European Journal of Epidemiology (2006) 21: 511–519
DOI 10.1007/s10654-006-9031-z


GENETIC EPIDEMIOLOGY

# Genomic Analysis of *Influenza A Viruses*, including Avian Flu (H5N1) Strains

Insung Ahn[1,2], Byeong-Jin Jeong[2,3], Se-Eun Bae[2], Jin Jung[2] & Hyeon S. Son[2,3]
[1]*Bioinformatics Team, Supercomputing Center, Korea Institute of Science and Technology Information, Yusong-Gu, Daejon, Korea;* [2]*Laboratory of Computational Biology & Bioinformatics, Graduate School of Public Health, Seoul National University, 28, Yongon-Dong, Chongno-Gu, Seoul, Korea;* [3]*Interdisciplinary Graduate Program in Bioinformatics, College of Natural Science, Seoul National University, Gwanak-Gu, Seoul, Korea*

Accepted in revised form 13 June 2006

**Abstract.** This study was designed to conduct genomic analysis in two steps, such as the overall relative synonymous codon usage (RSCU) analysis of the five virus species in the *orthomyxoviridae* family, and more intensive pattern analysis of the four subtypes of *influenza A virus* (H1N1, H2N2, H3N2, and H5N1) which were isolated from human population. All the subtypes were categorized by their isolated regions, including Asia, Europe, and Africa, and most of the synonymous codon usage patterns were analyzed by correspondence analysis (CA). As a result, *influenza A virus* showed the lowest synonymous codon usage bias among the virus species of the *orthomyxoviridae* family, and *influenza B* and *influenza C virus* were followed, while suggesting that *influenza A virus* might have an advantage in transmitting across the species barrier due to their low codon usage bias. The ENC values of the host-specific HA and NA genes represented their different HA and NA types very well, and this reveals that each *influenza A virus* subtype uses different codon usage patterns as well as the amino acid compositions. In NP, PA and PB2 genes, most of the virus subtypes showed similar RSCU patterns except for H5N1 and H3N2 (A/HK/1774/1999) subtypes which were suspected to be transmitted across the species barrier, from avian and porcine species to human beings, respectively. This distinguishable synonymous codon usage patterns in non-human origin viruses might be useful in determining the origin of *influenza A viruses* in genomic levels as well as the serological tests. In this study, all the process, including extracting sequences from GenBank flat file and calculating codon usage values, was conducted by Java codes, and these bioinformatics-related methods may be useful in predicting the evolutionary patterns of pandemic viruses.

**Key words:** Avian flu, Correspondence analysis, Genomic analysis, *Influenza A virus*, Synonymous codon usage

## Introduction

The first human infection from avian *influenza A virus* (H5N1 subtype), which resulted in 33% fatality, was reported in Hong Kong during the 1997 outbreak of bird flu in poultry [1, 2]. About 6 years later, an outbreak of avian flu virus (H5N1) occurred among poultry in eight other Asian countries, South Korea, China, Japan, Cambodia, Indonesia, Laos, Thailand, and Viet Nam. At that time, over 100 million birds died from bird flu or were killed by people in efforts to control the outbreak [3]. This outbreak began in South Korea and then spread to Viet Nam, Japan, Thailand, Cambodia, China, Laos, and Indonesia. Among these countries, only Viet Nam and Thailand reported human infections, which resulted in 80% and 66.7% fatality rates, respectively. These rates were much higher than that of the Hong Kong outbreak of 1997 [1], but the origin of the virus responsible has not yet been identified. On February 2006, European Union authorities held an urgent meeting after the H5N1 avian influenza virus was found in dead swans in Italy, Greece, Slovenia, and Austria. They agreed that Europe would be the next target for H5N1 avian flu in the near future, and the foothold for that pandemic would be Africa. Like other Asian countries where the human infections have occurred, the majority of chickens are kept in and around people's homes in Africa. These environmental conditions would provide more opportunities for H5N1 influenza viruses to contact with human beings [4]. The recent appearance of avian flu in Europe suggests that long-distance spread of the virus is also possible via migratory birds [5].

Influenza viruses are classified into three groups, such as A, B, and C groups, according to the antigenicity of their internal viral nucleoproteins. Type A viruses cause the worldwide epidemics of the respiratory disease influenza, and both type A and B viruses cause epidemics, when type C viruses cause only minor upper respiratory illness. The primary hosts of *influenza A viruses* are wild aquatic birds,

such as duck, terns and shore birds, and influenza B viruses mainly infect humans, with infections of non-human animals being reported only rarely [6, 7]. *Influenza A virus*, is a member of the *orthomyxoviridae* family with a negative-sense single-stranded RNA. Its genome consists of eight different segments, ranging in size from 890 to 2,341 bases [8]. Among the proteins encoded by these segments, two glycoproteins, such as hemagglutinin (HA; segment 4) and neuraminidase (NA; segment 6), play important roles in the immune response. Different subtypes of HA (H1 to H15) and NA (N1 to N9) are found in avian species. The HAs of human viruses bind to the terminal sialic acid of glycoprotein and glycolipid receptors containing α2,6 linkages to galactose, whereas HAs of avian isolates prefer α2,3 linkages. NA is not considered to be as important an antigenic determinant as HA, but NA, like HA, is regarded as a possible restrictive factor for viral growth. The major function of NA is to remove sialic acid from the HA and NA of progeny virus particles, intercellular glycoproteins, and host cell receptors, thus facilitating virus release from infected cells and from intercellular inhibitors [9–11]. In addition to HA and NA proteins, the three viral polymerase proteins, such as PB1 (polymerase basic protein1), PB2 (polymerase basic protein2), and PA (polymerase) encoded on segment 1, 2, and 3 form an enzyme complex that functions in both transcription and replication. To accomplish the replication process, virus requires that newly synthesized NP (nucleoprotein), which encoded on segment 5, to bind to either negative-strand or positive-strand RNAs that are to be used as template for full-length copying [8]. In this study, we analyzed the HA and NA genes which determine the host specificity, and two genes which form the enzyme complex for transcription and replication, including PA and PB2, as well as NP gene from the four subtypes of *influenza A virus* which has an important role in the replication process. PB1 gene was not used in this study because there were little or no full-length genomes available for four subtypes of *influenza A virus*es which infected human population.

Analyses of codon usage patterns have been used by some scientists to determine the origins of species in many fields. Synonymous codons are usually known to encode common amino acids for protein synthesis. These codons are not used randomly, but rather some codons are used more frequently than others [12–14]. In prokaryotes, such as thermophilic bacteria, highly expressed genes have codon usage patterns that are shifted toward a more restricted set of ''preferred'' synonymous codons compared to the codon usage of other less highly expressed genes within a genome [15], and codon usage patterns have a tendency to mirror the distribution of tRNA abundances [16, 17]. With respect to virus genomes, Gu et al. [18] reported that the relative synonymous codon usage values in Severe Acute Respiratory Syndrome (SARS) coronaviruses were virus-specific and that translational selection and gene length might have no effect on the codon usage pattern in these viruses [19]. Almost all codon usage studies using the correspondence analysis method, however, were conducted in bacteria or higher species and not in virus species.

We examined the differences in codon usage patterns among the species included in the *orthomyxoviridae* family first, and then, we also analyzed the differences among the H1N1, H2N2, H3N2, and H5N1 subtypes of the *influenza A virus* group which were isolated from human population. To investigate the codon usage patterns, the relative synonymous codon usage (RSCU) and the effective number of codons (ENC) for all species were calculated. The correspondence analysis was used to analyze the differences statistically.

## Materials and methods

### *Nucleotide sequences*

To investigate the overall synonymous codon usage patterns, 13,496 coding sequences (CDSs) from the *orthomyxoviridae* family, including *influenza A virus*, *influenza B virus*, *influenza C virus*, *thogoto virus*, and *dhori virus* species, were examined. Using Java codes, all the complete CDSs of these five species were extracted from GenBank flat files (version 151) in NCBI ftp site (ftp://ftp.ncbi.nih.gov/genbank), and pooled by their species names. Then, these data sets were used in the RSCU (relative synonymous codon usage) calculation and the statistical analysis for each species.

The full-length genes of HA (hemagglutinin), NA (neuraminidase), NP (nucleoprotein), PA (polymerase), and PB2 (polymerase basic protein 2) from the H1N1, H2N2, H3N2, and H5N1 subtypes of the *influenza A virus* which infected human population were also collected from 'NCBI Influenza Virus Resources (http://www.ncbi.nlm.nih.gov/genomes/FLU /Database /select.cgi?go = 1)' to analyze the differences among subtypes. Because there were little or no full-length genomes of PB1 (polymerase basic protein 1) which infected human population, we excluded PB1 gene in this study. Among the sequences for each gene from four influenza subtypes, only the most recently registered and complete sequences which were isolated from Asian, European, or African countries were used (Table 1). In the case of H5N1 subtype, we collected all the possible sequences, which were registered in 2005 to compare more intensively with other subtype viruses. Partial CDSs, complementary CDSs, and CDSs with more than one ambiguous sequence were excluded.

**Table 1.** The list of the region, subtype, length, country, year and accession numbers of gene sequences including hemagglutinin (HA), neuraminidase (NA), nucleoprotein (NP), polymerase (PA), and polymerase basic protein 2 (PB2) genes from H1N1, H2N2, H3N2, and H5N1 subtypes of *influenza A virus*

| Gene | Region | Subtype | Length | Country | Year | Accession # |
|---|---|---|---|---|---|---|
| HA | Asia | H1N1 | 1698 | Taiwan | 2002 | DQ249260 |
| | | H2N2 | 1689 | South Korea | 1968 | L11133 |
| | | H3N2 | 1701 | Taiwan | 2004 | DQ249261 |
| | | H5N1 | 1704 | China | 2005 | DQ371928 |
| | | H5N1 | 1704 | Viet Nam | 2005 | AB239125 |
| | Europe | H1N1 | 1701 | Switzerland | 1995 | AF386773 |
| | | H2N2 | 1689 | Germany | 1964 | L11126 |
| | | H3N2 | 1701 | Denmark | 2003 | AY531039 |
| NA | Asia | H1N1 | 1413 | South Korea | 2002 | AY297140 |
| | | H2N2 | 1410 | Taiwan | 1967 | AY209925 |
| | | H3N2 | 1410 | Taiwan | 2004 | DQ249255 |
| | | H5N1 | 1350 | Viet Nam | 2005 | AB239126 |
| | | H5N1 | 1350 | Thailand | 2005 | DQ360836 |
| | Europe | H1N1 | 1410 | Finland | 2002 | AJ518100 |
| | | H2N2 | 1410 | United Kingdom | 1967 | AY209924 |
| | | H3N2 | 1410 | Denmark | 2003 | AY531006 |
| | Africa | H1N1 | 1413 | South Africa | 1997 | AJ518096 |
| | | H2N2 | 1410 | South Africa | 1967 | AY209930 |
| | | H3N2 | 1410 | South Africa | 1998 | AJ457940 |
| NP | Asia | H1N1 | 1497 | Hong Kong | 1998 | AF258516 |
| | | H2N2 | 1497 | South Korea | 1968 | AY210103 |
| | | H3N2 | 1497 | Hong Kong | 1999 | AJ293924 |
| | | H5N1 | 1497 | Viet Nam | 2005 | DQ099780 |
| | | H5N1 | 1497 | Thailand | 2005 | DQ360840 |
| | Europe | H1N1 | 1497 | Ukraine | 1979 | X51972 |
| | | H2N2 | 1497 | United Kingdom | 1967 | AY210098 |
| | | H3N2 | 1497 | Switzerland | 1999 | AJ458276 |
| | Africa | H2N2 | 1497 | South Africa | 1967 | AY210094 |
| PA | Asia | H1N1 | 2151 | Hong Kong | 1998 | AF258519 |
| | | H2N2 | 2151 | South Korea | 1968 | M26079 |
| | | H3N2 | 2151 | Hong Kong | 1999 | AJ293922 |
| | | H5N1 | 2151 | Viet Nam | 2005 | DQ138184 |
| | | H5N1 | 2151 | Thailand | 2005 | DQ360839 |
| | Europe | H1N1 | 2151 | Russia | 1977 | CY009289 |
| | | H2N2 | 2151 | United Kingdom | 1967 | AY210004 |
| | | H3N2 | 2151 | France | 1997 | AF483603 |
| PB2 | Asia | H1N1 | 2280 | Hong Kong | 1998 | AF258525 |
| | | H2N2 | 2280 | South Korea | 1968 | M73524 |
| | | H3N2 | 2280 | Hong Kong | 1999 | AJ293920 |
| | | H5N1 | 2280 | Viet Nam | 2005 | DQ138181 |
| | | H5N1 | 2283 | Thailand | 2005 | DQ372598 |
| | Europe | H1N1 | 2280 | Ukraine | 1979 | M38277 |
| | | H2N2 | 2280 | United Kingdom | 1967 | AY209948 |
| | | H3N2 | 2280 | France | 1997 | AF483602 |

*Relative synonymous codon usage (RSCU)*

In a correspondence analysis, the RSCU value for each codon is usually used to prevent the amino acid composition from influencing the codon usage values for each gene [20]. The RSCU value is the number of times a particular codon is observed, relative to the number of times the codon would be observed in the absence of any codon usage bias. If there were no codon usage bias, the RSCU value would be 1.00. The RSCU was calculated as

$$\mathrm{RSCU}_{ij} = \frac{X_{ij}}{\frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}}$$

where $X_{ij}$ is the frequency of occurrence of the $j$th codon for the $i$th amino acid, and $n_i$ is the number of codons for the $i$th amino acid. For the correspondence analysis, each gene was represented as a 59-dimensional vector excluding start and stop codons and the UGG codon, which codes tryptophan without any synonymous codons. All 59 codons were then pooled and calculated in a contingency table

according to the species genome in which they were included. All the calculated RSCU values of the species from the *orthomyxoviridae* family, and of the H1N1, H2N2, H3N2, and H5N1 human subtypes from *influenza A viruses* were used to the further statistical analysis, such as correspondence analysis.

*Effective number of codons (ENC)*

In addition to the RSCU values per each amino acid, we also calculated the ENC value, which is the most common measure of codon bias [21]. ENC values range from 20, for the case in which only one codon is used for each amino acid, to 61, for the case in which all synonymous codons are used in equal frequency. ENC can be estimated as

$$\mathrm{ENC} = 2 + \frac{9}{\bar{F}_2} + \frac{1}{\bar{F}_3} + \frac{5}{\bar{F}_4} + \frac{3}{\bar{F}_6},$$

where $\bar{F}_k$ ($k = 2$, 3, 4, or 6) is the average of the $F_k$ values for $k$-fold degenerate amino acids. $F_k$ for each of the $k$-fold degenerate amino acids was estimated as

$$F_k = \frac{nS - 1}{n - 1},$$

where $n$ is the total number of codons for that amino acid, and

$$S = \sum_{i=1}^{k} \left(\frac{n_i}{n}\right)^2,$$

where $n_i$ is the number of occurrences of the $i$th codon for that amino acid [14]. All the calculated ENC values of the H1N1, H2N2, H3N2, and H5N1 human subtypes from *influenza A viruses* were divided into each gene and country group for more comparisons (Figure 3). All calculations of codon usage numbers, RSCU, and ENC values were performed using JAVA codes that we developed in the RedHat Linux 9.0 platform.

*Correspondence analysis*

Correspondence analysis is a type of multivariate analysis that represents associations as a table of frequencies or graphically as counts. For a contingency table with $I$ rows and $J$ columns, the plot produced by correspondence analysis would contain two sets of points: one set of $I$ points corresponding to the rows, and one set of $J$ points corresponding to the columns. The positions of the points would reflect the associations. The RSCU values for 59 codons, as described above, were used, and the correspondence analysis process was performed using the SAS statistical program, version 9.1 [22].

The usual output from a correspondence analysis includes the ''best'' two-dimensional representation of the data, along with the coordinates of the plotted points, and a measure (called the inertia) of the amount of information retained in each dimension [23]. The results provide information similar to that produced by factor analysis techniques and allow one to explore the structure of categorical variables included in the table. Genes or species that are strongly associated as measured by their chi-square distances will lie in a similar direction from the origin [23, 24]. The chi-square distance between two coordinates with the same row or column value, called the Euclidean distance [18, 24], has important statistical meaning, but there is no significant meaning between the row coordinate and the column coordinate. The SigmaPlot 8.0 program was used to create the graphical correspondence analysis plots using coordinates consisting of the first and second dimensional factors that resulted from the correspondence analysis. The correspondence analysis was performed using both the RSCU values of the species from the *orthomyxoviridae* family, and of the H1N1, H2N2, H3N2, and H5N1 human subtypes from *influenza A viruses* (Figures 2, 4).

*Other statistical analyses*

Linear regression analysis was also conducted to determine the correlation between the nucleotide content at the third codon position and the first axis (dim1) in the correspondence analysis results for the species from the *orthomyxoviridae* family. One-way ANOVA with Duncan's multiple range test was also used to compare the average values among the RSCU results. All of these analyses were performed using the SAS statistical software, version 9.1 [22].

## Results

*RSCU and correspondences analysis patterns of the orthomyxoviridae family*

To examine the synonymous codon usage patterns, we extracted all the complete CDSs from GenBank flat files and calculated the RSCU value for each species. Not only the *influenza A*, *B* and *C viruses*, but also the *thogoto* and *dhori virus* in the *orthomyxoviridae* family were analyzed to compare the overall codon usage patterns with influenza group. The result of the RSCU patterns from five species showed that all the viruses revealed the highest value at 'AGA' codon which encodes arginine (Figure 1). The *influenza B virus* resulted the highest bias in 'AGA' codon usage, and *influenza B virus*, *thogoto virus*, *influenza A*, and *dhori virus* were followed in descending order, and this RSCU pattern also continued in other synonymous codon groups. If there was no synonymous codon usage bias among the codons which encodes the same amino acid, the RSCU values for those codons would be 1.00. So, the higher RSCU values for each synonymous codon mean the higher codon

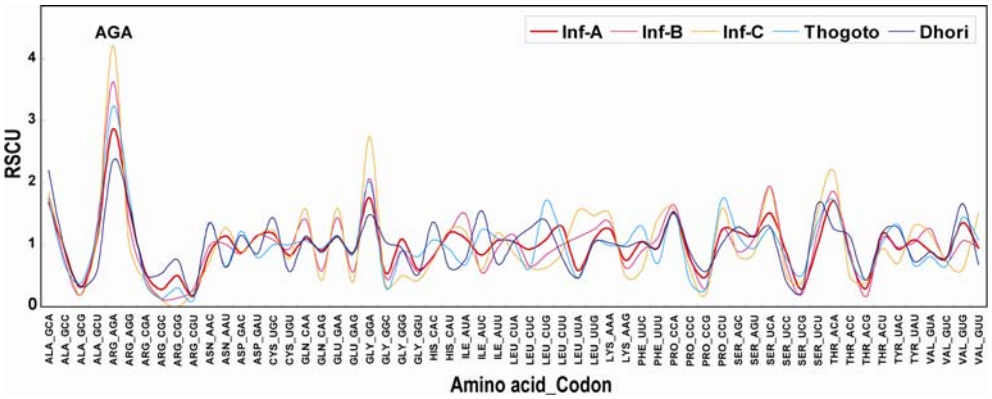


**Fig. 1** Relative synonymous codon usage (RSCU) values of viruses from the *orthomyxoviridae* family in 59 codons. All the legends and line colours are shown in the graph, and each amino acid and its codon nucleotides was shown in X-axis. The abbreviations above for each species are as follows: Inf-A, *influenza A virus*; Inf-B, *influenza B virus*; Inf-C, *influenza C virus*; DRV, *dhori virus*; TGV, *thogoto virus*.

usage bias in that amino acid. The *influenza B virus* also revealed the high synonymous codon usage bias in codon groups which encoded glutamine, glycine, and threonine. The *influenza A virus*, however, showed relatively the lowest codon bias among the three influenza virus group. To determine the bias for each species, we calculated the difference between the RSCU value of each species and the RSCU value of 1.00, and then averaged the differences for each virus group. As a result, the average difference from 1.00 for the *influenza A virus* group was 0.300, and the average differences for the *influenza B* and *influenza C virus* groups were 0.383 and 0.510, respectively. The difference for the *influenza A virus* group was statistically smaller than that for the *influenza C virus* group ($p < 0.05$). *Thogoto* and *dhori virus* showed reversal RSCU patterns codon groups for histidine and isoleucine.

In the next step, we also performed a correspondence analysis of *orthomyxoviridae* species using the RSCU values (Figure 2). The first-dimensional factor (dim1) in the correspondence analysis was significantly correlated with the base composition, especially with the GC content at the third codon position, showing the highest $R^2$ value (0.989; $p < 0.0005$) in the linear regression test (Table 2). Among the influenza virus groups, the *influenza A virus*, with low synonymous codon bias, was located on the opposite side from the *influenza B* and *influenza C virus*. *Thogoto* and *dhori viruses* were also located on the same side with the *influenza A virus*, but they were separately located from the *influenza A virus* on the basis of the dim2-axis, with showing the higher synonymous codon usage bias than the *influenza A viruses* group.

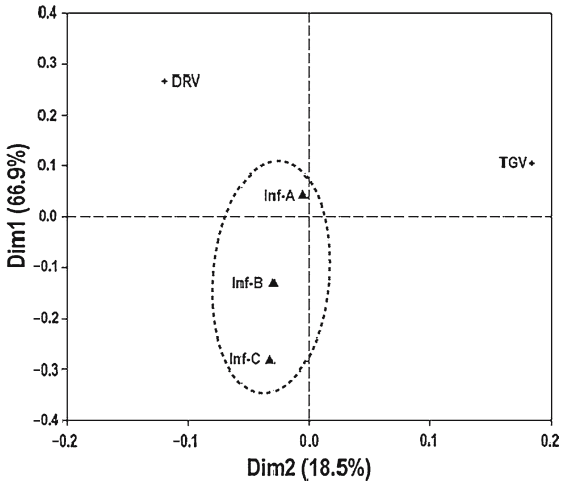


**Fig. 2** A plot of the values of the first and second axes of the *orthomyxoviridae* family. Dim1 and dim2 represent the values of the first and the second dimensional factors of each species, and the percentage in each parenthesis means the percent inertia of that axis in correspondence analysis. All the influenza viruses are clustered with blue-dotted line. The abbreviations above for each species are as follows: Inf-A, *influenza A virus*; Inf-B, *influenza B virus*; Inf-C, *influenza C virus*; DRV, *dhori virus*; TGV, *thogoto virus*.

### *ENC patterns among influenza A virus subtypes*

The nucleotide sequences of the HA, NA, NP, PA, and PB2 genes from four virus subtypes (H1N1, H2N2, H3N2, and H5N1), which were isolated from human population were used to calculate the ENC values for determining the codon usage bias (Figure 3). The detailed information for those sequences, including length, isolated region, year, and GenBank accession numbers for each virus subtype, are shown in Table 1. In HA gene, H1N1 subtypes from both Asia and Europe showed the lowest ENC values, and H3N2 subtypes revealed the highest score. ENC represents the effective number of codons in each gene, ranging between 20 and 61. So the ENC values greater than 20 means that more than one codon is used for each amino acid, and if the ENC value is 61, then it means that all the synonymous

**Table 2.** $R^2$ values and significance levels of linear regression tests between the first axis in correspondence analysis and the nucleotide composition on the third codon position of the *orthomyxoviridae* family

| Base composition | $R^2$ [a] | Pr > $F$ |
|---|---|---|
| $G_{3S}+C_{3S}$ [b] | 0.989 (+)[c] | 0.0005 |
| $A_{3S}$ | 0.932 (−) | 0.008 |
| $G_{3S}$ | 0.900 (+) | 0.014 |
| $C_{3S}$ | 0.969 (+) | 0.002 |
| $T_{3S}$ | 0.786 (−) | 0.045 |

[a] $R^2$ value of each linear regression analysis.
[b] Base composition on 3rd synonymous position.
[c] (+) means positive correlation, and (−) means negative correlation in linear regression test.

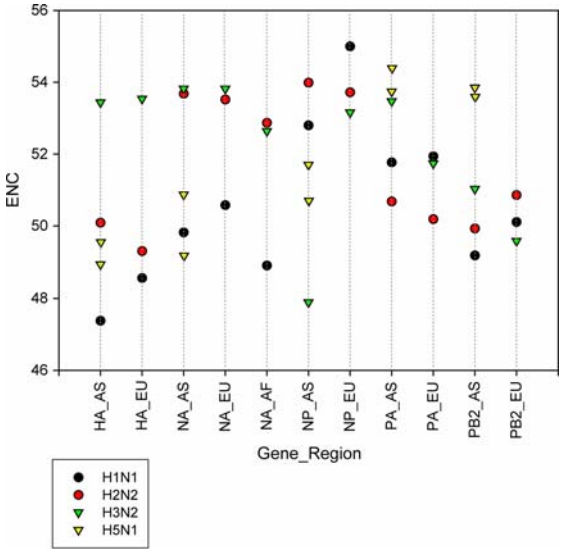


**Fig. 3** Effective number of codons (ENC) from hemagglutinin (HA), neuraminidase (NA), nucleoprotein (NP), polymerase (PA), and polymerase basic protein 2 (PB2) genes among H1N1, H2N2, H3N2, and H5N1 subtypes of *influenza A virus* isolated from human population. Each gene and isolated region name is shown in X-axis, and legends are shown at the bottom of the graph. The GenBank accession numbers for each species are shown in Table1. The abbreviations above for each isolated regions are as follows: AS, Asia; EU, Europe; AF, Africa.

codon are used in equal frequency. HA genes represented the differences among four different HA subtypes, including H1, H2, H3, and H5, but H1N1 subtype which was isolated from Switzerland in 1999 showed higher ENC value, with showing similar value with H2N2 subtype which was isolated from Germany in 1964. In NA gene, the ENC values were divided into two groups by their NA subtypes, such as N1 and N2, and NA gene which was isolated from South Africa in 1997 showed the lowest ENC value.

In NP gene, H3N2 subtypes which were isolated from Asian country (A/HK/1774/1999) showed the dramatic differences from that from European country, with showing the lowest ENC value, 47.8. Moreover, there were some differences in the order of ENC values between H3N2 subtypes from Asia (A/HK/1774/1999) and from Europe (A/Switzerland/9243/1999) in PA and PB2 genes when other subtypes of Europe maintained the order of ENC values like those of Asian countries. This kind of difference was also continued in the following synonymous codon usage analysis.

### *Correspondence analysis (CA) among influenza A virus subtypes*

Using the RSCU values for each gene in the four *influenza A virus* subtypes, we also performed a correspondence analysis to determine the codon usage patterns among the H1N1, H2N2, H3N2, and H5N1 subtypes (Figure 4). All these sequences were isolated from human population. For the HA gene, all the subtypes were distributed in the four quadrants of the correspondence plot, except for the H1N1 subtype isolated from Asia in 2002. But both two H1N1 subtype genes showed the similar low values on the basis of the dim1. This was reasonable because each subtype had different HA types, such as H1, H2, H3, and H5, respectively (Figure 4A). CA result of NA gene presented that subtypes which have the same type of NA gene, such as N1 and N2, were grouped together on the basis of the dim 1 (Figure 4B). Unlike the H2N2 and H3N2 subtypes, however, the H5N1 and H1N1 subtypes were located on the different quadrant along with the dim 2. In both HA and NA genes, there were no significant differences among the isolated regions, including Asia, Europe and Africa. In NP, PA and PB2 genes, H3N2 subtype showed different patterns between Asian and European genes, and these results were also appeared in the previous ENC analysis. Except for the H3N2 subtype genes isolated from Asian countries, H1N1, H2N2, and H3N2 subtypes were located on the opposite side from H5N1 subtype in NP, PA and PB2 genes (Figures 4C, D, E).

## Discussion

Huge amounts of genomic data have been collected and distributed on the Web by organizations such as the U.S. National Center for Biotechnology Information (NCBI) and the European Bioinformatics Institute (EBI). The NCBI regularly distributes genomic data sets as GenBank flat files via their ftp server. We collected the most recent version (version 151) of the GenBank flat files from the NCBI and extracted target sequences from the files by using Java codes. This study was designed to perform genomic analysis in two steps. The overall RSCU patterns analysis of the five virus species in the

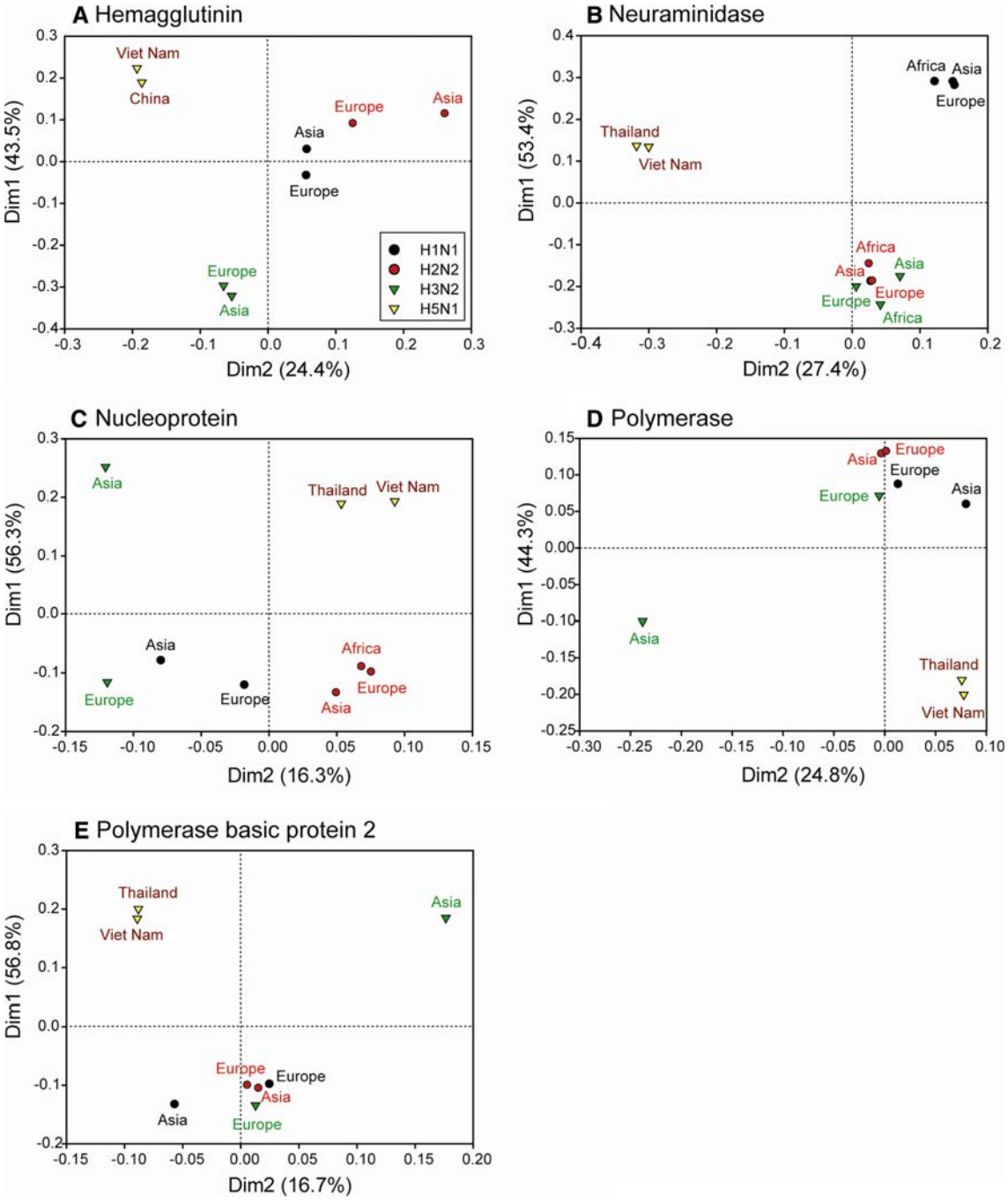


**Fig. 4** Plots of the values of the first and second axes of the hemagglutinin, neuraminidase, nucleoprotein, polymerase, and polymerase basic protein 2 genes in H1N1, H2N2, H3N2, and H5N1 subtypes of *influenza A virus*. Dim1 and dim2 represent the values of the first and the second dimensional factors of each species, and the percentage in each parenthesis means the percent inertia of that axis in correspondence analysis. The GenBank accession numbers for each virus subtype are shown in Table 1.

*orthomyxoviridae* family was conducted first, and then, four subtypes of *influenza A virus* which were isolated from human population were analyzed to determine the synonymous codon usage patterns more intensively.

Virus species in the *orthomyxoviridae* family were classified very well by their synonymous codon usage patterns (Figure 2). The *influenza A virus* group of viruses showed the lowest codon usage bias in the correspondence analysis. Pandemic diseases caused by RNA viruses, including *SARS coronavirus* and avian *influenza A virus*, have occurred worldwide during the 21st century. These two pandemic viruses are believed to have originated through the transfer of nonhuman virus species to humans. According to Subbarao and Shaw [2] and Puthavathana et al. [1], pandemic avian flu viruses appear to have crossed the species barrier, moving from avian hosts to humans.

After successfully transferring to the new host, the viruses were able to temporarily bypass the host immune response because the new host had no information about the new virus species. Among the virus species in the *orthomyxoviridae* family, the *influenza A virus* group revealed the lowest synonymous codon usage bias in our result, indicating that compared to other viruses in the *orthomyxoviridae* family, *influenza A viruses* are better able to adapt to new hosts with different synonymous codon usage environments. Many previous studies have reported that every bacteria and higher organism has its own synonymous codon usage pattern [12–15]. According to a previous study, the *influenza B* and *influenza C virus* groups do not usually infect nonhuman hosts, except for swine species, and also do not exhibit antigenic shift [8].

Among the *influenza A virus* envelope proteins, the HA and NA glycoproteins evoke protective immune responses in infected persons. *Influenza A viruses* usually exhibit antigenic drift by changing their HA and NA types in a population, thereby disturbing the host immune system. Although the most subtypes of *influenza A viruses* cause minor illness in human beings, they also can lead to severe cases of pneumonia or death, as was the case in the 1918 pandemic of the Spanish flu H1N1 subtype [25]. Antigenic shifts in the *influenza A virus* population have produced new subtypes associated with influenza epidemics or pandemics, such as those that occurred in 1918 (Spanish flu), 1957 (Asian flu), and 1968 (Hong Kong flu). The antigenic shift that occurred in 1957 involved changes in the structures of both HA and NA (H1N1 $\rightarrow$ H2N2), that in 1968 involved only one (H2N2 $\rightarrow$ H3N2) [8]. Therefore, we chose to compare the synonymous codon usage patterns in these four subtypes, including H1N1, H2N2, N3N2, and H5N1 subtypes, in this study.

The ENC values on both HA and NA genes represented their HA and NA types well, and this result reveals that each *influenza A virus* subtype uses different in codon usage patterns as well as the amino acid compositions when it synthesizes HA and NA glycoproteins (Figure 3). In NP, PA and PB2 genes, H3N2 subtypes which were isolated from Asia, especially Hong Kong in 1999 showed different patterns with same subtypes from Switzerland and France in 1999 and 1997, respectively. In HA and NA genes of H3N2 subtype from Hong Kong in 1999, however, did not showed this distinct pattern when they were compared with the H3N2 subtypes from European countries (data not shown). According to Gregory et al. [26] who sequenced this H3N2 subtype *influenza A virus* first, this virus (A/HK/1774/99) was isolated from 10-month-old girl, and it was antigenically related to early (1968–1975) human and swine H3N2 viruses. But unlike other H3N2 human viruses, it was clearly distinguishable from human H3N2 viruses isolated since 1979 and avian H3 viruses, and it was closely related in its genetic characteristics to H3N2 viruses prevalent in pigs in Europe during the 1990s. Pigs usually have the receptors for both human and avian *influenza A virus* HA glycoprotein, so they may act as an intermediate host in the emergence of novel human subtypes [27]. These distinguishable synonymous codon usage patterns of H3N2 subtype which was isolated from Hong Kong which were resulted in our study also proved this history very well (Figures 3, 4), and, in addition to the serological tests, this synonymous codon usage analysis might to be a useful method to identify the non-human origin of the virus in genomic levels. Actually, this H3N2 virus (A/HK/1774/99) also showed very similar codon usage patterns with H5N1 subtypes which are suspected to be transmitted from non-human species, avian species on the basis of the dim1 in CA plot (Figure 4C, D, E).

In conclusion, *influenza A virus* might have an advantage in transmitting from non-human origin to the human beings because they resulted to have the lowest synonymous codon usage bias than that of other species in the *orthomyxoviridae* family. The CA result of the four subtypes of *influenza A virus* group showed that the viruses which were transmitted from the non-human origin to human population use the synonymous codons in a different manner when compared to those of other human-origin viruses, so this distinguishable synonymous codon usage patterns in non-human origin viruses might be useful in determining the origin of *influenza A viruses* in genomic levels as well as the serological tests.

Thanks to the public genome databases, such as NCBI, EBI, and DDBJ, it became possible to analyze the genetic patterns among each country. In addition to the classical epidemiological method, genetic epidemiology using bioinformatics techniques has become more and more important these days. But most of the available data sets were mainly provided from Asia, Europe, and America except for Africa. According to the European Union authorities, Europe will be the next target for H5N1 avian flu in a near future, and the foothold for that pandemic would be Africa [4]. So, it will be necessary to pay more attention to these African countries for protecting Europe from the pandemic diseases, such as avian flu or SARS (severe acute respiratory syndrome), which initially occurred in Asia.

## Acknowledgments

We acknowledge the contribution of all the experimentalists who made their invaluable data publicly available. This work was supported by the Brain Korea 21 Project in 2006.

## References


1. Puthavathana P, Auewarakul P, Charoenying PC, et al. Molecular characterization of the complete genome of human influenza H5N1 virus isolates from Thailand. J General Virol 2005; 86: 423–433.
2. Subbarao K, Shaw MW. Molecular aspects of avian influenza (H5N1) viruses isolated from humans. Rev Med Virol 2000; 10: 337–348.
3. CDC News. 2005. Information about Avian Influenza (Bird Flu) and Avian Influenza A (H5N1) virus. Available at http://www.cdc.gov/flu/avian/gen-info/facts.htm.
4. Enserink M. H5N1 moves into Africa, European Union, deepening global crisis. Science 2006; 311: 932.
5. Lin YP, Shaw M, Gregory V, et al. Avian-to-human transmission of H9N2 subtype influenza A viruses: Relationship between H9N2 and H5N1 human isolates. Proc Natl Acad Sci 2000; 97: 9654–9658.
6. Dimmock, NJ, Easton, AJ, Leppard KN 2002. Introduction to modern virology. pp. 297–311. Blackwell publishing, 350 Main Street, Malden, MA 02148–5018, USA.
7. Liu JP. Avian influenza – a pandemic waiting to happen?. J Microbiol Immunol Infect 2006; 39: 4–10.
8. Voyles BJ 2002, The Biology of Viruses. 2nd edn. McGraw-Hill Companies, Inc. 1221 Avenue of the Americas, New York, USA, pp. 147–149, 338–341.
9. Matrosovich M, Zhou N, Kawaoka Y, Webster R. The surface glycoproteins of H5 influenza viruses isolated humans, chickens, and wild aquatic birds have distinguishable properties. J Virol 1999; 73: 1146–1155.
10. Plotkin JB, Dushoff J. Codon bias and frequency-dependent selection on the hemagglutinin epitopes of influenza A virus. Proc Natl Acad Sci 2003; 100: 7152–7157.
11. Rogers GN, Paulson JC. Receptor determinants of human and animal influenza virus isolates: differences in receptor specificity of the H3 hemagglutinin based on species of origin. Virology 1983; 127: 361–373.
12. Duret L. Evolution of synonymous codon usage in metazoans. Curr Opin Gene Dev 2002; 12: 640–649.
13. Mclnerney JO. Codon. Replicational and transcriptional selection on codon usage in Borrelia burgdorferi. Proc Natl Acad Sci 1998; 95: 10698–10703.
14. Moriyama EN, Hartl DL. Codon usage bias and base composition of nuclear genes in Drosophila. Genetics. 1993; 134: 847–858.
15. Lynn DJ, Singer GAC, Hickey DA. Synonymous codon usage in subject to selection in thermophilic bacteria. Nucleic Acids Res 2002; 30: 4272–4277.
16. Shields DC, Sharp PM. Synonymous codon usage in Bacillus stbtilis reflects both translational selection and mutational biases. Nucleic Acids Res 1987; 15: 8023–8040.
17. Stenico M, Lloyd AT, Sharp PM. Codon usage in Caenorhabditis elegans: delineation of translational selection and mutational biases. Nucleic Acids Res 1994; 22: 2437–2446.
18. Gu W, Zhou T, Ma J, Sun X, Lu Z. Analysis of synonymous codon usage in SARS Coronavirus and other viruses in the Nidovirales. Virus Res 2004; 101: 155–161.
19. Jenkins GM, Holmes EC. The extent of codon usage bias in human RNA viruses and its evolutionary origin. Virus Res 2003; 92: 1–7.
20. Sharp PM, Li WH. Codon usage in regulatory genes in Escherichia coli does not reflect for 'rare' codons. Nucleic Acids Res 1986; 14: 7737–7749.
21. Wright F. The 'effective number of codons' used in a gene. Gene 1990; 87: 23–29.
22. Cary NC 2004. *SAS*®R9.1.2 Qualification Tools User's Guide. SAS Institute Inc. .
23. Johnson RA, Wichern DW Applied Multivariate Statistical Analysis. 5th edn. Prentice-Hall, Inc. 2002 .
24. Perrière G, Thioulouse J. Use and misuse of correspondence analysis in codon usage studies. Nucleic Acids Res 2002; 30: 4548–4555.
25. Edler AA. Avian flu (H5N1): its epidemiology, prevention, and implications for anaesthesiology. J Clin Anesth 2006; 18: 1–4.
26. Gregory V, Lim W, Cameron K, et al. Infection of a child in Hong Kong by influenza A h3n2 virus closely related to viruses circulating in European pigs. J Gen Virol 2001; 82: 1397–1406.
27. Eijk M, White MR, Batenburg JJ, et al. Interactions of influenza A virus with sialic acids present on porcine surfactant protein D. Am J Respir Cell Mol Biol 2004; 30: 871–879.


*Address for correspondence:* Hyeon S. Son, Laboratory of Computational Biology and Bioinformatics, Graduate School of Public Health, Seoul National University, 28, Yongon-Dong, Chongno-Gu, Seoul, 110-799, Korea
Phone: +82-2-740-8864; Fax: +82-2-762-9105
E-mail: hss2003@snu.ac.kr.