Chinese Science Bulletin 2006 Vol. 51 No. 18 2236—2242

DOI: 10.1007/s11434-006-2118-9

Using maximum likelihood method to detect adaptive evolution of HCV envelope protein-coding genes

ZHANG Wenjuan[1], ZHANG Yuan[1] & ZHONG Yang[1,2]

1. School of Life Sciences, Fudan University, Shanghai 200433, China;
2. Shanghai Center for Bioinformation Technology, Shanghai 201203, China

Correspondence should be addressed to Zhong Yang (email: yangzhong@fudan.edu.cn)

Received January 25, 2006; accepted July 12, 2006

Abstract Nonsynonymous-synonymous substitution rate ratio (d_N/d_S) is an important measure for evaluating selective pressure based on the protein-coding sequences. Maximum likelihood (ML) method with codon-substitution models is a powerful statistic tool for detecting amino acid sites under positive selection and adaptive evolution. We analyzed the hepatitis C virus (HCV) envelope protein-coding sequences from 18 general geno/subtypes worldwide, and found 4 amino acid sites under positive selection. Since these sites are located in different immune epitopes, it is reasonable to anticipate that our study would have potential values in biomedicine. It also suggests that the ML method is an effective way to detect adaptive evolution in virus proteins with relatively high genetic diversity.

Keywords: adaptive evolution, positive selection, amino acid sites, hepatitis C virus, envelope proteins.

The basic process of adaptive evolution by natural selection is the replacement of one allele gene by another with a higher fitness in a population. Detecting the adaptive evolution would be helpful for better understanding of bio-evolutionary mechanism and corresponding variation in structure and function[1]. The nonsynonymous-synonymous substitution rate ratio (d_N/d_S) is an important indicator of selective pressure at the protein-coding gene, with $d_N/d_S = 1$ meaning neutral mutation, $d_N/d_S <1$ purifying selection, and $d_N/d_S>1$ diversifying positive selection (i.e. adaptive evolution). In comparison with a large amount of neutral mutations and purifying selections, positive selection is rare and hardly detected effectively because it often just occurs on a few of sites or during a short period[1]. In particular, for some protein-coding sequences with high genetic diversity that might result from relatively high mutation rates or long evolutionary history, it is much difficult to infer whether positive selection exists. For example, hepatitis C virus (HCV) is a type of RNA viruses with high mutation rates, and its genotypes have emerged in determining the clinic variation, main features of chronic infection and duration of antiviral therapy. Although the rapid variation of HCV has attracted the attention of virologists and evolution biologists, it is still unclear how HCV evades the host immune response and the mechanism of chronic infection[2].

Envelope glycoproteins E1 and E2 of HCV are involved in virus attaching to the host cell as well as in virus endocytosis and fusion with host membrane. E2 protein contains two highly variable regions called hypervariable regions 1 and 2 (HVR1 and HVR2) and two CD81-binding sites. As one of the receptors of E2 protein[3], CD81 is a bridge between virus and host cell. HVR1 is implicated in the SCARB1-mediated cell entry. It was reported that despite strong amino acid sequence variability related to strong pressures towards change, the chemicophysical properties and conformation of HVR1 were highly conserved. The conservation of positively charged residues located at specific sequence positions of HVR1 indicates that HVR1 is involved in interactions with negatively charged molecules on host cell surface. This possible interaction probably plays a role in host cell recognition and attachment[4]. HVR2 and CD81-binding sites may be involved in sensitivity and/or resistance to IFN-alpha therapy. It is therefore considered that HCV evades the host immune response through mutation in some amino acid sites of envelope proteins, which result in recognition error during HCV contacting host cell. These mutations will be fixed under pressure driven by host immune system environment and form the adaptive evolution of HCV genomes.

Previous studies focused on exploration of the adaptive evolution within an individual HCV subtype. For example, positively selected amino acid sites in the entire coding sequences of HCV subtype1b were identified[5]. The increasing availability of data storing in HCV databases allows us to analyze HCV genome evolution on a large scale of genetic diversity from more quantitative frameworks based on statistical inference[6,7]. Technologically, a number of advanced methods have been proposed to reconstruct ancestral sequences or estimate the parameters under different substitution models when calculating nonsynonymous-synonymous rate ratio. For example, when all the sites evolve independently in one substitution model, a parsimony approach can be used to infer ancestral sequences and compute substitution numbers of different types. However, estimation of parameters by the parsimony approach may be biased because it does not account for multiple substitutions on one site. The maximum likelihood (ML) method is another way to estimate model parameters and more strict. It is employed not only to select the best phylogenic tree that fit the real data[1], but also to detect select pressure on sites under different codon substitution models, especially for amino acid sites undergoing positive selection[8]. The purpose of this study is therefore to use the ML method[9] to infer adaptive evolution and positively selected amino acid sites of HCV envelope protein entire coding sequences containing all 6 HCV genotypes.

1 Materials and methods

1.1 Sequence data

The scientists that expert in the fields of HCV genetic variability and development of HCV sequence databases (such as the Hepatitis Virus Database (Japan), euHCVdb (France)[6], and Los Alamos (United States)[7]) meet to re-examine the status of HCV genotype nomenclature. HCV variants can be classified into 6 genotypes representing the 6 genetic groups defined by phylogenetic analysis[10]. The confirmed genotypes with 27 complete HCV genome sequences (18 subtypes) were defined according to the nomenclature stipulated in 2004 Heidelberg conference[10,11]. The proposal provides the framework by which the HCV databases store and provide access to data on HCV. Considering the computational workload of adaptive evolution detection and statistical significance of the data analysis, this study used the sequences of these 27 complete annotated HCV genomes (Table 1)[6,10]. The average numbers of amino acid sites for E1 and E2 were 192 and 367, respectively.

1.2 Data analysis

Amino acid sequences of E1 and E2 proteins were aligned using Clustal X 1.83[12−14]. The nucleic acid sequences were aligned according to the protein alignments with Tranalign program in the EMBOSSwin software package[15]. These nucleic acid sequences were retrieved from the GenBank database (release 155.0). In order to infer a reliable phylogeny of HCV geno/subtypes, we used the alignment result of 27 complete HCV polyprotein coding sequences for evolution tree reconstruction. Neighbor-joining method with Kimura-2 parameter model implemented in MEGA 3.1 was used for phylogenetic analysis[16−18]. Clade robustness was measured by bootstrap method with 1000 replicates.

Table 1 The entire coding sequences of HCV used in this study [a)]

Genotype		GenBank Accession No.
Genotype 1		
	1a	M62321, M67463, AF009606
	1b	D90208, M58335
	1c	D14853, AY051292
Genotype 2		
	2a	D00944, AB047639
	2b	D10988, AB030907
	2c	D50409
	2k	AB031663
Genotype 3		
	3a	D17763, D28917
	3b	D49374
	3k	D63821
Genotype 4		
	4a	Y11604
Genotype 5		
	5a	Y13184, AF064490
Genotype 6		
	6a	Y12083, AY859526
	6b	D84262
	6d	D84263
	6g	D63822
	6h	D84265
	6k	D84264

a) Classification of genotypes is available at http://euhcvdb.ibcp.fr/euHCVdb.

Nonsynonymous-synonymous substitution rate ratio ($\omega = d_N/d_S$) was calculated by site-specific models of codon substitution models according to the results of phylogeny tree and sequences alignment. An ω significantly greater than 1 means that the nonsynonmous mutations are fixed at a higher rate than synonymous mutations and the evolution of this site is driven by positive selection. The model with maximum likelihood ratio is considered as the best model to fit the data. The likelihood-ratio test (LRT) was used to compare twice the log-likelihood differences between two nested models and with a χ^2 distribution to identify the statistics significance. The degrees of freedom (d*f*) used in LRT were equal to the difference in the number of parameters between the two models[19]. We used three pairs of models to form three LRTs: M0 (one-ratio) and M3 (discrete), M1a (nearly neutral) and M2a (positive selection), and M7 (β) and M8 (β & ω). The simplest model, M0, assumes one ω for all sites. Model M1a (nearly neutral) allows two classes of sites with $0<\omega_0<1$ and $\omega_1 = 1$ in proportions p_0 of conserved sites and $p_1 = 1-p_0$ of neutral sites, respectively. Based on M1a, M2a (positive selection) adds an additional class of sites with frequency $p_2 = 1-p_0-p_1$ and an ω_2 estimated from the data. M3 (discrete) uses an unconstrained discrete distribution to model heterogeneous ω ratios among sites. M7 (β) assumes a $\beta(p,q)$-distribution for $0\leqslant\omega\leqslant1$. M8 ($\beta$ & ω) adds to M7 an extra category, with proportion p_1 of sites with ω_1, while the rest of sites (at frequency $p_0 = 1-p_1$) have ω from the $\beta(p,q)$-distribution between 0 and 1. This model can be compared with M7 to test the presence of positive sites using a likelihood-ratio test (LRT)[19]. In this study, site-specific models were used with Codeml in the PAML 3.14b package[9]. We tested positive selection over sites of coding sequences by comparing twice the log-likelihood differences between M1a vs. M2a and M7 vs. M8 with a χ^2 distribution in the LRT.

2 Results

2.1 *Detection of selective pressure at single amino acid site of E1 protein-coding sequence*

Phylogenic tree shown in Fig. 1 was consistent with the phylogeny analysis based on complete genome sequences published previously[2] and the one available in http://hcv.lanl.gov/content/hcv-db/Distances/HCV_variability.html. The results of identifying positively selected amino acid sites in the coding region of E1 are summarized in Table 2. The LRTs of adaptive evolution suggested that the model of one ω ratio for all sites (M0) was rejected when compared with model M3 (2δL = 569.66, P<0.01, df = 4). The LRT statistic for comparing M1a (nearly neutral) and M2a (positive selection) showed that M2a did not have precedence over M1a (2δL = 0, df = 2). Indeed, model M2a and M1a had the same likelihood value and the estimations of parameters under these models were similar. In M2a, p_1 and p_2 could be combined into one because $\omega_2 = 1$. In this way, M2a was equivalent to M1a. Therefore, we could not infer positive selection from this comparison. Model M8 was significantly prior to M7 (2δL = 6.24, P<0.05, df = 2). Model M3 provided three proportions of sites, p_0, p_1 and p_2 with ω ratio of 0.0141, 0.1134, and 0.3231

Table 2 Parameter estimates and likelihood scores for the coding regions of E1 and E2 proteins under different sites models

Pro	Model	p a)	lnL	Estimates of parameters	LRT	2ΔlnL	P (<0.05)	Positively selected sites
E1	M0: one-ratio	1	–9414.4282	$\omega = 0.0871$				none
	M1a: nearly neutral	1	–9331.1145	$p_0 = 0.9104, p_1 = 0.0896$				not allowed
	M2a: positive selection	3	–9331.1145	$p_0 = 0.9104, p_1 = 0.0846$ $p_2 = 0.005, \omega_2 = 1.0000$	M1a vs. M2a	0		none
	M3: discrete	5	–9129.6006	$p_0 = 0.4510, p_1 = 0.4019, p_2 = 0.1471$ $\omega_0 = 0.0141, \omega_1 = 0.1134,$ $\omega_2 = 0.3231$	M0 vs. M3	569.6552	√	
	M7: β	2	–9122.4869	$p = 0.5627, q = 4.6297$				not allowed
	M8: β & ω	4	–9119.3651	$p_0 = 0.9928, p = 0.6054, q = 5.5253$ $p_1 = 0.0072, \omega = 1.0000$	M7 vs. M8	6.2436	√	none
							P (<0.005)	
E2	M0: one-ratio	1	–17744.5092	$\omega = 0.0936$				none
	M1a: nearly neutral	1	–17293.4300	$p_0 = 0.8087, p_1 = 0.1913$				not allowed
	M2a: positive selection	3	–17277.9643	$p_0 = 0.8054, p_1 = 0.1844$ $p_2 = 0.0102, \omega_2 = 7.3185$	M1a vs. M2a	30.9314	√	1E (P>0.5), 21T (P>0.95) 8N, 14A (P>0.99)
	M3: discrete	5	–16891.4965	$p_0 = 0.4535, p_1 = 0.3335, p_2 = 0.2130$ $\omega_0 = 0.0051, \omega_1 = 0.0915,$ $\omega_2 = 0.3742$	M0 vs. M3	1706.026	√	
	M7: β	2	–16866.1685	$p = 0.2936, q = 1.7793$				not allowed
	M8: β & ω	4	–16839.6503	$p_0 = 0.9892, p = 0.3279, q = 2.4348$ $p_1 = 0.0108, \omega = 4.9507$	M7 vs. M8	53.0364	√	21T (P>0.95), 1E 8N, 14A (P > 0.99)

a) p is the number of free parameters for the ω ratios.

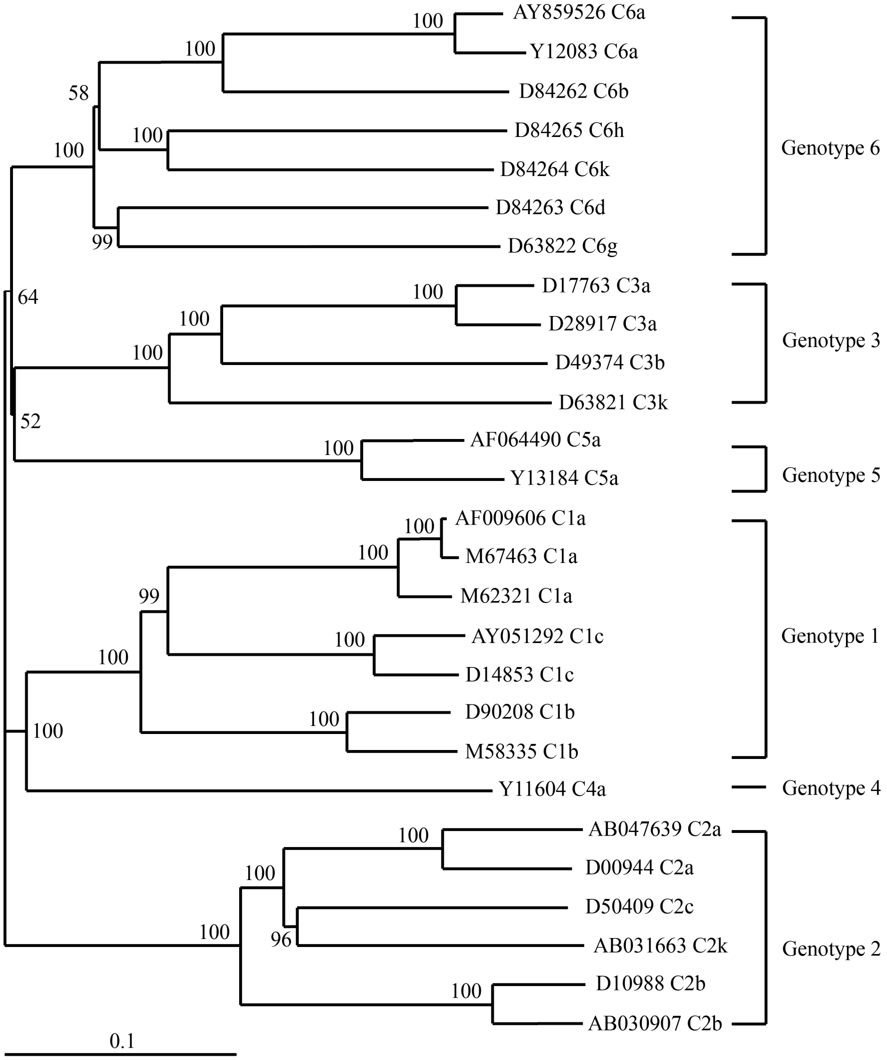

Fig. 1. Phylogenetic tree of the principal genotypes of HCV that used in this study. The tree was constructed by using the neighbor-joining method as implemented in MEGA 3.1 package.

respectively. It suggested a large proportion of sites (~85%) under strong purifying selection. Another piece of evidence for E1 gene being negatively selected was that ω in M8 was no greater than 1.

Models that allow for positively selected sites are M2a and M8 in the three pairs of nested models. However, neither of these two models suggested the existence of sites of E1 protein under positive Darwinian selection.

2.2 Detection of selective pressure at single amino acid site of E2 protein-coding sequence

The results of identifying positively selected amino acid sites in the coding region of E2 are also summarized in Table 2. The LRT statistics for comparing M0 and M3 significantly supported M3 ($2\delta L = 1706.0255$, $P<0.005$, df = 4). It suggested that heterogeneous ω ratios existed among sites. The M1a-M2a ($2\delta L = 30.9314$, $P<0.005$, df= 2) and M7-M8 ($2\delta L = 53.0364$, $P<0.005$, df = 2) comparisons indicated that M2a and M8 used for identifying positively selected sites were preferential models.

Three proportions of sites in M3, p_0, p_1 and p_2 with ω ratio of 0.0051, 0.0915, and 0.3742 respectively presented a large proportion of sites (~75%) in E2 protein under purifying selection. About 99% sites in M8 were with ω values between 0 and 1. However, different from those in E1 protein, positively selected sites with ω ratios greater than 1 were detected in M2a (ω =

7.3185) and M8 ($\omega = 4.9507$). These sites were 1E, 8N, 14A and 21T which all located in HVR1 of E2 protein.

2.3 Immune epitope analysis on sites under adaptive evolution

To assess the potential impact of the adaptive mutations, sites of E2 protein under positive selection were mapped onto immune epitope against HCV based on the epitope maps from HCV immunology database (http://hcv.lanl.gov/content/immuno/immuno-main.html)[20] (Table 3). All of the amino acid sites under adaptive evolution were located in B-cell epitopes of rat. Only one site (21T) was found in T-cell epitopes of human and transgenic mouse. Two 14A and 21T were also located in T-helper epitopes of human. It probably suggested that humoral immune response plays a key role in the immune clearance and exert more selective pressure on HCV replication than cell mediated response.

3 Discussion

Detecting adaptive evolution is a bioinformatics exploration based on the knowledge of genetics and statistics. d_N and d_S as well as their ratio ω which measures the selective pressure at the amino acid level provide powerful tools for better understanding of the effect of natural selection on molecular evolution. An ω significantly greater than 1 means that nonsynonymous mutations offer fitness advantages and this lineage (in lineage-specific models) or this critical amino acid site in the protein (in the site-specific models) are considered under positive selection driven by environment. Though ω ratio is a sensitive measure of positive selection, both lineage-specific models and site-specific models may lack power in detecting positive selection if adaptive evolution occurs at a few time points and affects a few amino acids. We need more robust statistic tools to test the hypothesis models[19,21]. Maximum likelihood method and LRT could help to identify the best codon substitution models to fit the real data, and some models such as M2a and M8 have been successfully used for detecting positive selection. We took HCV envelope glycoprotein as an example to explore the adaptive evolution driven by immune environment pressure of coding sequences of 27 HCV containing 18 geno/subtypes and found that a number of amino acid sites were under positive selection and ML could be employed for identifying the adaptive evolution of RNA virus on a large scale of genetic diversity.

Brown *et al.*[22] cloned HCV E1E2 full-length nucleotide sequences generated from serum samples of 4 chronically infected patients and identified 11 amino acid sites undergoing patient-specific adaptive evolution. In this study, we detected 4 amino acids sites of E2 protein under positive selection. Two of these sites were proved in Brown's work. Note that a region including the N-terminal 27—31 amino acid sites in E2 is known to be the most variable and is called hypervariable region 1 (HVR1)[23,24]. This region is surface-exposed[25] and has been proposed as a major target of the immune response probably because its hypervariable is correlated with immune evasion[26−28]. For all of the positively selected amino acid sites located in HVR1 of E2 protein and in some immune epitopes, adaptive evolution of HCV could be the consequence of the environment pressure directly driven by host immune response. Recent studies revealed more information about how HCV escaping from host immune system response, but more comprehensive and careful research should be done to make clear the role

Table 3 Epitope summary table of positively selected amino acid sites in the entire coding region of HCV envelop glycoproteins

Position[a)]	Position on HCV-H77[b)]	Protein	Function							
			B-cell epitope (Mab ID)[c)]	species	T-cell epitope	HLA	species	T-helper epitope	HLA	species
1E**	1 (384)	E2 (HVR1)	7/59 (IgM), 6/82a (IgG1), 6/16 (IgG1)	rat						
8N**	8 (391)	E2 (HVR1)	7/59 (IgM), 6/82a (IgG1), 6/16 (IgG1)	rat						
14A**	14 (397)	E2 (HVR1)	9/27	rat				√	DRB1*1101	human
21T*	21 (404)	E2 (HVR1)	9/27	rat	√	A2	human transgenic-mouse	√	DRB1*1101	human

a) Positively selected sites (*, $P > 95\%$; **, $P > 99\%$); b) the viral strain H77 (GenBank Accession No. M67463) is usually used as a reference strain. The amino acid position is numbered according to HCV-H77. And the number in the parentheses denotes its corresponding position of HCV polyprotein; c) Mab ID gives the name of the monoclonal antibody and with synonyms in parentheses.

of immune evasion in HCV chronically infection and explain the mechanism of HCV evolution involving immunology and virology. In other words, the positively selected sites' location indicated the immunogenicity of these sites and they might be candidate vaccination targets against HCV. The composite vaccines containing these different amino acid residues at the positively selected sites located in immune epitopes would be effective to preventing proliferation of escape mutants[5]. No amino acid sites exhibited positive selection within E1 protein in this study. It was consistent with the report from Brown *et al.*[22]. The possible reason was that E1 was unlikely to be surface-exposed[29] and not a major target for the host antibody response. It was reported that E1 protein was a poor natural immunogen for humoral response[30]. In other words, E1 protein was not under strong selective pressure of adaptive evolution driven by immune response.

Suzuki and Gojobori[5] identified 13 positively selected amino acid sites in the entire coding region of HCV subtype 1b by parsimony method. Four of these sites were located in E1 and three located in E2. It is different from the results obtained from this study. The possible reasons may be: (1) the strategies to reconstruct ancestral sequences are different. Adaptsite, the program employed in Suzuki and Gojobori's work, uses maximum parsimony method to perform reconstruction while Codeml uses a likelihood reconstruction. Thus, the reconstructed ancestral states may be different. In general, the two implementations produce similar results in dataset with high similarity among sequences. However, ML provides a more reliable result when used to analyze small-size dataset with relatively high diversity in sequence similarity[31,32]. They focused on HCV subtype 1b[5,33], and deleted any sequences with gap by pairwise-alignment with reference sequence (HCV-JS) to obtain dataset with highly similar sequences. Our study used sequences containing 18 HCV geno/subtypes with a relatively large scale of genetic diversity; thus the ML method was adopted; (2) the methods to estimate branch length are different. Adaptsite uses a neighbor-joining algorithm to estimate branch length[34] while Codeml uses a codon model M0 to do it; (3) for codons that are neighbors of stop codons, Adaptsite and Codeml count sites differently, e.g. TAC and TAT, Adaptsite counts $S = 1$ and $N = 2$, while Codeml gives 0.429 and 2.571; (4) missing data are handled differently. Adaptsite requires dataset without gaps[34] while Codeml implemented in this work allows sequences data with some gaps; and (5) mutation rates in RNA viruses are several orders of magnitude higher than those in DNA based life-forms. By limiting genome size and content, RNA virus genome can avoid deleterious mutation accumulation. It brings virus genomes to be inclined to concerted evolution or parallel evolution and own similar codon substitution patterns[35]. Therefore, it is not difficult to understand why most of amino acid sites undergoing purifying selection of genetic constraints though HCV envelope proteins still possess high mutation rates. Our study focused on a larger scale of genetic diversity than previous work. This study might probably miss particular positively selected amino acid sites of individual subtype but produce more general sites under positive Darwinian selection to all HCV genotypes.

The parsimony method of Suzuki and Gojobori[33] and the maximum likelihood method developed by Nielsen and Yang[8] are two widely used methods for detecting natural selection in homologous protein-coding sequences. However, they have their own pros and cons. The former may fail to infer positively selected sites when the branches of the phylogenetic tree are long because the maximum parsimony method is not fit for multiple substitutions. In contrast, multiple substitutions in the Nielsen and Yang method may be corrected by assuming the codon substitution model. Suzuki also attempted to employ ML to modify previous method[36]. Our study showed an application of ML method in detecting adaptive evolution in HCV envelope protein-coding sequences based on 18 geo/subtypes. It provided an instance for similar work of high diversity homologous genes analysis. Indeed, using ML method to infer adaptive evolution has been an effective strategy for some emerging viruses such as SARS-CoV. In this way we had successfully explored the adaptive evolution of SARS-CoV spike protein[37].

Acknowledgements The authors would like to thank the Bioinformatics Working Group in School of Life Sciences at Fudan University, and Masami Hasegawa and Ying Cao in the Institute of Statistical Mathematics (Tokyo) for helpful discussions. This work was supported by the National Basic Research Programs (973) (Grant Nos. 2002CB512801 & 2003CB715904), and New Century Excellent Talents in University Program .

References

1 Nei M, Kumar S. Molecular Evolution and Phylogenetics. New

York: Oxford University Press Inc. 2000

2 Simmonds P. Genetic diversity and evolution of hepatitis C virus—15 years on. J Gen Virol, 2004, 85: 3173－3188

3 Voisset C, Dubuisson J. Functional hepatitis C virus envelope glycoproteins. Biol Cell, 2004, 96: 413－420

4 Penin F, Combet C, Germanidis G, et al. Conservation of the conformation and positive charges of hepatitis C virus E2 envelope glycoprotein hypervariable region 1 points to a role in cell attachment. J Virol, 2001, 75: 5703－5710

5 Suzuki Y, Gojobori T. Positively selected amino acid sites in the entire coding region of hepatitis C virus subtype 1b. Gene, 2001, 276: 83－87

6 Combet C, Penin F, Geourjon C, et al. HCVDB: Hepatitis C virus sequences database. Appl Bioinfo, 2004, 3: 237－240

7 Kuiken C, Yusim K, Boykin L, et al. The Los Alamos HCV sequence database. Bioinformatics, 2005, 21: 379－384

8 Nielsen R, Yang Z. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics, 1998, 148: 929－936

9 Yang Z. PAML: A program for package for phylogenetic analysis by maximum likelihood. CABIOS, 1997, 15: 555－556

10 Simmonds P, Bukh J, Combet C, et al. Consensus proposals for a unified system of nomenclature of hepatitis C virus genotypes. Hepatology, 2005, 42: 962－973

11 Bowden D S, Berzsenyi M D. Chronic hepatitis C virus infection: Genotyping and its clinical role. Future Microbiol, 2006, 1: 103－112

12 Higgins D G, Sharp P M. CLUSTAL: A package for performing multiple sequence alignment on a microcomputer. Gene, 1998, 73: 237－244

13 Thompson J D, Gibson T J, Plewniak F, et al. The Clustal-X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res, 1997, 25: 4876－4882

14 Jeanmougin F, Thompson J D, Gouy M, et al. Multiple sequence alignment with Clustal X. Trends Biochem Sci, 1998, 23: 403－405

15 Rice P, Longden I, Bleasby A. EMBOSS: The european molecular biology open software suite. Trends Genet, 2000, 16: 276－277

16 Saitou N, Nei M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol Biol Evol, 1987, 4: 406－425

17 Kumar S, Tamura K, Nei M. MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. Briefings Bioinfor, 2004, 5: 150－163

18 Kimura M. A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. J Mol Evol, 1980, 16: 111－120

19 Yang Z, Nielsen R, Goldman N, et al. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics, 2000, 155: 431－449

20 Yusim K, Richardson R, Tao N, et al. The Los Alamos hepatitis C immunology database. Appl Bioinform, 2005, 4: 217－225

21 Yang Z, Bielawski J P. Statistical methods for detecting molecular adaptation. Trends Ecol Evol, 2000, 15: 496－502

22 Brown R J, Juttla V S, Tarr A W, et al. Evolutionary dynamics of hepatitis C virus envelope genes during chronic infection. J Gen Virol, 2005, 86: 1931－1942

23 Hijikata M, Kato N, Ootsuyama Y, et al. Hypervariable regions in the putative glycoprotein of hepatitis C virus. Biochem Biophys Res Commun, 1991, 175: 220－228

24 Weiner A J, Brauer M J, Rosenblatt J, et al. Variable and hypervariable domains are found in the regions of HCV corresponding to the flavivirus envelope and NS1 proteins and the pestivirus envelope glycoproteins. Virology, 1991, 180: 842－848

25 Yagnik A T, Lahm A, Meola A, et al. A model for the hepatitis C virus envelope glycoprotein E2. Proteins, 2000, 40: 355－366

26 Weiner A J, Geysen H M, Christopherson C, et al. Evidence for immune selection of hepatitis C virus (HCV) putative envelope glycoprotein variants: Potential role in chronic HCV infections. Proc Natl Acad Sci USA, 1992, 89: 3468－3472

27 Farci P, Shimoda A, Wong D, et al. Prevention of hepatitis C virus infection in chimpanzees by hyperimmune serum against the hypervariable region 1 of the envelope 2 protein. Proc Natl Acad Sci USA, 1996, 93: 15394－15399

28 Zibert A, Kraas W, Meisel H, et al. Epitope mapping of antibodies directed against hypervariable region 1 in acute self-limiting and chronic infections due to hepatitis C virus. J Virol, 1997, 71: 4123－4127

29 Allison S L, Schalich J, Stiasny K, et al. Mutational evidence for an internal fusion peptide in flavivirus envelope protein E. J Virol, 2001,75: 4268－4275

30 Fournillier A, Wychowski C, Boucreux D, et al. Induction of hepatitis C virus E1 envelope protein-specific immune response can be enhanced by mutation of N-glycosylation sites. J Virol, 2001, 75: 12088－12097

31 Yang Z, Kumar S, Nei M. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics, 1995, 141: 1641－1650

32 Zhang J, Nei M. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. J Mol Evol, 1997, 44: S139－S146

33 Suzuki Y, Gojobori T. A method for detecting positive selection at single amino acid sites. Mol Biol Evol, 1999, 16: 1315－1328

34 Suzuki Y, Gojobori T, Nei M. ADAPTSITE: Detecting natural selection at single amino acid sites. Bioinformatics, 2001, 17: 660－661

35 Holmes E C. Error thresholds and the constraints to RNA virus evolution. Trends Microbiol, 2003, 11: 543－546

36 Suzuki Y. New methods for detecting positive selection at single amino acid sites. J Mol Evol, 2004, 59: 11－19

37 Zhang Y, Zheng N, Hao P, et al. Reconstruction of the most recent common ancestor sequences of SARS-CoV S gene and detection of adaptive evolution in the spike protein. Chin Sci Bull, 2004, 49: 1311－1313