.058	87
1c	EF115984	0.061	0.031	0.046	68
1d	EF115989	0.030	0.015	0.023	33
1j	AY434129	0.046	0.020	0.033	49
1k	AY434113	0.054	0.026	0.040	60
2e	EF116024	0.109	0.096	0.103	153
2m	AY754634	0.085	0.075	0.080	119
2r	EF116059	0.086	0.057	0.072	107
3b	AF279121	0.068	0.064	0.066	98
3g	EF116087	0.069	0.042	0.055	83
3h	AF279120	0.174	0.132	0.153	228
3i	AY434138	0.120	0.086	0.103	154
4b	EF116138	0.088	0.069	0.078	117
4l	EF116139	0.039	0.021	0.030	44
4q	AY434126	0.062	0.052	0.057	86
6e	EF116196	0.040	0.031	0.035	53
6h	EF116156	0.081	0.058	0.069	103
6l	EF116159	0.101	0.053	0.077	115
6o	AY894524	0.098	0.098	0.098	146
6r	EF116153	0.072	0.047	0.059	89
6s	EF116169	0.014	0.097	0.056	83
7a	AY434115	0.217	0.270	0.243	363
2f	AY834974	0.059	0.047	0.053	79
6k	AY834938	0.096	0.080	0.088	131
6n	AY834939	0.117	0.074	0.096	143

1g	EF694452	0.069	0.047	0.058	86
4a	AB103457	0.043	0.036	0.040	59
4m	EF694517	0.067	0.059	0.063	94
4o	|EF694422	0.063	0.052	0.058	86
1b	AF515988	0.050	0.047	0.048	72
1i	L48495	0.072	0.042	0.057	85
2c	AF515981	0.055	0.058	0.056	84
2i	AF515968	0.062	0.047	0.054	81
2j	DQ220919	0.066	0.031	0.048	72
4d	AJ291258	0.052	0.063	0.058	86
4h	AJ291249	0.043	0.036	0.040	59
4n	AY743101	0.087	0.063	0.075	112
4c	L29614	0.048	0.042	0.045	67
4g	L29618	0.076	0.086	0.081	120
1m	AF037235	0.031	0.026	0.028	42
2a	D10648	0.051	0.015	0.033	50
6q	AY735101	0.128	0.114	0.121	181
2l	AY257465	0.116	0.079	0.098	146
6m	AB103135	0.111	0.042	0.076	114
3k	AB444475	0.140	0.097	0.118	177
5a	DQ164544	0.166	0.138	0.152	227
2b	DQ666241	0.088	0.058	0.073	108
3a	DQ663603	0.124	0.097	0.111	165
6c	AB027610	0.104	0.098	0.101	151
6f	DQ640386	0.082	0.058	0.070	104
6i	DQ640367	0.041	0.026	0.033	49
6j	AB027608	0.064	0.052	0.058	87
4r	AY577585	0.069	0.058	0.063	95
1a	AF268586	0.050	0.036	0.043	65
2k	AB081066	0.076	0.081	0.078	117
6d	DQ155517	0.056	0.036	0.046	69
6p	DQ155504	0.082	0.042	0.062	92

Table 2. Tabulated results of divergence time of each genotype.

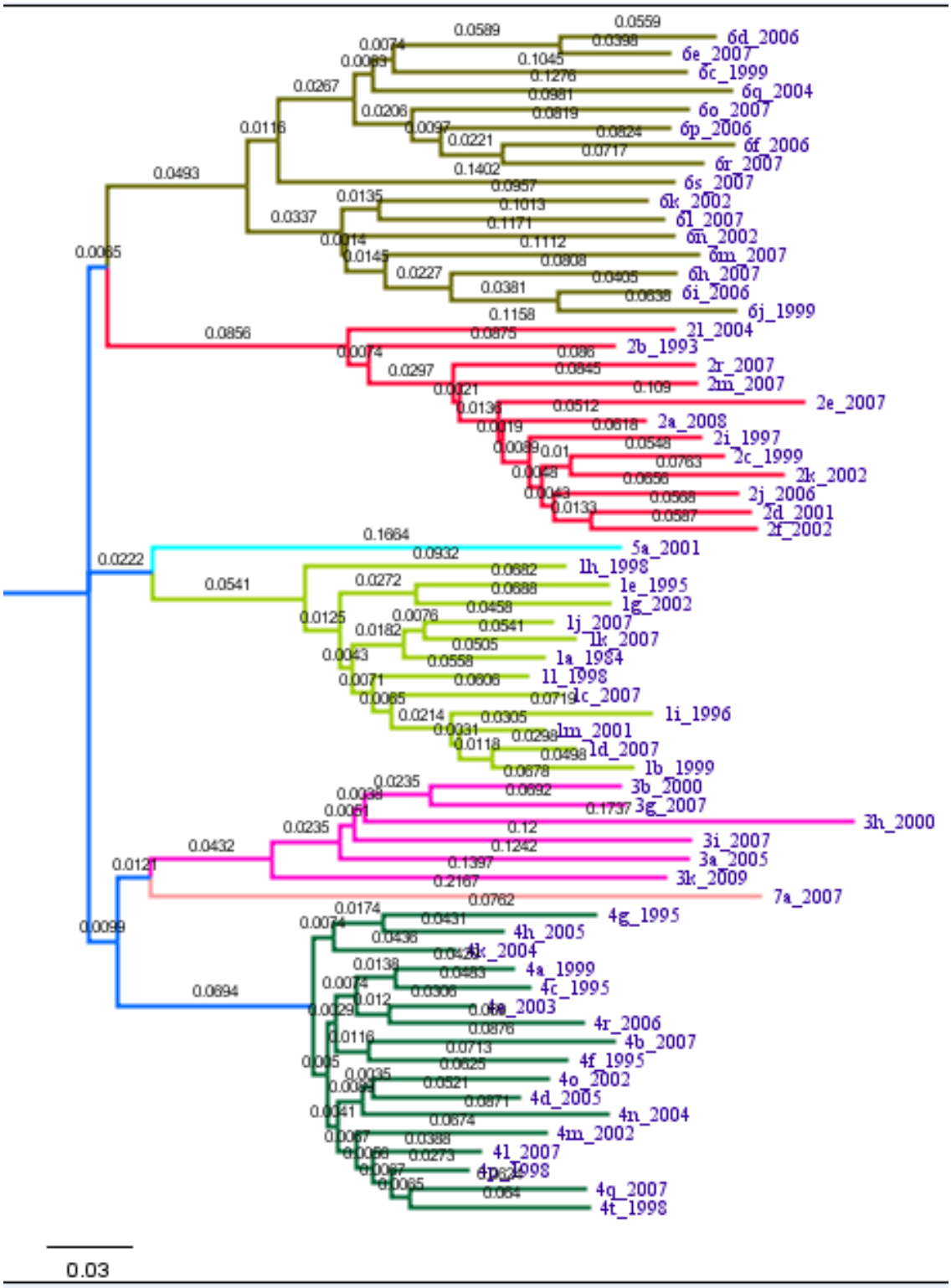

Figure 1: N-J tree

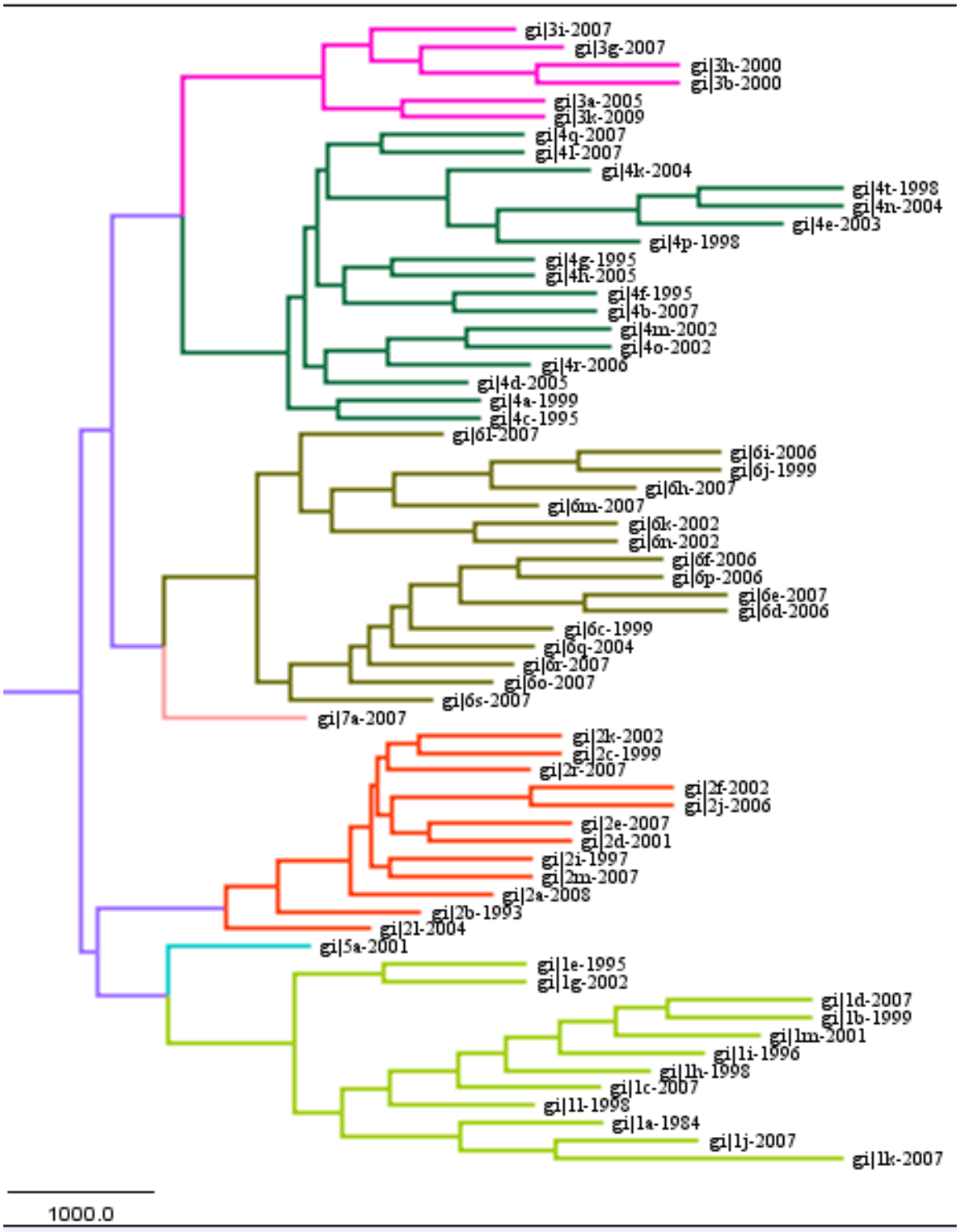

Figure 2: Consensus tree constructed from Dnapars

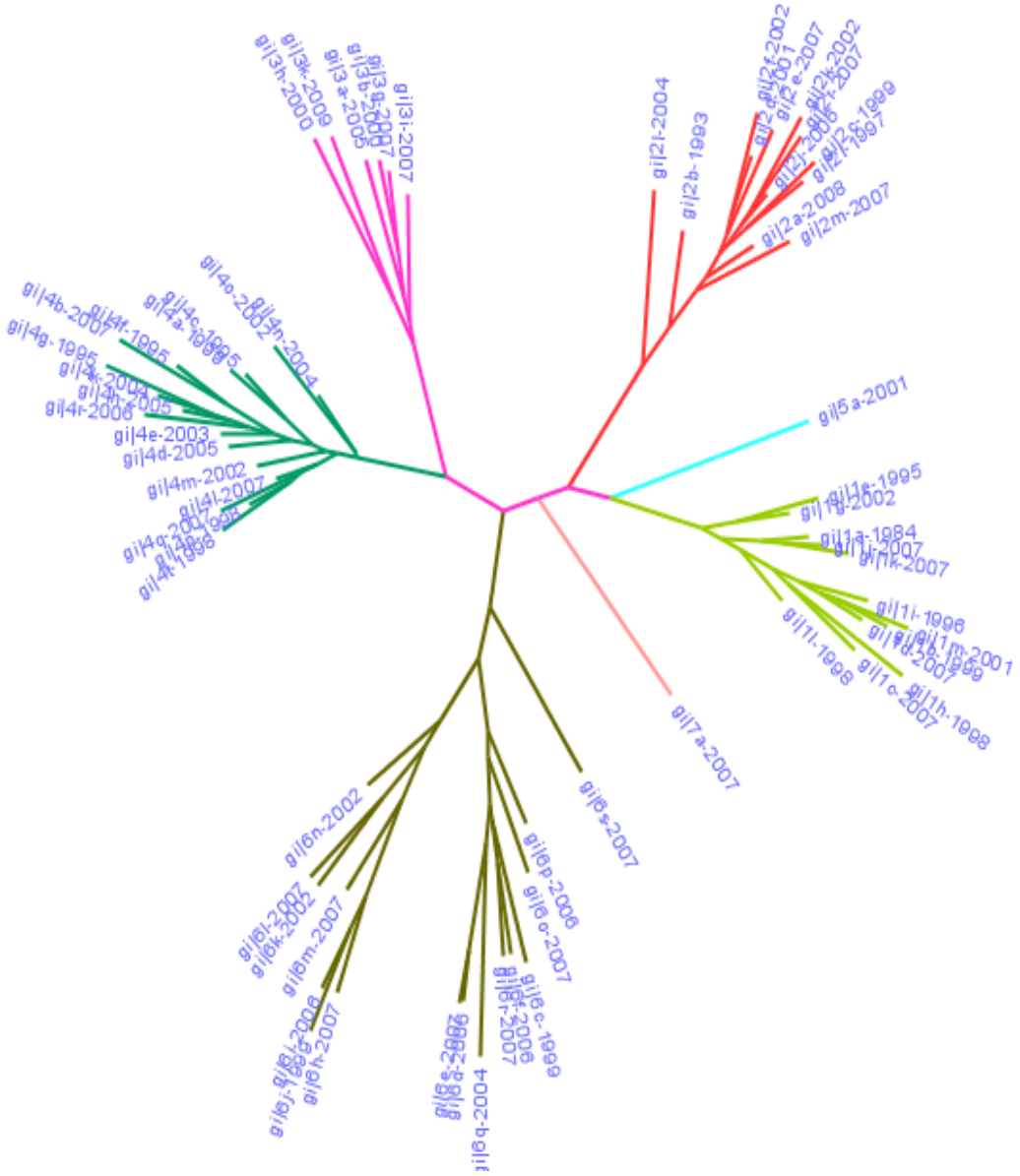

Figure 3: Phylogenetic tree prepared from Dnaml.

Reference

1. Kato N (2000). "Genome of human hepatitis C virus (HCV): gene organization, sequence diversity, and variation". *Microb. Comp. Genomics* 5 (3): 129–51.

2. Lindenbach B, Rice C (2005). "Unravelling hepatitis C virus replication from genome to function". *Nature* 436 (7053): 933–8.

3. Kumar V, Fausto N, Abbas A (editors) (2003). *Robbins & Cotran Pathologic Basis of Disease* (7th ed.). Saunders. pp. 914–7.

4. Simmonds P, et.al (2005). "Consensus proposals for a unified system of nomenclature of hepatitis C virus genotypes". *Hepatology* 42 (4): 962–73.

5.Murphy et.al.(2007)"Use of sequence analysis of the NS5B region for routine genotyping of hepatitis C virus with reference to C/E1 and 5' untranslated region sequences". J Clin Microbiol. 2007 Apr;45(4):1102-12. Epub 2007 Feb 7.

6. Morgan, G.J. (1998). "Emile Zuckerkandl, Linus Pauling, and the Molecular Evolutionary Clock, 1959-1965". Journal of the History of Biology 31 (2): 155–178.

77. HCV sequence database: Kuiken C, Yusim K, Boykin L, Richardson R. The Los Alamos HCV Sequence Database. Bioinformatics(2005), 21(3):379-8

7. Larkin M.A., Blackshields G., Brown N.P., Chenna R., McGettigan P.A., McWilliam H., Valentin F., Wallace I.M., Wilm A., Lopez R., Thompson J.D., Gibson T.J. and Higgins D.G. (2007) ClustalW and ClustalX version 2.Bioinformatics 23(21): 2947-2948.

8. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, and Kumar S (2011) MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Molecular Biology and Evolution (submitted)

9. Felsenstein J. PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle. 1993. Available through http://evolution.genetics.washington.edu/phylip.

10.Vlad I. Morariu, Balaji Vasan Srinivasan, Vikas C. Raykar, Ramani Duraiswami, and Larry S. Davis. Automatic online tuning for fast Gaussian summation. Advances in Neural Information Processing Systems (NIPS), 2008.

11. Yasuhito Tanaka, Kousuke Hanada, Masashi Mizokami, Anthony E. T. Yeo, J. Wai-Kuo Shih, Takashi Gojobori, and Harvey J. Alter. 2002. A comparison of the molecular clock of hepatitis C virus in the United States and Japan predicts that hepatocellular carcinoma incidence in the United States will increase over the next two decades. 15584–15589. PNAS vol. 99, no. 24.