Journal of Multidisciplinary Engineering Science and Technology (JMEST)
ISSN: 2458-9403
Vol. 7 Issue 11, November - 2020

Calculation of RNA Viruses Mutation Rates Based on 96-parameter Mutational Model

Xuejuan Zhu, Wenjun Xia*

School of Mathematical Sciences, Jiangsu University, Zhenjiang, 212013, PR China

***Abstract*—Viruses are infectious agents that cannot reproduce without a host cell.Their survival depends on their ability to enter host cells, replicate themselves, and avoid the host's immune system. One therefore expects that virus and host genomes should have common features.In this paper, we have analyzed the dinucleotide patterns of RNA viruses and found that many RNA viruses have insufficient dinucleotide CpG expression like their hosts. We also used the 96-parameter mutation model to construct an RNA mutation rate calculation model, and analyzed the average mutation rate of each nucleotide in the RNA virus only under the influence of its own surrounding nucleotides.We found that different types of RNA viruses mutate at different rates.**

Keywords—RNAvirus; 96 parameter mutational model; Dinucleotide; evolution;mutation rate

I. INTRODUCTION

One of the most important structural abnormalities in the genomic DNA sequence of mammals and other vertebrates is the under-representation of CpG dinucleotides [1,4,12]. The accepted explanation for this is based on the fact that deamination of 5-methylcytosine in methylated CpG dinucleotides produces a cytosine(mC) to thymine mutation at a high frequency that is not efficiently corrected by the DNA editing system [1,2,5,6].Therefore, the frequency of CpG dinucleotides in DNA is reduced.As an infectious pathogen that only reproduces in host cells, the evolution of viruses is closely related to the nature and fate of their hosts.It is expected that the virus and its host genome should have similar characteristics. Therefore, in this article, we have analyzed the dinucleotide patterns of different types of RNA viruses and found that most viruses have similar dinucleotide patterns to their hosts [8-11,13,14].

As we all know, RNA viruses use RNA as their genetic material, and the structure of RNA is a single-stranded structure, which is unstable relative to the double-stranded structure of DNA. Therefore, compared with DNA, the genetic material (RNA) of RNA viruses is more likely to mutate under the action of the external environment [3]. If the influence of external factors is not considered, it is assumed that each nucleotide in the RNA sequence of the genetic material of the RNA virus is only affected by its surrounding nucleotides when it is mutated. So how do RNA viruses mutate?

Simmonds[1] developed an extended model of sequence evolution that allows separate mutation rates for each type of transition and transversion in each dinucleotide context against a background, separately optimized mean transition / transversion ratio (k).This model generalizes Duret and Galtier's model[7] ,in which k was fixed at 2.1 and only one context dependent mutation, ($C \to T, G$) was allowed to take a higher mutation rate. (This rate was based on observational data available at the time of the study on sequence variability in human DNA sequences.) This model allowed up to 96 different dinucleotide context-dependent mutations and used model error calculations to allow each to be systematically optimized rather than empirically assigned.Applying this model to biological DNA and RNA data sets can reproduce the observed low and high expression of dinucleotides.And it also can predict context-dependent mutations and rates. These changes and occurrences are biologically reasonable and consistent with the results and inferences obtained through different methods[1].

Next, in this paper, we will use the model(96-parameter mutation model) proposed by Simmonds to simulate and analyze the mutation process of several RNA viruses from the original sequence to the current sequence. In this mutation process, we only consider base context-dependent mutations and do not consider mutations under the influence of other external factors.

II. METHODS

A. Data resources

The genome sequences from RNA viruses were the main subject of the investigation of this paper.These RNA viruses belonged to 9 different species, with the majority of samples from Coronaviridae (n = 7), Orthomyxoviridae (n = 15), Picornavirus (n = 9),and Flaviviridae(n = 7). Those complete genome sequence were downloaded from GenBank(https://www.ncbi.nlm.nih.gov/genbank/).

B. Calculation of RNA sequence mutation rate

As for a random RNA sequence without neighboring effect ,according to Tamura's(1992) model,the base substitution rates in RNA chain satisfies the four-state Markov chain has the following form[7]:

$$Q = \begin{array}{c} \\ C \\ G \\ U \\ A \end{array} \begin{array}{c} \begin{array}{cccc} C & G & U & A \end{array} \\ \begin{bmatrix} * & \theta & k(1-\theta) & 1-\theta \\ \theta & * & 1-\theta & k(1-\theta) \\ k\theta & \theta & * & 1-\theta \\ \theta & k\theta & 1-\theta & * \end{bmatrix} \end{array}$$

Where Q_{YW} for distinct Y and W is the default rate of transformation from nucleotide Y to nucleotide W. k is the transition to transversion ratio.And θ is the equilibrium $G+C$ content. we hypothesize that the transformation rate of a nucleotide can be influenced independently by its two neighbouring nucleotides.For trinucleotide XYZ the mutation rate from Y to W (X,Y,Z,W represent a specific nucleotide)could be calculated by [1]:

$$r(X,Y\to W,Z) = f(X,Y\to W,M)Q_{YW}f(M,Y\to W,Z)$$

where $f(X,Y\to W,M)$ (M is C,G,U *or* A) is the factor giving the variation of mutation rate of $Y\to W$ when the upstream nucleotide is X , and $f(M,Y\to W,Z)$ is the factor giving the variation of mutation rate of $Y\to W$ when the downstream nucleotide is Z ,and both factors contribute independently.

RNA chains contain 4 different nucleotide,for a specific nucleotide Y has 3 possible transitions to a different nucleotide W ,thus there are 12 possible transitions $Y\to W$. Since there are 4 possible upstream nucleotide X , There are 48 factors of the form $f(X,Y\to W,M)$. Similarly , there are 48 factors of the form $f(M,Y\to W,Z)$ .Thus this yields a model with $4*3*4*2+1=97$ independent factors. In summary,the total mutation rate of a nucleotide Y is given by:

$$R(X,Y\to W,Z) = \sum_{i=1}^{3} f(X,Y\to W,M)Q_{YW}f(M,Y\to W,Z) \quad (1)$$

For example , base C ,there are 3 possible transitions to a different nucleotide $C\to A, C\to G$ *and* $C\to U$ .If the upstream of base C is the base G and the downstream is the base U ,then the mutation rate of base C is related by:

$$\begin{aligned} R(G,C\to W,U) &= f(G,C\to A,X)Q_{CA}f(X,C\to A,U) \\ &+ f(G,C\to U,X)Q_{CU}f(X,C\to U,U) \\ &+ f(G,C\to G,X)Q_{CG}f(X,C\to G,U) \end{aligned}$$

Therefore, the average mutation rate of each base in an RNA chain of length N can be calculated by the following formula:

$$\begin{aligned} \bar{R}(X,Y\to W,Z) &= \frac{\sum_{j=1}^{N} R(X,Y\to W,Z)}{N} \\ &= \frac{\sum_{j=1}^{N}\sum_{i=1}^{3} f(X,Y\to W,M)Q_{YW}f(M,Y\to W,Z)}{N} \end{aligned} \quad (2)$$

C. 96 parameter mutational model

For any specified set of mutational rates $r(X,Y\to W,Z)$,this model can simulate the mutational process starting from some arbitrary compositions until an equilibrium is reached .The method used is as follows[1,7].

Let d_{ij} be the proportion of dinucleotide ij in all dinucleotides at time u , n_j be the proportion of nucleotide j in all nucleotide at time u ,and for any i , $n_j(u) = \sum_i d_{ij}(u) = \sum_i d_{ji}(u)$ .We support that the instantaneous rate of change any nucleotide depends only on its state and the states of its two neighboring nucleotides.Let $t_{ijk}(u)$ be the proportion of trinucleotide ijk at time u , it follows that:

$$\begin{aligned} & d_{xy}(u+du) \\ &= d_{xy}(u) + \sum_{i,j,k} t_{ijk}(u)\sum_{l} r(i,j\to l,k)b((x,y),(i,j\to l,k))du \end{aligned} \quad (3)$$

where $r(i,j\to l,k)$ is the rate of transformation from trinucleotide (ijk) to trinucleotide (ilk),that is $f(i,j\to l)Q_{ij}f(j\to l,k)$, $b((x,y),(i,j\to l,k))$ is the number of change of dinucleotide(xy) during the transformation of trinucleotide(ijk) to trinucleotide (ilk).For example, $b((C,G),(A,C\to U,G))$ is -1, i.e., one dinucleotide (CpG) is lost while converting trinucleotide ACG to trinucleotide AUG .Thus,the value of might be -2,-1,0,1 ,2. Trinucleotide frequencies can $b((x,y),(i,j\to l,k))$ be inferred from dinucleotide ones:

$$t_{ijk}(u) = \frac{d_{ij}(u)d_{jk}(u)}{n_j(u)} \quad (4)$$

Equation (3) written for all possible (x,y) forms a system of 16 differential equations to describe the

instantaneous dynamic process of RNA sequences' evolution.From this nonlinear system we can observe that it always converges from any different initial composition to the same equilibrium state.Let F denote the vector of k and all the factors $f_i, i = 1,2,3,\cdots,96$,where second parameter f_0 allows k to vary,let $k = k_0 f_0, k_0 = 2.1$ .The other 96 parameters f_i are $f(X,Y \to W)$ and $f(Y \to W,Z)$ and have initial values of 1.0 .For given values of θ and F ,the steady state dinucleotide proportions can be found by substituting (4)to(3)and the 16 differential equations from a initial composition to the stable state.To measure the difference between the real data and the prediction of models more precise,we compared every value of $XpY_{o/e}$ of real data with the value of $XpY_{o/e}$ for the same value of G + C context obtained from the simulation, $XpY_{o/e} = d_{XY}/(n_X n_Y)$ .The latter can be obtained by using linear interpolation. Let the x-axis is G + C context and the y-axis is $XpY_{o/e}$,for every (x, y) from real data we can get two adjacent points (x_0, y_0) and (x_1, y_1) from the outcome of simulation, $x \in [x_0, x_1]$,thus the interpolant is given by: $y' = y_0 + (x - x_0)\frac{y_1 - y_0}{x_1 - x_0}$. Assume there are N samples,then we calculate the total sum of squares(TSS) of 16 dinucleotides by $TSS = \sum_{i=1}^{N}\sum_{j=1}^{16}(y_j' - y_j)^2$ .Finally we get the root mean square(RMS) error by

$$RMS = \sqrt{\frac{TSS}{16N}} = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{16}(y_j' - y_j)^2}{16N}} \quad (5)$$

The goal is to f_i find the minimum value of RMS and the corresponding value of the factors F ,which gives the best fit to the data.First we find the minimum value of RMS for k and each single factor in turn,and choose the factor that gives the best fit.Then we repeat the calculation allowing the value of k and the previously selected factor and each other factor in turn to vary from their default value, and find which other factor allows the best reduction in RMS value.Then this is repeated to select at each step the best addition factor to add to the previously selected ones allowing at every step the re-optimisation of the values of k and f_i for each of the previously selected nucleotides.This procedure was repeated with further context-dependent mutations until there was no further reduction in RMS distance[1].

III. Results and Discussion

A. CpG deficiency in RNA viruses

RNA viruses have evolved in vertebrate hosts and carry the genomic characteristics of their host-specific environment. For example, many RNA viruses have strong CpG defects similar to their host genomes. We analyzed more than 50 RNA viruses using the index of CpG deficiency $CpG_{o/e}$ (the index is expected to be 1 with no deficiency or excess,less than 1 if deficiency and greater than 1 if excess), and the results are shown in Figure 1. The $CpG_{o/e}$ values of these RNA viruses are basically at the bottom of the $CpG_{o/e}$ distribution graph. This indicates that these viruses exhibit CpG suppression related to their G + C composition[13-15]. As shown in Fig. 1: Different viruses are distributed in different positions in the figure. Coronavirus are mainly distributed in the upper part of the picture, while picornaviruses are mainly distributed in the lower part of the picture. This indicates that different types of viruses have different degrees of CpG deficiency, which may be related to the specific environment of their host [8-11].

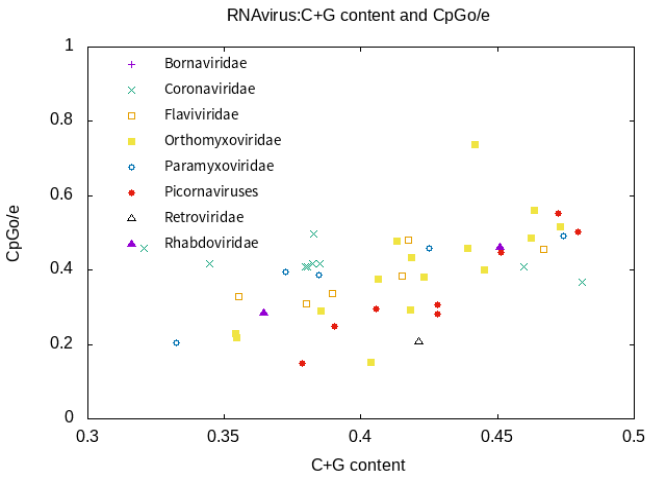

Fig. 1. Different types of RNA viruses have different combinations of viral genome C+G% content and observed/expected CpGo/ ratio

B. Compare mutation rates between different viruses

As we all know, for the original RNA sequence (that is, the RNA sequence is CGUA) without being affected by external factors, the mutation (that is, base substitution, deletion or insertion) of each base in the sequence is random and non-directional during each mutation process. In this paper, we mainly simulated how the RNA sequence of a certain virus evolved from the original sequence through mutation when the evolution rate of a nucleotide is influenced by the neighbouring nucleotides.

First, we used the 96-parameter mutation model to analyze 50 different RNA viruses (the 50 viruses come from 9 different virus categories).This can simulate the

mutational process from the original RNA sequence to any RNA virus sequence and obtained the predicted values of 96 context-dependent mutations in the evolution of gene sequences. Then use the mutation rate calculation formula to obtain the average mutation rate of each base in the mutation process of each RNA sequence.And The results are shown in Fig. 2. These RNA viruses are distributed in various parts of the virus base average mutation rate graph and the base mutation value of these RNA viruses is greater than 2. This is consistent with the fact that RNA viruses have unstable single-stranded structures and are prone to base mutations. It also shows that different types of RNA viruses have different degrees of base mutation capabilities.The mutation rate derived from other experimental data may be different from our results. This also shows from the side that the RNA sequence is not only affected by the neighbouring nucleotides during the mutation process, but also affected by other external factors.For example,from the Fig.2 we can see that the coronaviruses show the highest mutation rate among all the viruses we studied.However,in actual situations,the mutation rate of coronaviruses is not fast,which may be due to the coronaviruses has a more complicated genetic mutation error coerrection mechanism[3].This error correction mechanism can correct some errors in the RNA replication process.

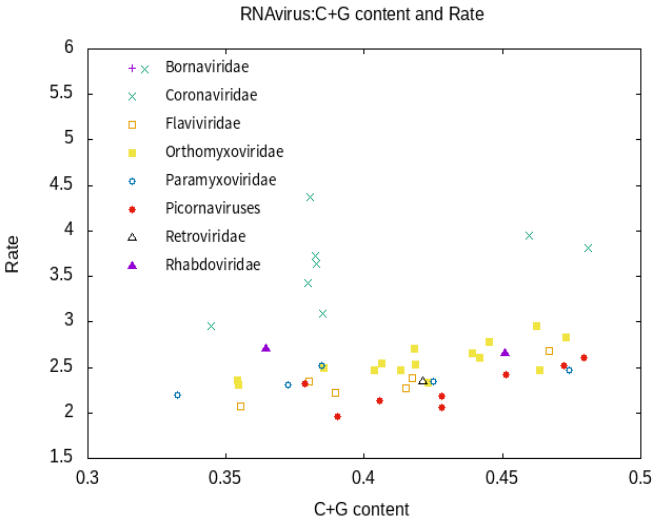

Fig. 2. Mutation rate of various viruses

TABLE I. MUTATION RATE OF VARIOUS VIRUSES

	SeqName	Rate	C+G	CpG
	Filoviridae:			
1	zaire Ebola virus	2.443028	0.411129	0.603683
2	Lake Victoria marburgvirus	2.439031	0.382921	0.530005
	Bornaviridae:			
3	Borna-disease-virus1	2.238724	0.501909	0.652825
4	Borna-disease-virus2	2.621557	0.50101	0.653063
	Coronaviridae:			
5	HCoV-229E	3.639146	0.382619	0.496455
6	HCoV-OC43	3.816713	0.480932	0.480932
7	HCoV-NL63	2.960315	0.344609	0.416797
8	HCoV-HKU1	5.779956	0.320591	0.458391
9	SARS-CoV	3.951595	0.459531	0.407877
10	MERS-CoV	3.741056	0.558245	0.411348
11	SARS-CoV-2	3.427979	0.379728	0.407702
12	RmYN02	3.719508	0.382394	0.416280
13	Bat-coronavirus-RaTG13	4.367920	0.380372	0.408752
14	Pangolin-cov	3.093104	0.384997	0.417661
	Paramyxoviridae:			
15	Human-parainfluenza-virus1	2.309548	0.372436	0.394037
16	Human-parainfluenza-virus2	2.520780	0.384566	0.38745
17	Measles-virus	2.471405	0.473952	0.490783
18	Mumps-virus	2.340554	0.424922	0.457216
19	Human-respiratory-syncytial-virus	2.191165	0.332348	0.205534
	Picornaviruses:			
20	Hepatitis-A-virus	2.320599	0.378577	0.148964
21	Hepatitis-C-virus	2.490588	0.582314	0.730821
22	Human-rhinovirus89	1.956751	0.39038	0.249615
23	Human-rhinovirusB	2.131649	0.405713	0.296582
24	Human-rhinovirusC	2.059348	0.427948	0.307993
25	Human-enterovirusA	2.608330	0.479428	0.50248
26	Human-enterovirusB	2.519290	0.472324	0.551722
27	Human-enterovirusC	2.426053	0.451425	0.448362
28	Human-enterovirusD	2.183199	0.428049	0.280901
	Retroviridae:			
29	Human-immunodeficiency-virus1	2.347086	0.421196	0.206086
	Rhabdoviridae:			
30	Rabies-virus	2.655982	0.450788	0.460571
31	Human-metapneumovirus	2.707610	0.364229	0.28339
	Flaviviridae:			
32	Dengue-virus	2.679634	0.466791	0.455458
33	Human-spumaretrovirus	2.339411	0.455458	0.308658
34	Malakal-virus	2.075011	0.355219	0.329517
35	Simian-foamy-virus	2.222635	0.389552	0.336987
36	Vesicular-stomatitis-Indiana-virus	2.387972	0.417346	0.480389
37	Visna_Maedi-virus	2.276517	0.414973	0.38366
	Orthomyxoviridae:			
38	Avian-paramyxovirus6	2.469744	0.463476	0.559823
39	Influenza-A-virus1	2.783140	0.445175	0.399481
40	Influenza-A-virus2	2.333528	0.422867	0.380658
41	Influenza-A-virus3	2.657023	0.439146	0.457327
42	Influenza-A-virus4	2.544923	0.406393	0.376698
43	Influenza-A-virus5	2.957343	0.462258	0.487363
44	Influenza-A-virus6	2.527520	0.418637	0.432855
45	Influenza-A-virus7	2.831532	0.473042	0.5164
46	Influenza-A-virus8	2.602947	0.441667	0.736557
47	Influenza-B-virus2	2.490686	0.385214	0.289024
48	Influenza-B-virus4	2.707920	0.418172	0.291859
49	Influenza-B-virus8	2.472600	0.413321	0.477466
50	Influenza-C-virus2	2.312080	0.354525	0.217366
51	Influenza-C-virus3	2.353184	0.353991	0.230163
52	Influenza-C-virus4	2.463496	0.403455	0.150776

REFERENCES

[1] Simmonds, Peter, et al. ”Modelling mutational and selection pressures on dinucleotides in eukaryotic phyla-selection against CpG and UpA in cytoplasmically expressed RNA and in RNA viruses.” BMC genomics 14.1 (2013): 610.

[2] Greenbaum, Benjamin D., et al. ”Patterns of evolution and host gene mimicry in influenza and other RNA viruses.” PLoS Pathog 4.6 (2008): e1000079.

[3] Yao, Hangping, et al."Patient-derived SARS-CoV-2 mutations impact viral replication dynamics and infectivity in vitro and with clinical implications in vivo." Cell Discovery 6.1 (2020): 1-16.

[4] Russell GJ, Walker PM, Elton RA, Subak-Sharpe JH: Doublet frequency analysis of fractionated vertebrate nuclear DNA. J Mol Biol 1976,108:1‑23.

[5] Bird AP: DNA methylation and the frequency of CpG in animal DNA.Nucleic Acids Res 1980, 8:1499‑1504.

[6] Coulondre C, Miller JH, Farabaugh PJ, Gilbert W: Molecular basis of base substitution hotspots in Escherichia coli. Nature 1978, 274:775‑780.

[7] Duret L, Galtier N: The covariation between TpA deficiency, CpG deficiency, and G+C content of human isochores is due to a mathematical artifact. Mol Biol Evol 2000, 17:1620–1625.

[8] Karlin S, Doerfler W, Cardon LR (1994) Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses? J Virol 68: 2889–97.

[9] Auewarakul P (2004) Composition bias and genome polarity of RNA viruses.Virus Res 109: 33–7.

[10] Rima BK, McFerran NV (1997) Dinucleotide and stop codon frequencies in single-stranded RNA viruses. J Gen Virol 78: 2859–70.

[11] Sewatanon J, Srichatrapimuk S, Auewarakul P (2007) Compositional bias and size of genomes of human DNA viruses. Intervirology 50: 123–32.

[12] Russell GJ, Walker PM, Elton RA, Subak-Sharpe JH: Doublet frequency analysis of fractionated vertebrate nuclear DNA. J Mol Biol 1976,108:1‑23.

[13] Karlin S, Doerfler W, Cardon LR: Why is CpG suppressed in the genomesof virtually all small eukaryotic viruses but not in those of large eukaryotic viruses? J Virol 1994, 68:2889‑2897

[14] Belalov IS, Lukashev AN: Causes and implications of codon usage bias in RNA viruses. PLoS One 2013, 8:e56642

[15] Rima BK, McFerran NV (1997) Dinucleotide and stop codon frequencies in single-stranded RNA viruses. J Gen Virol 78: 2859‑70.