J. Microbiol. Biotechnol. (2010), **20**(1), 63–70
doi: 10.4014/jmb.0908.08004
First published online 24 October 2009

JOURNAL OF MICROBIOLOGY AND BIOTECHNOLOGY

Comparative Study of the Nucleotide Bias Between the Novel H1N1 and H5N1 Subtypes of Influenza A Viruses Using Bioinformatics Techniques

Ahn, Insung[1] and Hyeon Seok Son[2,3]*

[1]*Supercomputing Center, Korea Institute of Science and Technology Information, Daejon 305-806, Korea*
[2]*Laboratory of Computational Biology and Bioinformatics, Institute of Health and Environment, Graduate School of Public Health, Seoul National University, Seoul 110-799, Korea*
[3]*Interdisciplinary Graduate Program in Bioinformatics, College of Natural Science, Seoul National University, Seoul 151-742, Korea*

Received: August 4, 2009 / Revised: September 22, 2009 / Accepted: September 23, 2009

Novel influenza A (H1N1) is a newly emerged flu virus that was first detected in April 2009. Unlike the avian influenza (H5N1), this virus has been known to be able to spread from human to human directly. Although it is uncertain how severe this novel H1N1 virus will be in terms of human illness, the illness may be more widespread because most people will not have immunity to it. In this study, we compared the codon usage bias between the novel H1N1 influenza A viruses and other viruses such as H1N1 and H5N1 subtypes to investigate the genomic patterns of novel influenza A (H1N1). Totally, 1,675 nucleotide sequences of the hemagglutinin (HA) and neuraminidase (NA) genes *of* influenza A virus, including H1N1 and H5N1 subtypes occurring from 2004 to 2009, were used. As a result, we found that the novel H1N1 influenza A viruses showed the most close correlations with the swine-origin H1N1 subtypes than other H1N1 viruses, in the result from not only the analysis of nucleotide compositions, but also the phylogenetic analysis. Although the genetic sequences of novel H1N1 subtypes were not exactly the same as the other H1N1 subtypes, the HA and NA genes of novel H1N1s showed very similar codon usage patterns with other H1N1 subtypes, especially with the swine-origin H1N1 influenza A viruses. Our findings strongly suggested that those novel H1N1 viruses seemed to be originated from the swine-host H1N1 viruses in terms of the codon usage patterns.

Keywords: Influenza A virus, novel H1N1 subtype, H5N1 subtype, base composition, genomics

**Corresponding author*
Phone: +82-2-740-886; Fax: +82-2-762-9105;
E-mail: hss2003@snu.ac.kr

Since the World Health Organization declared a public health emergency of international concern, and raised the pandemic influenza phase from 4 to 5 in April 2009, about 69 countries have officially reported 21,940 cases of the novel influenza A (H1N1) infection, including 125 deaths [World Health Organization. 2009. http://www.who.int/csr/don/ 2009_06_05/ en /print.html]. This novel H1N1 subtype is known to contain genes closely related to swine influenza [Trifonov *et al.* 2009. http://www.eurosurveillance.org/images/dynamic/EE/V14N17/art19193.pdf]. The hemagglutinin (HA) gene of this novel virus is known to be similar to that of the swine influenza virus isolated from the U.S.A., whereas the neuraminidase (NA) gene was similar to that of swine influenza viruses isolated from Europe [M.S.Bronze. 2009. http://emedicine.medscape.com/article/1673658-print]. However, the exact origin of this new strain is not yet determined. Historically, H1N1 subtypes of influenza A viruses were responsible for the global flu pandemics in 1918, which killed about one-third of the world's population with >2.5% fatality rates [8, 19]. After the first global pandemic, epidemiologically severe outbreaks occurred periodically from 1928 to 1957 with severe H1N1 epidemic occurring in 1950–1951 [13]. In 1977, H1N1 subtypes re-emerged, and then have been co-circulated with H3N2 subtypes until now [6].

Each animal or plant species has been known to have its own nucleotide bias, especially synonymous codon usage patterns [11, 16, 17, 21]. In thermophilic bacteria, for example, the highly expressed genes were known to shift their codon usage toward a more restricted set of preferred synonymous codons, compared with less highly expressed genes within the genome, indicating that the codon usage bias could mirror tRNA abundance [11, 16]. According to Zhou and Li [21], the codon usage patterns of mitochondrial genes were more conserved in GC content, and there was no correlation between GC12 and GC3. T and A ending

codons were detected as the preferred codons in plant mitochondrial genomes. In our previous study, synonymous codon usage patterns among RNA viruses such as influenza A viruses and HIV-1s were divided into each region, subtype, host, or occurring-year group, with an expectation that there might be some correlations between the nucleotide patterns and the direction of viral variations on the codon basis [1–3]. In this study, we compared the codon usage bias between the novel H1N1 influenza A viruses and other viruses such as H1N1 and H5N1 subtypes, to investigate the genomic patterns of novel influenza A (H1N1). All the target sequences were divided into different groups according to their subtypes and host species, such as human, avian, and swine, for the intensive analysis.

Materials and Methods

Nucleotide Sequences

The hemagglutinin and neuraminidase genes of influenza A virus (H1N1 and H5N1 subtypes) that had occurred from 2004 to 2009 were collected from the Influenza Virus Resource at the National Center for Biotechnology Information [4]. In the phylogenetic analysis in Fig. 1, the nucleotides of the HA (25 sequences) and NA (21 sequences) genes of the swine-origin H1N1 subtypes were used to investigate the evolutionary correlations with the HA (53 sequences) and NA (47 sequences) genes of the novel human-host influenza A viruses (H1N1). To compare the nucleotide patterns of the novel H1N1 viruses with others, we used 887 and 788 nucleotide sequences of HA and NA genes, respectively, of both H1N1 and H5N1 influenza A viruses isolated from various host species. All the GenBank accession numbers of those sequences are shown in Appendix I and II. Downloaded FASTA format sequences were parsed into each category such as accession number, subtype, gene, host, occurring year, and other parameters using JAVA codes, and the MySQL database management system was used to construct all the local databases on the Linux operating system. Only complete coding sequences without abnormal sequences that include unknown characters, except for A, G, C, or U, or not started with the "AUG" codon were selected for this study.

Phylogenetic Analysis

In order to determine the phylogenetic correlations among viral genes, multiple sequence alignments using the ClustalW ver. 1.83 [10] were performed. The IUB matrix was used for the DNA weight matrix, and the gap opening and extension penalty were 40.0 and 0.8, respectively. Phylograms for swine-origin and novel human-origin influenza A viruses (H1N1) were created by the PAUP* ver. 4.0 program [18] using the neighbor-joining method with 1,000 times bootstrapping process. All the analyzing processes were conducted on both the Linux and Windows operating systems.

Indices of Nucleotide Bias

To examine the nucleotide bias among each subtype and host group, the % guanine and cytosine contents on the first (1st GC%), second (2nd GC%), and third (3rd GC%) codon positions, as well as the effective number of codons (ENC), were calculated from all the target sequences. The ENC values range from 20, for the case in which only one codon is used for each amino acid, to 61, when all synonymous codons are used in equal frequency. ENC can be calculated as given below:

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

where $\overline{F}_k$ (k=2, 3, 4, or 6) is the average of the F_k values for *k*-fold degenerate amino acids [12]. The F value in this equation means the average homozygosity or probability of two randomly chosen codons for each amino acid. Thus, if all synonymous codons are used in equal frequency, ENC will be 61, whereas the opposite extreme condition would be if a single codon is used for each amino acid, yielding an ENC value of 20 [20]. 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 each 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 each amino acid. All the calculations were performed using JAVA codes on Linux.

Relative Synonymous Codon Usage (RSCU)

Synonymous codon usage patterns of target genes were examined by calculating the RSCU values. Because RSCU is the observed frequency of a codon in the gene divided by the frequency expected if all the synonymous codons are used randomly, it is known to minimize the bias from the amino acid composition [15]. The RSCU is calculated as

$$RSCU_{ij} = \frac{obs_{ij}}{\frac{1}{n_i}\sum_{j=1}^{n_i} obs_{ij}}$$

where obs_{ij} is the observed number 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 described below, each sequence was represented as a 59-dimensional vector excluding start and stop codons as well as the UGG codon, which produces a tryptophan without any other synonymous codons.

Correspondence Analysis (CA)

The CA method, a kind of multivariate analysis, was used to analyze the major trend in terms of synonymous codon usage variation among target genes. CA is a compositional method, because the perceptual map is based on the association between objects and a set of descriptive characteristics [7]. In this study, we assigned the subtypes of influenza A viruses, which were grouped on the basis of their host species as objects, and the RSCU values of target genes as descriptive characteristics factors. The biplot graph of the CA results includes the major two-dimensional representations of the data, and a measure of the amount of information retained in each dimension. Thus, nucleotide sequences that are strongly associated on the basis of RSCU values will be plotted in a similar direction from the origin [9, 14], and the distance between two coordinates within the same row or column represents the x^2 distance. JAVA codes were used in creating

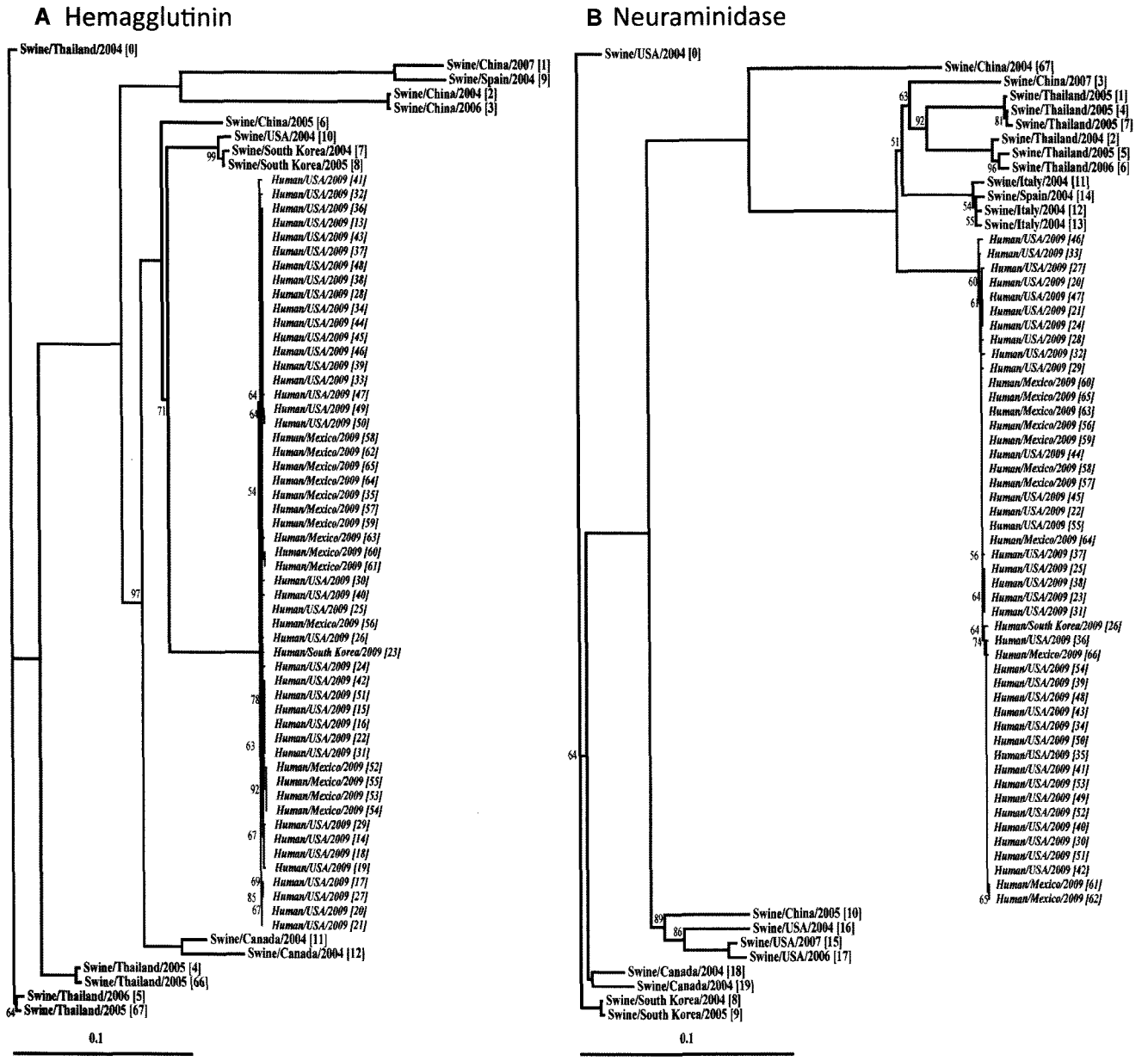

Fig. 1. Phylograms of the (**A**) hemagglutinin and (**B**) neuraminidase genes of swine-origin and novel human-origin influenza A virus (H1N1).

Each host, country, and occurring-year information is shown as taxon name with its data number presented as a parenthesized number. Novel human-origin influenza A viruses (H1N1) are shown in italic characters, and the GenBank accession number of each taxon is provided in Appendix I. Trees were derived by the neighbor-joining method using the PAUP* 4.0 program with bootstrapping analysis of 1,000 iterations. The branch lengths are drawn to scale, with the bar indicating 0.1 nucleotide replacements per site. Bootstrap values (%) that are not 100% are represented in each node.

the input files for the SAS statistical program [5] using the nucleotide sequences extracted from the MySQL database we constructed.

RESULTS

Phylogenetic Analysis

In order to determine whether the novel 68 H1N1 strains that were registered to GenBank before 14 May 2009 were related to other swine-origin H1N1 subtypes isolated from 2004, we performed phylogenetic analysis of HA and NA genes using the neighbor-joining method (Fig. 1). Among the novel H1N1 subtypes, there were less or no significant differences, indicating that the novel viruses that occurred in the U.S.A., Mexico, and Korea were spread from a common strain. In each gene, HA of the novel H1N1 subtypes showed relatively close correlations with swine-origin H1N1 viruses isolated from U.S.A. in 2004 (EU139832),

and Korea in 2004 and 2005 (EU798778, EU798779), whereas NA sequences were from Italy (EU045393, EU045388, EU045389) and from Spain (CY010582) in 2004.

Nucleotide Composition

The overall average of guanine and cytosine contents (%G+C contents) on the first (1st GC%), second (2nd GC%), and third (3rd GC%) codon positions of the target genes were investigated in both H1N1 and H5N1 subtype influenza A viruses (Fig. 2). We divided all the sequences into each subtype and host group to determine whether there were any specific patterns among the % G+C contents on three codon positions. In the HA gene, the novel H1N1 viruses showed different nucleotide compositions from other groups, with very similar %G+C contents resulting on the 2nd and 3rd codon positions (Fig. 2A). In the other groups, the 3rd GC% revealed higher values than the 2nd GC%, showing similar values with the 1st GC% except for the H1N1_Swine group. In NA genes, however, the %G+C distributions among three codon positions in all groups showed very similar patterns, and the overall values of %G+C contents of H5N1 subtypes were slightly higher than those of H1N1 strains (Fig. 2B).

Correspondence Analysis Using the RSCU Patterns

To investigate the nucleotide patterns of HA and NA genes among all the target H1N1 and H5N1 influenza A viruses, we calculated the RSCU values of each target sequence. The biplot graph of the CA results, which includes the major two-dimensional representations of the data, is shown in Fig. 3. Linear regression tests between each dimension of the CA result and other parameters of nucleotide bias such as the %G+C contents on each codon position as well as the ENC values were also performed (Table 1). As a result, dim1 of the CA results for both the HA and NA genes showed the highest correlations with %G+C contents on the 2nd codon position, with resulting R^2 values of 0.777 and 0.765, respectively, indicating that it represents the amino acid differences among viruses (Table 1). Dim2 of the HA and NA genes, however, showed significantly high R^2 values in %G+C contents on the 3rd codon position (0.374) and ENC (0.498), respectively, indicating that it represents the codon differences within each synonymous codon group.

In Fig. 3, all the target sequences were divided into each viral group according to their subtypes and host species. Dotted circle areas filled with green color represent the

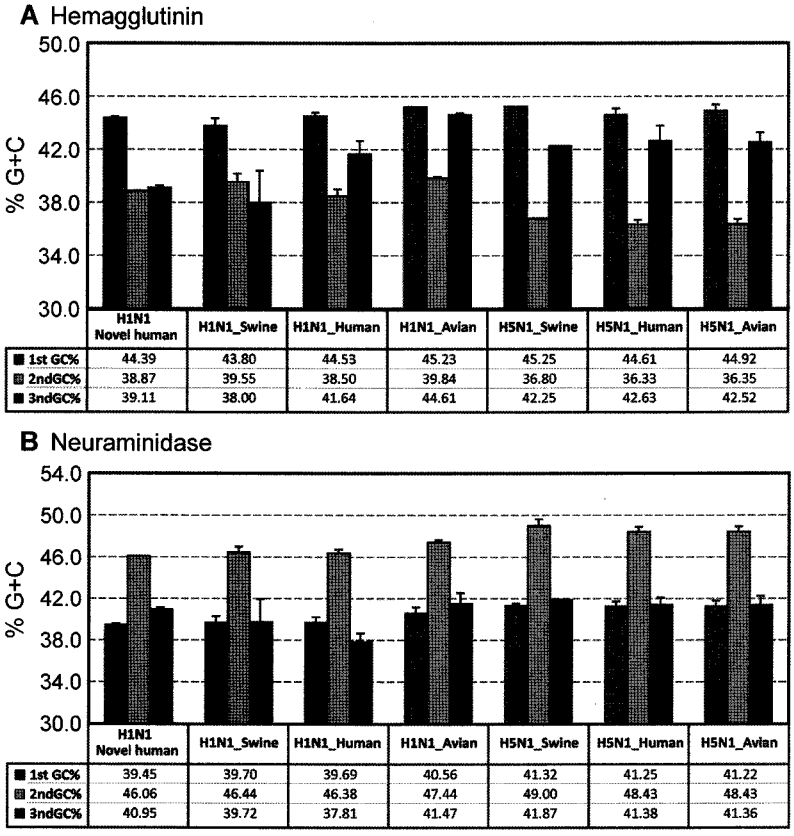

Fig. 2. Nucleotide composition (%G+C contents) on the first (1st GC%), second (2nd GC%), and third (3rd GC%) codon positions of the (**A**) hemagglutinin and (**B**) neuraminidase genes as a measure of overall average in both H1N1 and H5N1 subtype influenza A viruses. 95% Confidence bars are shown and actual means are below each bar. Each host name is shown with its influenza subtype on the X-axis, and all the accession numbers are provided in Appendix II.

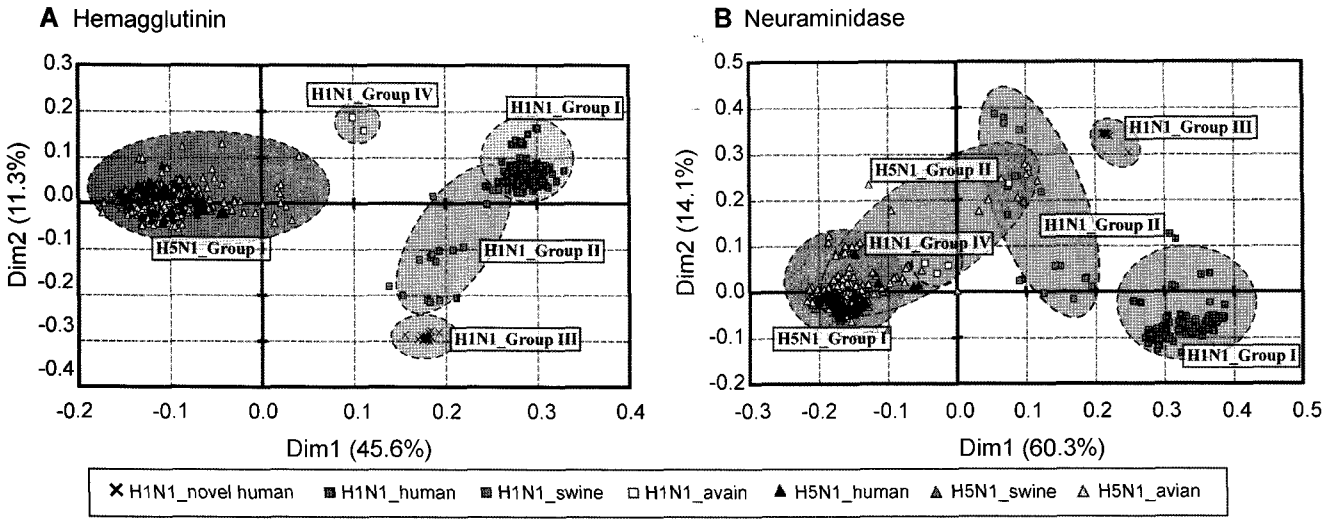

Fig. 3. Scatterplots of the correspondence analysis results using the relative synonymous codon usage values of the (**A**) hemagglutinin and (**B**) neuraminidase genes of influenza A virus subtypes H1N1 and H5N1 isolated from human, swine, and avian species.
Dim1 and dim2 represent the values of the first- and second-dimensional factors of each sequence. The percentage in each parenthesis indicates the percent inertia of each axis given by correspondence analysis, and all the accession numbers are provided in Appendix II.

H1N1 subtypes, whereas those of color pink mean H5N1 groups. In both HA and NA genes, the H1N1 and H5N1 subtype groups were located on the opposite site along the X-axis (dim1), indicating the different amino acid compositions between those two groups. In the NA gene, however, avian-origin H1N1 (H1N1_Group IV) and H5N1 (H5N1_Gourp II) subtypes were distributed near the opposite subtype groups. The novel H1N1 subtypes (H1N1_Group III) were located on the same side with other H1N1 subtypes along the X-axis, but there was no H1N1 group that showed the exact same pattern as H1N1_Group III. On the basis of the Y-axis (dim1) in the CA plots, most of the H5N1 subtypes were located near the origin of the Y-axis, whereas the novel H1N1 group (H1N1_Group III) was highly biased on dim2. The swine-origin H1N1 subtypes (H1N1_Group II) were broadly distributed from the origin to the H1N1_Group III area. In both HA and NA genes, human-origin H1N1 subtype (H1N1_Group I) showed different synonymous codon patterns with the novel viruses (H1N1_Group III).

To determine more intensively the characteristics of the novel H1N1 subtypes that occurred in 2009, we selected three NA sequences such as the novel H1N1 subtype [A/New York/2009 (FJ984362)[, swine-origin H1N1 subtype [A/swine/Ohio/2007 (EU409949)[, and human-origin H1N1 subtype [A/North Carolina/2008 (EU779651)], and compared the RSCU differences in each codon (Table 2). Each number and calculated RSCU value is shown, and the highest RSCU value in each synonymous codon group in each species is presented as an underlined-bold number. In terms of the most frequently used codons, the novel H1N1 subtype revealed similar patterns with other two sequences. However, it showed slightly different RSCU patterns in synonymous codon groups for alanine, aspartate, glycine, isoleucine, leucine, serine, and valine from A/swine/Ohio/2007, whereas it revealed different patterns in alanine,

Table 1. R^2 values and significance levels of linear regression test between dim1 and dim2 as given by correspondence analysis, and % G+C contents at each codon position and effective number of codons (ENC) of hemagglutinin and neuraminidase genes of influenza A virus H1N1 and H5N1 subtypes as well as the novel H1N1 influenza virus.

Gene	CA[a] result	1st GC%[b]	2nd GC%[b]	3rd GC%[b]	ENC[c]
Hemagglutinin	Dim1	0.1443**	**0.7777****	0.2019**	0.2469**
	Dim2	0.0147*	0.0349**	**0.3736****	0.3425**
Neuraminidase	Dim1	0.6646**	**0.7655****	0.3905**	0.0063*
	Dim2	0.0149*	0.0259**	0.1068**	**0.4976****

[a]Correspondence analysis.
[b]% Contents of guanine-cytosine at the 1st (GC_{1st}), 2nd (GC_{2nd}) and 3rd (GC_{3rd}) codon position.
[c]Effective number of codons.
*P<0.05.
**P<0.0001.

Table 2. Synonymous codon usage profile of the neuraminidase genes of influenza A viruses, including the H1N1 subtypes of newly emerged (A/New York/2009), swine (A/swine/Ohio/2007), and human (A/North Carolina/2008) influenza A viruses.

		A/New York/2009[a]		A/swine/Ohio/2007[b]		A/North Carolina/2008[c]	
AA[d]	Codon	N[e]	RSCU[f]	N	RSCU	N	RSCU
ALA	GCA	18	**2.12**[g]	6	1.60	8	**1.60**
	GCC	9	1.06	2	0.53	4	0.80
	GCG	2	0.24	0	0.00	0	0.00
	GCU	5	0.59	7	**1.87**	8	**1.60**
ARG	AGA	14	**4.67**	12	**4.24**	11	**4.12**
	AGG	4	1.33	3	1.06	3	1.12
	CGA	0	0.00	1	0.35	1	0.38
	CGC	0	0.00	0	0.00	0	0.00
	CGG	0	0.00	0	0.00	0	0.00
	CGU	0	0.00	1	0.35	1	0.38
ASN	AAC	14	0.68	13	0.74	15	0.81
	AAU	27	**1.32**	22	**1.26**	22	**1.19**
ASP	GAC	13	1.00	8	0.70	11	0.96
	GAU	13	1.00	15	**1.30**	12	**1.04**
CYS	UGC	6	0.80	9	0.95	8	0.89
	UGU	9	**1.20**	10	**1.05**	10	**1.11**
GLN	CAA	6	0.80	6	0.86	11	**1.83**
	CAG	9	**1.20**	8	**1.14**	1	0.17
GLU	GAA	24	**1.37**	12	**1.26**	11	**1.10**
	GAG	11	0.63	7	0.74	9	0.90
GLY	GGA	13	1.30	20	**1.78**	21	**1.91**
	GGC	4	0.40	8	0.71	9	0.82
	GGG	14	**1.40**	9	0.80	8	0.73
	GGU	9	0.90	8	0.71	6	0.55
HIS	CAC	7	0.93	3	0.75	3	0.75
	CAU	8	**1.07**	5	**1.25**	5	**1.25**
ILE	AUA	12	0.97	25	**1.63**	22	**1.50**
	AUC	6	0.49	6	0.39	10	0.68
	AUU	19	**1.54**	15	0.98	12	0.82
LEU	CUA	14	**1.83**	4	1.14	4	1.14
	CUC	5	0.65	0	0.00	0	0.00
	CUG	10	1.30	4	1.14	2	0.57
	CUU	1	0.13	2	0.57	2	0.57
	UUA	6	0.78	6	**1.71**	6	1.71
	UUG	10	1.30	5	1.43	7	**2.00**
LYS	AAA	27	**1.29**	12	**1.20**	13	**1.13**
	AAG	15	0.71	8	0.80	10	0.87
PHE	UUC	11	**1.16**	10	**1.18**	8	0.94
	UUU	8	0.84	7	0.82	9	**1.06**
PRO	CCA	9	**1.89**	10	**1.90**	6	1.20
	CCC	4	0.84	2	0.38	2	0.40
	CCG	5	1.05	2	0.38	4	0.80
	CCU	1	0.21	7	1.33	8	**1.60**
SER	AGC	9	1.15	10	1.13	8	0.96
	AGU	7	0.89	15	**1.70**	14	1.68
	UCA	17	**2.17**	13	1.47	16	**1.92**
	UCC	4	0.51	4	0.45	3	0.36
	UCG	1	0.13	2	0.23	1	0.12
	UCU	9	1.15	9	1.02	8	0.96
THR	ACA	23	**2.49**	12	**1.55**	10	1.29
	ACC	2	0.22	7	0.9	11	**1.42**
	ACG	3	0.32	1	0.13	0	0.00
	ACU	9	0.97	11	1.42	10	1.29
TYR	UAC	14	**1.04**	7	**1**	8	**1.14**
	UAU	13	0.96	7	**1**	6	0.86
VAL	GUA	17	**1.89**	6	0.86	4	0.57
	GUC	6	0.67	6	0.86	4	0.57
	GUG	7	0.78	6	0.86	9	1.29
	GUU	6	0.67	10	**1.43**	11	**1.57**

[a]GenBank Accession No. FJ984362 [novel influenza A (H1N1)].
[b]GenBank Accession No. EU409949 [swine influenza A (H1N1)].
[c]GenBank Accession No. EU779651 [human influenza A (H1N1)].
[d]AA stands for amino acid.
[e]N stands for count of codons.
[f]RSCU stands for the relative synonymous codon usage.
[g]Underlined codons are those most frequently used among each synonymous codon group.

glutamine, glycine, isoleucine, phenylalanine, proline, threonine, and valine. There were different patterns between A/swine/Ohio/2007 and A/North Carolina/2008 in the glutamine, leucine, phenylalanine, proline, serine, threonine, and tyrosine. In terms of the GC or AT composition on the 3rd codon position, however, all the sequences commonly preferred to use the AT-ended codons.

Discussion

In this study, we calculated and compared various kinds of nucleotide patterns such as % G+C contents on each codon position, relative synonymous codon usage (RSCU) values, and the effective number of codons (ENC) using the 1,675 sequences of HA and NA genes of both H1N1 and H5N1 influenza A viruses. According to the result of the phylogenetic analysis, the novel H1N1 influenza viruses were branched out from other swine-origin H1N1 viruses, creating a distinct taxon group in both HA and NA genes (Fig. 1). We also generated the neighbor-joining trees using known H1N1 influenza A viruses isolated from human- and avian-hosts, but the novel H1N1 group showed far more distances from those viruses (data not shown), supporting the results from Bronze [M.S.Bronze. 2009. http://emedicine.medscape.com/article/1673658-print], which reported that the HA and NA genes of the novel H1N1 strains were similar to those of swine-origin H1N1 viruses isolated from the U.S.A. and Europe, respectively. We found that the HA and NA genes of these novel viruses were relatively close correlations with swine-origin viruses isolated in 2004, rather than more recently isolated ones.

To compare the genetic differences between the novel H1N1s and other H1N1 and H5N1 subtype viruses more intensively, we calculated the overall average of % G+C contents on each codon position of all the target genes (Fig. 2). Interestingly, the HA genes of novel H1N1 viruses showed very similar frequency ratios on the first, second, and third codon positions with those of swine-origin H1N1 subtypes, resulting as 36:32:32 and 36:33:31, respectively. That of the human-origin H1N1 viruses, however, revealed different patterns, resulting as 36:30:33, and it was more similar to those of H5N1 subtypes (Fig. 2A). In NA genes, the frequency ratios on three codon positions revealed reverse patterns with those of HA genes, showing the highest value in each second codon position, and the novel H1N1 viruses also showed more similar frequency ratios with other swine-origin H1N1 viruses than human-origin viruses (Fig. 2B). In this study, we confirmed that although the novel H1N1 influenza A viruses were branched out and generated a distinct group from other swine-origin H1N1 viruses in the phylogenetic analysis, they showed very similar % G+C pattern ratios among three codon positions with swine-origins than all the other viruses. Our results suggested that the novel H1N1s were more affected from swine-origin H1N1s, on the basis of not only the overall sequences of HA and NA genes, but also the nucleotide compositions such as %G+C contents, than other similar viruses.

Moreover, we also found that, on the basis of the synonymous codon usage patterns, the novel H1N1 viruses showed the opposite codon usage patterns with the H5N1 subtypes along with the X-axis (dim1), and although they were distinctly located among H1N1 influenza A viruses, they also showed relatively similar patterns with swine-origin H1N1 viruses, especially in the HA genes (Fig. 3A). Unlike the H5N1 subtypes which were gathering thickly near the origin of the Y-axis, the H1N1 viruses were broadly distributed along the dim 2 (Y-axis), indicating that there were diverse synonymous codon patterns among them. The novel H1N1s showed highly biased synonymous codon usage patterns in both HA and NA genes, revealing the opposite pattern to the human-origin H1N1 viruses (H1N1_Group I). In terms of the overall composition of GC or AT bases on the 3rd codon position, however, both the human- and swine-origin H1N1 subtypes showed similar patterns with the novel H1N1s, commonly preferring to use the AT-ended codons (Table 2). On the other hand, the NA genes of avian-origin H1N1 (H1N1_Group IV) and H5N1 subtypes (H5N1_Gourp II) were distributed near or within the opposite subtype groups in NA genes, indicating that avian-hosts might function as a messenger between two different influenza A subtypes such as H1N1 and H5N1 (Fig. 3B).

Since early 2009, many countries in the world have been shocked by the novel H1N1 influenza A pandemic. This novel virus is known to contain genes closely related to swine influenza [Trifonov *et al.* 2009. http://www.eurosurveillance.org/images/dynamic/EE/V14N17/art19193.pdf], but the exact origin of this new strain is not yet determined. In this study, we compared the codon usage patterns of this novel species with other similar viruses, as well as the phylogenetic analysis, and we found that the novel H1N1 influenza A viruses showed the most close correlations with the swine-origin H1N1 subtypes than other H1N1 viruses, in the result from not only the phylogenetic analysis, but also the analysis of nucleotide compositions. Although the genetic sequences of novel H1N1 subtypes were not exactly the same as the other H1N1 subtypes, the HA and NA genes of novel H1N1s showed very similar codon usage patterns with other H1N1 subtypes, especially with the swine-origin H1N1 influenza A viruses. Our findings strongly suggested that those novel H1N1 viruses seemed to be originated from the swine-host H1N1 viruses in terms of the codon usage patterns. More studies about the transmission pathway of these novel subtypes from the swine to the human population are required to trace additional variations among the novel viruses.

Acknowledgments

This study was supported by the fund of the Korea Institute of Science and Technology Information. We also acknowledge the invaluable contribution of the researchers who have made their data publicly available.

REFERENCES

1. Ahn, I. and H. S. Son. 2006. Epidemiological comparisons of codon usage patterns among HIV-1 isolates from Asia, Europe, Africa and the Americas. *Exp. Mol. Med.* **38:** 643–651.
2. Ahn, I. and H. S. Son. 2007. Comparative study of the hemagglutinin and neuraminidase genes of influenza A virus H3N2, H9N2, and H5N1 subtypes using bioinformatics techniques. *Can. J. Microbiol.* **53:** 830–839.
3. Ahn, I., B. J. Jeong, S. E. Bae, J. Jung, and H. S. Son. 2006. Genomic analysis of influenza A viruses, including avian flu (H5N1) strains. *Eur. J. Epidemiol.* **21:** 511–519.
4. Bao, Y., P. Bolotov, D. Dernovoy, B. Kiryutin, L. Zaslavsky, T. Tatusova, J. Ostell, and D. Lipman. 2008. The influenza virus resource at the National Center for Biotechnology Information. *J. Virol.* **82:** 596–601.
5. Cary, N. C. 2004. *SAS 9.1.2 Qualification Tools User's Guide*, pp. 3–18. SAS Institute Inc., North Carolina.
6. Cox, N. J., R. A. Black, and A. P. Kendal. 1989. Pathways of evolution of influenza A (H1N1) viruses from 1977 to 1986 as determined by oligonucleotide mapping and sequencing studies. *J. Gen. Virol.* **70 (Pt 2):** 299–313.

7. Hair, J. F., R. E. Anderson, R. L. Tatham, and W. C. Black. 1998. *Multivariate Data Analysis*, pp. 519–570. 5th Ed. Prentice Hall, Upper Saddle River, New Jersey.
8. Johnson, N. P. and J. Mueller. 2002. Updating the accounts: Global mortality of the 1918–1920 "Spanish" influenza pandemic. *Bull. Hist. Med.* **76:** 105–115
9. Johnson, R. A., D. W. Wichern, and E. C. Holmes. 2002. *Applied Multivariate Statistical Analysis*, pp. 700–745. 5th Ed. Prentice Hall, Upper Saddle River, New Jersey.
10. Larkin, M. A., G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, *et al.* 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* **23:** 2947–2948.
11. Lynn, D. J., G. A. Singer, and D. A. Hickey. 2002. Synonymous codon usage is subject to selection in thermophilic bacteria. *Nucleic Acids Res.* **30:** 4272–4277.
12. Moriyama, E. N. and D. L. Hartl. 1993. Codon usage bias and base composition of nuclear genes in Drosophila. *Genetics* **134:** 847–858.
13. Nelson, M. I., C. Viboud, L. Simonsen, R. T. Bennett, S. B. Griesemer, St. K. George, *et al.* 2008. Multiple reassortment events in the evolutionary history of H1N1 influenza A virus since 1918. *PLoS Pathog.* **4:** e1000012.
14. Perrière, G. and J. Thioulouse. 2002. Use and misuse of correspondence analysis in codon usage studies. *Nucleic Acids Res.* **30:** 4548–4555.
15. Sharp, P. M. and W. H. Li. 1986. Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for 'rare' codons. *Nucleic Acids Res.* **14:** 7737–7749.
16. Singer, G. A. and D. A. Hickey. 2003. Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. *Gene* **317:** 39–47.
17. Suzuki, H., R. Saito, and M. Tomita. 2009. Measure of synonymous codon usage diversity among genes in bacteria. *BMC Bioinformatics* **10:** 167.
18. Swofford, D. L. 1999. *PAUP*. Phylogenetic Analysis using Parsimony (*and Other Methods)*, Version 4. Sinauer Associates, Sunderland, Massachusetts.
19. Taubenberger, J. K. and D. M. Morens. 2006. 1918 influenza: The mother of all pandemics. *Emerg. Infect. Dis.* **12:** 15–22.
20. Vicario, S., E. N. Moriyama, and J. R. Powell. 2007. Codon usage in twelve species of Drosophila. *BMC Evol. Biol.* **7:** 226.
21. Zhou, M. and X. Li. 2008. Analysis of synonymous codon usage patterns in different plant mitochondrial genomes. *Mol. Biol. Rep.* [Epub ahead of print.]