Host Specific Codon Usage Pattern of H1N1 Influenza A Viruses

Ming-Wei Su,[1] P. C. Chen,[2] Woei C. CHU[1*]
[1]Institute of Biomedical Engineering
National Yang Ming University, Beitou, Taipei 112, Taiwan
[2]*Department of Medical Imaging and Radiological Sciences, I-Shou University, Taiwan*
[*]wchu@ym.edu.tw

Hanns S. YUAN
Institute of Molecular Biology
Academia Sinica
Nang Kang, Taipei 115, Taiwan

***Abstract*—Codon usage preference and the highly expressed genes have strong correlations occur in many organisms. Codon usage preference of viruses may evolve much similar with its infected host to increase the fitness. In this study we investigated differences in codon usage preferences among influenza A H1N1 viruses which infected avian, swine and human, and may cause major pandemic around world. The relative synonymous codon usage (RSCU) indices of HA gene in H1N1 viruses were calculated and we further incorporate the principal component analysis (PCA) to characterizing different host infected viruses. Host-specific codon usage pattern of H1N1 viruses were reported in this study. In 2009 influenza A H1N1 virus was a major epidemic challenge of disease control department. We propose to use the codon-based method to gain a better understanding of the features of virus genome and evolutionary processes.**

Keywords-influenza A H1N1 virus; host; codon usage; principal component analysis

I. INTRODUCTION

Proteins are composed of 20 amino acids, and each amino acid is encoded by triplets of mRNA called codons. Because mRNA composed 4 nucleotides, so there are 64 possible triplet combinations to encode only 20 amino acids, most amino acids are encoded by more than one codon. Those codons that code the same amino acids are referred to as synonymous codons show in figure 1.

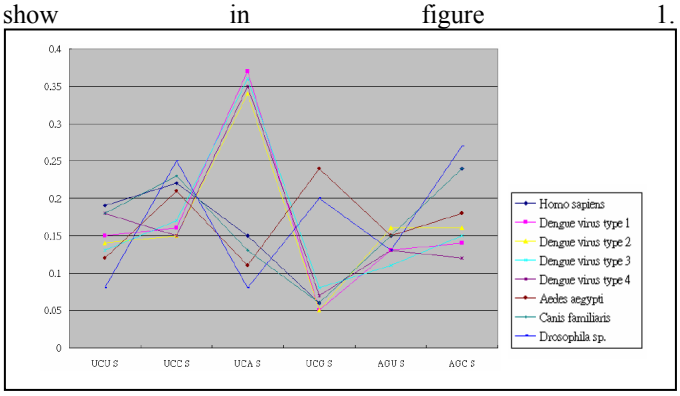

Figure 1. Synonymous codons of serine between different species.

Among different species we can easily find that the codons are not used with equal frequencies to code serine. For example, in dengue virus type 1, the codon UCA is much more frequently used than the other five codons. In contrast the codon UCG appears to be rarely used. The choice of a preference codon is different from one organism to another. It is observed that some codons are used more frequently than others in different organisms. Codon usage preference and the highly expressed genes have stronger selective correlation were first observed in bacteria [1, 2]. We hypothesize that virus codon usage preference may evolve differently due to the infected host species. Synonymous codon usage was nonrandom, and different genome had different "preferred" codons for a given amino acid. When sequence data are accumulated, how to explain this nonrandom use of synonymous codon became an important issue to biologists.

Influenza A virus is a negative-stained RNA virus majorly infected in human, swine and avian and caused the 1918 Spanish flu pandemic that killed 50 to 100 million people around the word and cause the major pandemic in 2009. Swine has suggested being an intermediary host species between human and avian influenza viruses [3, 4]. In this study, we make choice of different host infected H1N1 viruses to analysis their codon usage preference. The surface glycoprotein hemagglutinin (HA) is the host binding protein to initialize virus infection. HA proteins preferentially bind to specific receptors govern the virus particle into dissimilar host species [5]. Due to the importance of HA protein, we chosen HA gene sequences to facilities our study of virus and host codon usage relationships.

In Escherichia coli, Saccharomyces cerevisiae, Caenorhabditis elegans, Drosophila and Arabidopsis thaliana, highly expressed genes have a strong selective preference for codons with a high concentration of the corresponding tRNA molecule, whereas genes expressed at lower levels display a more uniform pattern of codon usage [2, 6]. These earlier results indicate that the codon usage preference of viruses also may influenced by the host environment. Thus these studies raise the possibility that the codon usage preference of a virus may be used as an indicator of its host species.

An analysis of H3N2, H9N2 and H5N1 showed that there are no obvious differences in codon usage patterns among virus isolated from human, avian and swine host [7]. Synonymous

978-1-4244-4713-8/10/$25.00 ©2010 IEEE

codon usage between different subtypes of influenza A viruses is different and mutation pressure is most important in synonymous codon usage [8]. We first revealed the host-specific codon usage pattern in influenza A virus H1N1. In the Principal Component Analysis (PCA), three different host infected H1N1 viruses are clustering according to its infected host species to demonstrate the host-specific codon usage pattern in H1N1 viruses.

II. Material and Method

A. Sequence Analysis

The HA gene sequences of 971 influenza A H1N1 viruses were genetically and ecologically diverse, epidemic in human, avian and swine collected from NCBI influenza virus resource [9]. Whole data sets have 68 avian host H1N1 sequences, 761 human host H1N1 sequences and 142 swine host H1N1 sequences. Sequences aligned by clustalw [10] software, and further genome composition analysis using DnaSP [11] software program.

B. Relative Synonymous Codon Usage (RSCU)

In order to analyze codon usage patterns of a virus genome, the first step is to calculate its Relative Synonymous Codon Usage (RSCU) values [6]. RSCU value for a specific codon (i) is given by

$$\mathrm{RSCU_i} = Xi / \sum(\mathrm{X_I} / n)$$

Where Xi is the counting numbers of ith codon for a given amino acid; $\sum \mathrm{X_I}$ is the sum of the occurrence numbers for all the synonymous codons in a certain amino acid; and n is the number of synonymous codons for a specific amino acid [6]. Methionine and tryptophan are only associated with one single codon, together with the three stop codons, were excluded from the codon usage preference analysis. A final 971 rows of virus strain and 59 columns of RSCU index to constructed the influenza A H1N1 codon profile matrix show in figure 2.

RSCU values

n different virus strains

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{159} \\ x_{21} & x_{22} & \cdots & x_{259} \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{n59} \end{bmatrix}$$

Figure 2. Codon profile matrix.

This matrix was further feed to the principal component analysis to confer the codon usage analysis.

C. Principal Component Analysis (PCA)

Principal component analysis (PCA) [12] is a multivariate analysis method, where the columns of the data matrix are generally a set of variables and the rows are a relatively homogeneous sample objects. PCA can be used for the differentiation of genes according to their codon usage. It is a method of identifying patterns in data, and presenting the data set in such a way to highlight their similarities and differences. It gives a more direct observation the space distribution of the complete data. PCA implementation can be divided into six steps:

STEP 1: Get data set from measurements.

STEP 2: Subtract the mean from each of the data dimensions, thus producing a zero-mean data set.

STEP 3: Calculate the covariance matrix.

STEP 4: Calculate the eigenvectors and eigenvalues of the covariance matrix.

STEP 5: Determine how many components is used to forming a feature vector.

STEP 6: Deriving the new data set.

III. Results

A. Codon Usage Preferences of H1N1 Viruses

The coding region of each viral strain was extracted and the RSCU values were calculated by our developed Perl script. Average of RSCU values of H1N1 viruses HA gene among three different kinds of host infected viruses. Codon AGA has the highest RSCU value among all three different host viruses indicate that arginine is the most biased amino acid in codon usage preference. We can found that in two-fold degenerate amino acids the codon usage bias were unobvious than others. Besides amino acid phenylalanine, AT-ended codons were generally used more preferentially in all three different host viruses among all amino acids. Avian host H1N1 viruses used slight more GC-end codons than other host infected viruses. Human and swine H1N1 virus isolates reveal much similar codon usage preference than avian H1N1 viruses.

B. Principal Component Analysis of H1N1 viruses

To study codon usage preference of 971 influenza A H1N1 viruses we utilize Principal Component Analysis (PCA) to visualizing all 59 codons in a three dimensional codon usage space. We applied the PCA on the H1N1 HA gene RSCU values for all three different host viruses to realize the codon usage relationships among influenza A H1N1 viruses. In order to avoid the amino acid compositional effect, each virus is represented as a 59 dimensional RSCU vector to facility the principal component analysis. To explore the codon usage pattern among H1N1 viruses, a 971 × 59 RSCU matrix was processed by PCA to calculate the principal components (PCs) on the purpose of highlight the similarities and differences in codon usage patterns in all three different host H1N1 viruses. Figure 3 shows the accumulated variance explanted by first ten principal components.

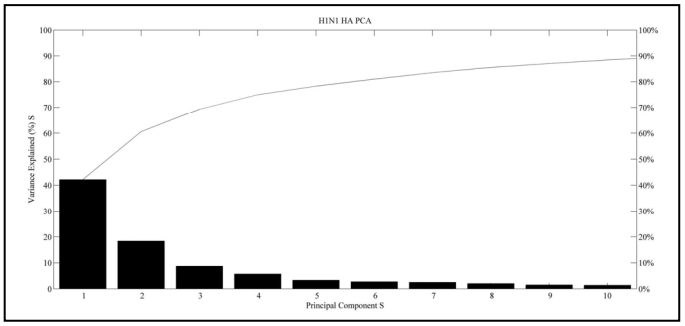

Figure 3. Accumulate variance of first ten principal components.

First PC explained 46.83% of the variance among 59 RSCU indices. The first two PCs accounted for 60.57% of the variance and the first three PCs accounted for 69.3% of the variance in original dataset. Influenza A H1N1 codon usage space formed by first three principal component axes shows in figure 4.

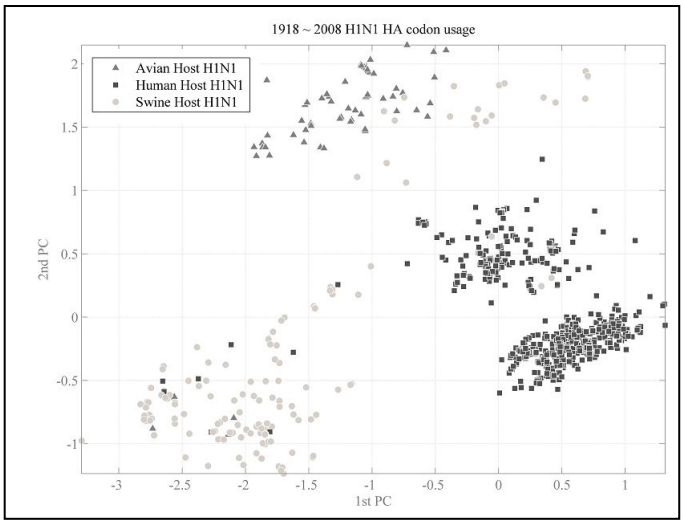

Figure 4. PCA of three different host infected H1N1 virus HA genes.

Avian host H1N1 virus showed in figure 4 plotted in triangular, human host H1N1 virus showed in square and swine H1N1 viruses represented by circle. Swine and human infected virus groups can be differentiated by first principal component axis. Second and third principal component can help us to further separate avian host H1N1 viruses more clearly. Figure showed that the codon usage preference categorized by the first three principal components possessed sufficient information to figure out the host specific codon usage pattern of influenza A virus H1N1. First three principal component axes primary affected by six-fold and four-fold degenerate amino acids due to the obviously bias of codon usage. In figure 4 avian host H1N1 viruses cluster, we can also found another small group of avian-like swine H1N1 viruses mainly composed by European isolates.

IV. DISSCUSSION

Influenza A H1N1 viruses are distributed worldwide and cause epidemic disease in various animal host including human, avian and swine. Virulence may influences by viral replication efficiency, transmissibility, host adaptation and tissue tropism. Viruses can overcome the host species barriers by evolutional constrain affected on the virus-host interaction to sustained transmission in a new host species [13]. Transmission of influenza viruses from swine to human occurred occasionally in past years. Different with 2009 pandemic, most of these cross species infections human can be seen as the dead-end host and well not further transmitted to second contact human to make epidemic. In 2009 H1N1 viruses of swine origin, first document case was on 11 March in Mexico City and as of 4 May over 1000 confirmed cases around 21 countries [14]. An earlier study in bacteriophages revealed that they usually presented the same codon usage preference of their bacterial hosts [15]. Aims of our study is to investigate the possible similarities of codon usage between viruses and its host, and to conferred the prospect of shaping codon usage preference inside the viral genomes which might influenced by the host. HA gene is hypothesized to be the major factor for species specificity of Influenza A viruses because it is the receptor binding protein [4].

Principal component analysis identified three different host infected H1N1 virus cluster. The first three principal components are dominated by six-fold and four-fold amino acids due to the most biased in codon usage preference. Host specific codon usage preference was first determined among influenza A H1N1 viruses in our study, and not occurs in H3N2, H9N2 and H5N1 subtypes [7]. In swine host viruses cluster, there are also appear five avian and eight human host H1N1 viruses, point out that the swine may be more sensitive in different host influenza viruses than human and avian. This result further supported that different host influenza viruses might occur gene reassortment in swine to generate novel synthesize strain causing pandemic. Another small group of swine viruses occur in the avian cluster, this group of swine viruses represented the evolutional history of avian H1N1 virus introduced into swine populations in Europe around 1979 so called avian-like H1N1 viruses [16].

Virus evolution influenced by host species is a very complex and interest issue, in this study we use virus and host codon usage as an approach to reveal the virus host adapt evolution. Influenza A virus is negative strain RNA virus, rely on the higher mutation rate and purifying selection adapt to a certain host species. Comprehensive codon usage analysis we proposed in this study are very helpful to understand the process governing the virus evolution, especially the role of random mutation and selection.

ACKNOWLEDGMENT

This work was supported by research grants from Ministry of Education, aim for the Top University Plan; the Summit Projects of Academia Sinica, Taiwan, and the National Science Council, Taiwan (NSC 97-2320-B-010-003-MY3).

REFERENCES

[1] M. Gouy and C. Gautier, "Codon usage in bacteria: correlation with gene expressivity," Nucleic Acids Res, vol. 10, pp. 7055-74, Nov 25 1982.

[2] R. Grantham, C. Gautier, M. Gouy, R. Mercier, and A. Pave, "Codon catalog usage and the genome hypothesis," Nucleic Acids Res, vol. 8, pp. r49-r62, Jan 11 1980.

[3] I. H. Brown, "The epidemiology and evolution of influenza viruses in pigs," Vet Microbiol, vol. 74, pp. 29-46, May 22 2000.

[4] D. van Riel, V. J. Munster, E. de Wit, G. F. Rimmelzwaan, R. A. Fouchier, A. D. Osterhaus, and T. Kuiken, "H5N1 Virus Attachment to Lower Respiratory Tract," Science, vol. 312, p. 399, Apr 21 2006.

[5] S. Karlin and J. Mrazek, "What drives codon choices in human genes?," J Mol Biol, vol. 262, pp. 459-72, Oct 4 1996.

[6] P. M. Sharp, T. M. Tuohy, and K. R. Mosurski, "Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes," Nucleic Acids Res, vol. 14, pp. 5125-43, Jul 11 1986.

[7] T. Zhou, W. Gu, J. Ma, X. Sun, and Z. Lu, "Analysis of synonymous codon usage in H5N1 virus and other influenza A viruses," Biosystems, vol. 81, pp. 77-86, Jul 2005.

[8] Y. Bao, P. Bolotov, D. Dernovoy, B. Kiryutin, L. Zaslavsky, T. Tatusova, J. Ostell, and D. Lipman, "The influenza virus resource at the National Center for Biotechnology Information," J Virol, vol. 82, pp. 596-601, Jan 2008.

[9] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," Nucleic Acids Res, vol. 22, pp. 4673-80, Nov 11 1994.

[10] J. Rozas, J. C. Sanchez-DelBarrio, X. Messeguer, and R. Rozas, "DnaSP, DNA polymorphism analyses by the coalescent and other methods," Bioinformatics, vol. 19, pp. 2496-7, Dec 12 2003.

[11] G. Perriere and J. Thioulouse, "Use and misuse of correspondence analysis in codon usage studies," Nucleic Acids Res, vol. 30, pp. 4548-55, Oct 15 2002.

[12] I. T. Jolliffe, Principal Component Analysis, 2nd ed., 2002.

[13] J. Cohen, "Swine flu outbreak. Out of Mexico? Scientists ponder swine flu's origins," Science, vol. 324, pp. 700-2, May 8 2009.

[14] T. Kunisawa, S. Kanaya, and E. Kutter, "Comparison of synonymous codon distribution patterns of bacteriophage and host genomes," DNA Res, vol. 5, pp. 319-26, Dec 31 1998.

[15] T. Ito, "Interspecies transmission and receptor recognition of influenza A viruses," Microbiol Immunol, vol. 44, pp. 423-30, 2000.

[16] C. Scholtissek, H. Burger, P. A. Bachmann, and C. Hannoun, "Genetic relatedness of hemagglutinins of the H1 subtype of influenza A viruses isolated from swine and birds," Virology, vol. 129, pp. 521-3, Sep 1983