

Codon Bias of Influenza A Viruses and Their Hosts

Honors Thesis in Computer Science

Keli C. Fancher

Table of Contents

Title Page	i
Table of Contents	ii
Lists of Equations, Tables, and Figures	iii
Acknowledgments.....	iv
Abstract	1
1. Introduction	2
2. Materials and Methods	4
2.1 Sequence Data	4
2.2 GC Content	5
2.3 Effective Number of Codons - ENC	5
2.4 Relative Synonymous Codon Usage - RSCU.....	5
2.5 Synonymous Codon Usage Order - SCUO	6
2.6 Codon Volatility	7
2.7 Dinucleotide Odds Ratio	8
3. Results.....	9
3.1 Overview	9
3.2 GC content.....	9
3.3 ENC, SCUO, and Codon Volatility.....	12
3.4 Association of ENC and GC Content	13
3.5 Dinucleotide Odds Ratio	14
3.6 RSCU and Odds Ratio	15
4. Conclusion	17
Works Cited.....	18

Lists of Equations, Tables, and Figures

List of Equations	
Equation 1: GC Content.....	5
Equation 2: ENC	5
Equation 3: RSCU.....	6
Equation 4-8: SCUO	6-7
Equation 4: Normalized Codon Frequency	6
Equation 5: Amino Acid Entropy	6
Equation 6: Amino Acid SCUO	6
Equation 7: Composition Ratio	7
Equation 8: Overall SCUO	7
Equation 9: Codon Volatility.....	7
Equation 10: Dinucleotide Odds Ratio.....	8
List of Tables and Figures.....	
Table 1: Summary of Datasets	4
Figure 1: GC Content across Influenza Genomes.....	11
Table 2: Influenza GC Content Summaries.....	11
Figure 2: ENC, SCUO, and Volatility values.....	12
Figure 3: ENC vs GC3 Content of the four viruses.....	13
Table 3: Highest and Lowest Common Dinucleotides.....	14
Table 4: Ordered list of codons by RSCU values	16

Acknowledgments

I would firstly like to thank Dr. Hu for his work as Chair on my Honors Committee and his constant guidance throughout this entire project, likewise I would like to thank the remaining professors on the committee, Dr. Irmgard Howard and Dr. Laurie Dashnau, for their assistance and time. In addition I would like to thank Houghton College for its financial support as well as Francisco Lobo for helpful discussion on the subject matter. Finally, I would like to thank my fellow Honors Student, Erin Bard, for his assistance in revising this work.

Abstract

The Influenza A virus remains an annual and worldwide health concern due to its fast evolutionary rate. There are two major forms of Influenza evolution: drift, caused by mutations, and shift, which results from the exchange of genetic information between two gene segments. Using six synonymous codon usage bias indexes (GC content, ENC, SCUO, Codon Volatility, RSCU, and Odds Ratio), this study revealed the evolutionary drift patterns in the Influenza A viruses of avian, human, and swine origins as well as those of their hosts. We found that the variation of GC content across the 11 genes in Influenza A uniquely determines the viral origins in avian, human, and swine hosts. As was previously noticed in the Flaviviridae virus, a codon's RSCU value of the Influenza viruses is positively correlated to the Odds Ratio of the dinucleotides contained within that codon. Additionally, the RSCU values of avian, human, and swine viruses and their corresponding hosts are similar, which is also true of Odds Ratio. Furthermore, the GC content, ENC, SCUO, and Codon Volatility are similar across the avian, human, and swine hosts; however, the RSCU and Odds Ratio of the hosts are distinct. Our findings expanded the knowledge on codon bias of Influenza viruses and their hosts.

1. Introduction

Due to its frequent evolution, the Influenza A Virus continues to be a worldwide health concern. There are two primary ways in which the Influenza A Virus evolves: drift and shift. Evolutionary shift occurs when two Influenza viruses exchange genetic information, and the resulting viruses usually represent a very large and rapid evolutionary change. This rapid change will sometimes generate a cross-species shift, as occurred in the 2009 H1N1 pandemic [1,2]. Conversely, evolutionary drift tends to occur slowly as mutations accumulate over time. These gradual changes are caused by the high error rate in the Influenza's RNA replication.

Evolutionary drift in Influenza is commonly evaluated through the study of codon usage. A codon is a combination of three out of the four possible nucleic bases in gene sequences: Adenine - 'A', Cytosine - 'C', Guanine - 'G', and Thymine - 'T'. There are 64 possible, 4^3 , trinucleotide codons, each of which encodes one of the 21 amino acids. As such (since 64 codons encode only 21 amino acids), there is degeneracy in the genetic code that results in synonymous codons, or codons that encode the same amino acid, which are not commonly used with equal frequency. Since synonymous codons are not utilized with equal frequency, their usage bias is often applied to detect whether mutations have occurred. Studying the usage bias of codons helps to reveal evolutionary information about individual genes or genomes, which can be used to design DNA primers and detect horizontal transfer events [3]. It is also possible that unique trends exist within the synonymous codon bias of individual virus/host pairs, which will enable the recent genetic history of a virus to be identified based on these unique patterns.

Many different methods of codon usage bias detection have been developed [3-15]. We focus on 6 frequently used bias detectors: GC content, ENC, RSCU, SCUO, Codon Volatility, and Odds Ratio. Among these, GC content, ENC, and RSCU are the simplest to compute and have been widely used in previous works [3-8]. In 2002, Zeeberg [5] developed a bias indicator that utilized Shannon Informational Theory within its calculations. Leading up to this new indicator, results were only applicable within a genome, but the Zeeberg technique, based on information theory, allowed codon bias to be compared across genomes. J. Zhou et al. [6,7] further improved Zeeberg's method and created the Synonymous Codon Usage Order (SCUO). SCUO was an improvement over

Zeeberg's approach in that it applied maximum entropy to normalize the results. The same year they continued their studies with SCUO and created a quantitative relationship between SCUO and the GC content [8]. In 2003-2004, Joshua B. Plotkin et al. [9,10] developed another information theory based bias detector. This new detector utilized codon volatility, where the volatility of a codon is the probability that a random point mutation will result in a nonsynonymous codon (a codon that does not encode the same amino acid). Later, Jianzhi Zhang [11] showed that codon volatility can be increased slightly by strong frequency-dependent selection, as well as other unrelated factors; whereas the volatility is not effected by directional positive selection. Zhang concluded that codon volatility could only have limited use for detecting positive selection. Another bias detector that is often utilized is the dinucleotide odds ratio [12-15], in which dinucleotides, a pair of nucleotide bases, are evaluated instead of codons. Benjamin D. Greenbaum et al. [14] found that viruses tend to mimic the hosts in that the same dinucleotides are pervasively over- or under- represented in Influenza. Francisco P. Lobo [15] furthered this methodology within the Flaviviridae family of viruses and their hosts to show similar findings.

In this study we analyzed the codon usage bias of Influenza A viruses of avian, human, and swine origins compared with that of their hosts, using the aforementioned methods. It has been previously proposed [14,15] that a connection exists between the synonymous codons used in viral genome sequences and the host that they infect in both Influenza and Flaviviridae. We intend to further this research in Influenza A by broadening the scope to include avian, human, and swine viruses and their infected hosts, as well as using additional codon bias measurements.

2. Materials and Methods

2.1 Sequence Data

The data used in this study included gene sequences in avian, human, swine, and the 2009 H1N1 pandemic viruses, as well as the mRNA transcripts from their corresponding hosts (avian, human, and swine). All available and fully sequenced Influenza strains (n=3007) of avian, human, or swine origin were obtained from the National Center for Biotechnology Information (NCBI) [16], which included 7 H subtypes (H1, H2, H3, H4, H5, H6, and H7) and 7 N subtypes (N1, N2, N3, N6, N7, N8, and N9). These sequences were then using MAFFT (Multiple Alignment using Fast Fourier Transform) [17], the N and H subtypes were aligned based on subtype. Only unique sequences were selected for evaluation, resulting in a total of 32,566 sequences. The host mRNA transcripts were obtained from NCBI using RefSeq [18]. Table 1 lists the dispersion of all sequences in both the viral and host datasets we used in this study.

Initial sequence processing consisted of verifying open reading frames and multiple sequence alignment using MAFFT. After completion Python scripts were created to perform the codon bias calculations and analyze the results.

Table 1 Summary of Datasets

Gene\Sequence Count	Human	Avian	Swine	2009 H1N1 Pandemic	Total
PB1	1810	973	103	111	2997
PB2	1808	972	103	112	2995
NA	1798	973	103	112	2986
PA	1810	973	103	112	2998
M1	1810	973	103	112	2998
M2	1729	951	103	112	2895
NP	1791	973	103	111	2978
NS1	1810	972	103	112	2997
NS2	1805	961	103	112	2981
HA	1810	972	103	113	2998
PB1-F2	1801	858	80	4	2743
Total	19782	10551	1110	1123	32566
Transcript Count	Human	Avian	Swine	-	Total
Host mRNA Transcripts	39130	19266	7967	-	66363

The dispersion of the viral and host datasets utilized in this study. The 32566 total sequences belonging to the viral datasets are shown as they are divided among the four datasets and their 11 genes. The number of host mRNA transcripts per host are also shown for each host used in our study.

2.2 GC Content

The frequency of which a Guanine – ‘G’ or a Cytosine – ‘C’ nucleotide appears at the third position of the codons in a gene is the GC3 content. This measurement has been shown through previous studies to correlate very strongly with the codon usage bias of a gene [19]. The GC content provides a simple technique to verify other codon usage bias indices because of its strong correlation with the usage bias on the whole gene. The GC content is usually found as in Equation 1, where G, C, A, and T represent the number of times that Guanine, Cytosine, Adenine and Thymine appear at a specific position in the codon. This gives the percentage of the GC content of the entire gene. The GC content can also be narrowed to just the third nucleotide position (GC3) by only counting the frequency of the bases at every third nucleotide.

$$\text{Equation 1: GC content} = \frac{G+C}{A+C+T+G} \times 100$$

2.3 Effective Number of Codons - ENC

The effective number of codons (ENC) is another simple metric used to quantify the synonymous codon usage bias of a gene; Wright [20] first proposed it in 1990. ENC estimates the absolute synonymous codon usage bias, which will range from 20, when only one codon is used per amino acid, to 61, when all synonymous codons are used with equal frequency.

$$\text{Equation 2: ENC} = 2 + \frac{9}{F2} + \frac{1}{F3} + \frac{5}{F4} + \frac{3}{F6}$$

Presented in Equation 2 is the formula for ENC, in which $F2$ is the probability that two randomly chosen codons for an amino acid, possibly encoded by two distinct codons, are identical. Likewise, $F3$ is the probability that three randomly chosen codons for an amino acid with three synonymous codons are identical, and so forth. This yields an easy-to-understand representational value for the synonymous codon dispersion within a gene. However, ENC is still quite limiting in that it does not provide specific details on codon usage frequency.

2.4 Relative Synonymous Codon Usage - RSCU

RSCU is an approach for examining the synonymous codon usage in a gene without the influence of amino acid composition [3]. RSCU values are applied to find the relative use of every unique codon in a gene. The RSCU of the j th codon for the i th amino acid is

defined as Equation 3, where x_{ij} is the frequency of the j th codon for the i th amino acid and n_i is the number of alternative synonymous codons for the i th amino acid.

$$\text{Equation 3: } \text{RSCU}_{ij} = \frac{x_{ij}}{\frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}}$$

Since RSCU values are calculated for each codon relative to all synonymous codons, it provides a more detailed analysis than the previous methods, which only calculate one result for an entire sequence.

2.5 Synonymous Codon Usage Order - SCUO

Synonymous codon usage order is based on information theory and was developed to analyze patterns in DNA sequences [6,7]. To implement SCUO, a codon table, which contains all amino acids that have more than one codon, is created. This allows the referencing of the j th synonymous codon for the i th amino acid, where $1 \leq i \leq 19$ and $1 \leq j \leq n_i$ and where n_i represents the number of synonymous codons for the i th amino acid. In each of these instances, x_{ij} will represent the occurrence of the j th synonymous codon for amino acid i . The following sequence of equations describes the SCUO calculation.

First, the frequency of the i th degenerate codon of amino acid i is found by normalizing x_{ij} as follows.

$$\text{Equation 4: } p_{ij} = \frac{x_{ij}}{\sum_{j=1}^{n_i} x_{ij}}$$

According to Shannon's information theory [6], the entropy of the i th amino acid is calculated next (Equation 5) where n_i represents the number of synonymous codons for the j th amino acid. The maximum entropy will occur when every codon is used with equal frequency.

$$\text{Equation 5: } H_i = - \sum_{j=1}^{n_i} p_{ij} \log p_{ij}$$

Next, the normalized difference between the maximum entropy and the observed entropy for the j th amino acid in each sequence is calculated. This value, O_i (Equation 6), will be the SCUO for the j th amino acid in each sequence.

$$\text{Equation 6: } O_i = \frac{H_i^{max} - H_i}{H_i^{max}}$$

The composition ratio of the j th amino acid in each sequence is calculated as F_i , (Equation 7), where the sum is taken from 1 to 18 to account for the amino acids that are encoded by more than one codon.

$$\text{Equation 7: } F_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{\sum_i^{18} \sum_{j=1}^{n_i} x_{ij}}$$

Finally, the average SCUO for each sequence is represented as in Equation 8.

$$\text{Equation 8: } O = \sum_{i=1}^{n_i} F_i O_i$$

2.6 Codon Volatility

In 2003 J. Plotkin and J. Dushoff [9] compared the codon usage in the HA, NA, and NP genes of the Influenza A viral subtype H3N2 from the year 1968 to the year 2000. Their studies revealed that while comparing the codon bias of these genes errors could easily accrue. This, they determined, was due to the fact that one of the genes could contain many more codons encoding volatile amino acids than the other genes.

To account for the impact of amino acid composition on codon usage calculation they incorporated the volatility of codons, defined below, into the measurement of the codon usage. Their approach is based on the presupposition that each codon has nine possible single-nucleotide mutational neighbors, where each of the three-nucleotide bases can individually mutate into one of the three other bases. Some of these single-nucleotide mutational neighbors will be synonymous mutations and the rest will be non-synonymous. This method is better suited for the study of codon bias compared to other metrics because, in the context of influenza mutations, the interest is in codon usage biased toward increasing or decreasing the amino acid substitution rate [9]. Focusing on the volatility of a codon, or how easily a codon change would produce a new amino acid, allows for better comparison between species.

The volatility of a codon measures the degree to which a random single-nucleotide mutation will change the corresponding amino acid and is calculated as Equation 9, where $d(x,y)$ is the Hamming distance between codons x and y .

$$\text{Equation 9: } V(c) = \sum_{i=1}^9 d(\text{acid}(c_i), \text{acid}(c))$$

The Hamming distance is defined as $d(\text{acid}(c_i), \text{acid}(c)) = 0$ if $\text{acid}(c_i)$ encodes the same amino acid as $\text{acid}(c)$ otherwise, if it encodes a different amino acid, it is defined as 1 and a substitution occurs.

The volatility of a codon (with a Hamming distance metric), $V_H(c)$, quantifies the degree to which a random nucleotide mutation will cause an amino acid substitution. Assuming that all nucleotides will have an equal rate of mutation and are equally exchangeable, the volatility of a codon is the ratio of point-mutational neighbors to total possible single point mutations. For example, TTG in the amino acid Leu will have a volatility of 6/8, since 6 of its 8 non-stop codon neighbors are non-synonymous [11].

2.7 Dinucleotide Odds Ratio

The Odds Ratio calculation is commonly used to evaluate dinucleotides, pairs of nucleotides, in gene sequences. Odds Ratio is a likelihood of observing a dinucleotide in a sequence and is calculated as in Equation 10.

$$\textbf{Equation 10: } P_{xy} = \frac{f_{xy}}{f_x f_y}$$

In Equation 10 x and y represent the nucleotides that form dinucleotide xy ; and f_x, f_y, f_{xy} denote the frequencies of nucleotide x , nucleotide y , and dinucleotide xy respectively. Karlin et al. [21] showed that dinucleotides with an odds ratio that is outside of the range $[0.78, 1.25]$ could be considered as being more under- or over- represented than normal.

3. Results

3.1 Overview

All of the above-mentioned measurements (GC content, ENC, SCUO, Codon Volatility, RSCU, and Odds Ratio) were calculated on the four viral (avian, human, 2009 H1N1, and swine) and three host (avian, human, and swine) data sequences. The GC content, ENC, SCUO, and Codon Volatilities of the avian, human, and swine host mRNA transcripts were not reported in this study because the results were similar across the hosts. The only measurements that provided distinct results across the avian, human, and swine hosts were RSCU and Odds Ratio (sections 3.5 and 3.6).

3.2 GC content

We summarized our comparison of the GC content in avian, human, swine, and 2009 H1N1 viruses in Figure 1 and Table 2. Figure 1 illustrates the GC, GC1, GC2, and GC3 content, represented by blue, green, red, and yellow respectively, at each gene of different origin. From the chart it is clear that the 4 measurements of GC content are similar across the avian, human, and swine hosts, represented by the blue trends, and as presumed, the GC3 content (yellow) displays the most diversity between the genomes. The NA gene has one of the most diverse GC contents across all four influenza hosts; this was anticipated as NA, as well as HA, is a surface protein of the virus and undergoes the most selection pressure. Surprisingly, however, the HA gene does not show the same bias, although the PB1-F2 gene does. Table 2 was constructed from the results in Figure 1. The mean and variation were calculated for the GC content of these genes of different host origins. The overall GC content of a codon (The top four rows in Table 3) has the lowest deviation of the 4 GC contents (GC, GC1, GC2, and GC3).

Nevertheless, a distinct pattern of GC content in the three primary viral species (avian, human, and swine) was discovered. For each of these the variation, across the 11 genes in Influenza, of the single-nucleotide GC content uniquely identifies its host. Avian viruses deviated in the range of 3.57 to 3.93 over the genes, with an error rate of 0.19; human viruses deviated in the range of 3.95 to 4.22, with an error rate of 0.12; and swine viruses deviated in the range of 4.51 to 6.57, with an error rate of 1.03. The standard deviation of the swine group is an order of magnitude greater than the avian or human groupings, which we have attributed to the composition of the swine sequences. In the

case of swine, it appears that the third nucleotide in a codon generally has a much higher GC content compared to that of the other two positions, which results in a comparatively larger standard deviation.

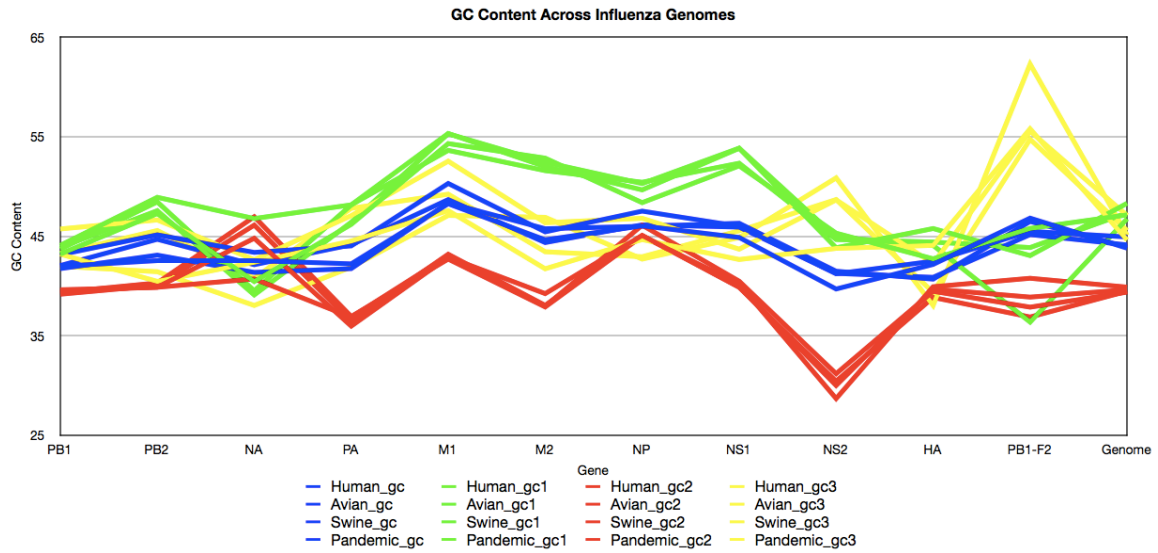


Figure 1 GC Content across Influenza Genomes

The mean GC content for every gene in the avian, human, swine, and 2009 H1N1 sequences. The x-axis represents the GC content at each of the 11 genes as well as the GC content over the entire genome. The blue trend lines represent the complete GC content; whereas the green, red, and yellow trend lines represent the GC content at the first, second, and third codon positions respectively.

Table 2 Influenza GC Content Summaries

Influenza GC Content		
Host	Mean	STDEV
2009 H1N1 GC	44.13	2.27
Avian GC	44.93	2.40
Human GC	43.80	2.42
Swine GC	43.91	2.57
STDEV	-	0.12398
Avian GC2	39.38	3.57
Avian GC3	47.09	3.83
Avian GC1	48.31	3.93
STDEV	-	0.18807
Human GC3	44.63	3.95
Human GC2	39.63	3.98
Human GC1	47.15	4.22
2009 H1N1 GC2	39.47	4.16
2009 H1N1 GC3	45.58	4.16
2009 H1N1 GC1	47.35	4.22
STDEV	-	0.12155
Swine GC2	39.89	4.51
Swine GC1	46.75	5.68
Swine GC3	45.10	6.57
STDEV	-	1.03356

The calculated arithmetic mean and standard deviation for each species and GC-type from the same viruses as in Figure 1. Each row represents a mean GC content and STDEV for a viral species across all 11 genes (shown in Figure 1).

3.3 ENC, SCUO, and Codon Volatility

Figure 2 shows the plotted ENC, SCUO, and volatility values for each gene as well as the trend lines for each host. Since ENC is an inverse probabilistic measurement of two randomly selected synonymous codons being identical, the lower the value the higher the bias; whereas in SCUO and volatility, the higher the value the more biased the virus. Observe in Figure 2 that only the ENC values for the M2 and the PB1-F2 genes are consistently highly biased (notice their extreme variance above the trend lines). The same trend is observed with SCUO in Figure 2b. However, Figure 2c shows that only PB1-F2 is strongly biased based on codon volatility, compared to the trend line of the whole genome.



Figure 2 ENC, SCUO, and Volatility values of the four viruses and their trend lines

3a (ENC) represents bias based on lowest values due to ENC's inverse probabilistic nature; M2 and PB1-F2 exhibit the greatest bias respective to all viruses. Whereas SCUO and Volatility (3b-c) identify a biased gene based on higher values; therefore, 3b (SCUO) shows that M2 and PB1-F2 are highly biased compared to all four viruses but 3c (Volatility) only displays bias for the PB1-F2 gene compared to the trend lines.

3.4 Association of ENC and GC Content

GC content has previously been shown to correlate strongly with the effective number of codons (ENC) [3]. Figure 3 depicts the relationship between GC3 content and ENC across the avian, human, swine, and 2009 H1N1 viruses. As seen in Figure 3, the avian viruses follow a distinctly linear trend line, whereas a virus from either human or swine origin follow a trend line that is non-linear. Also, the 2009 H1N1 viruses follow a trend line that is most similar to the human viruses. Since the 2009 H1N1 pandemic is considered to be a reassortment from avian, human, and swine viruses [22], this shows that the ENC and GC3 content in Influenza A depends more heavily on the infected host than on the virus' recent genetic history.

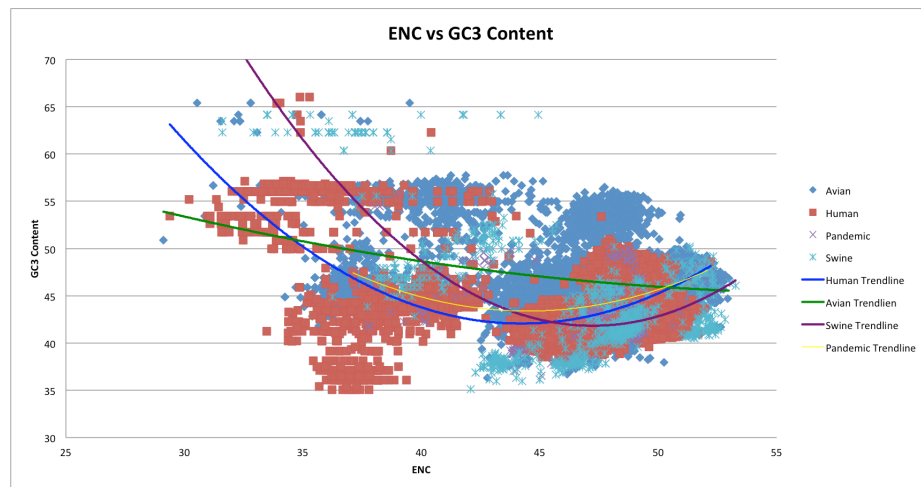


Figure 3 ENC vs GC3 Content of the four viruses

ENC values plotted against the GC3 content values from the four viral types. Trend lines were superimposed on the graph for each of the four viruses to exhibit correlations between the separate viral species. The avian viruses clearly show a linear trend, whereas the other three viruses, human, 2009 H1N1 pandemic, and swine, show a distinctly non-linear trend. Also, a close relationship can be seen between the human and 2009 H1N1 pandemic trend lines, inferring that the viral-host relationship in GC3 and ENC play a more important role than a genes recent genetic history.

3.5 Dinucleotide Odds Ratio

The Odds Ratio of the 16 possible dinucleotides revealed that at least 50% (depending on the dataset) were outside of normal range ([0.78,1.25]) [21] and can be classified as under- or over- represented. The extremities of these under- and over-represented dinucleotides are reported in Table 3, through which several distinct patterns that can be observed. Across the four viruses and three hosts, the CG and TA doublets have the lowest Odds Ratio. Also, CG, which has the lowest Odds Ratio in the human, 2009 H1N1 viruses, and all hosts, has been widely shown to be severely under-represented in vertebrate genomes as well as in viruses infecting vertebrate hosts [14,15]. It is interesting to note that while the most under-represented dinucleotide in all hosts is CG (by large range), it is only the most under-represented in the viruses infecting humans, whereas both the avian and swine viruses are more TA under-represented (swine by a large extent). In all cases the most severely over-represented dinucleotide is TG.

The CG suppression in vertebrate genomes is attributed to the methylation that occurs in vertebrate hosts, in which the cytosine nucleotide is often methylated into thymine [15]. This process often results in a mutation of the dinucleotide CG into TG and it is suggested that vertebrates have a tendency to use the duplets TG and CA with higher frequency than CG and TA to avoid the mutations that can occur during the methylation process. Both of these scenarios would also explain the high occurrences of TG and CA across every virus and host sequence.

Table 3 Highest and Lowest Common Dinucleotides

Host		Low	2nd Low	2nd High	High
Human	Dinucleotide	CG	TA	CA	TG
	Value	0.60	0.71	1.78	2.08
Avian	Dinucleotide	TA	CG	TC	TG
	Value	0.63	0.64	1.76	2.04
Swine	Dinucleotide	TA	CG	CA	TG
	Value	0.66	0.76	1.72	2.04
2009 H1N1	Dinucleotide	CG	TA	TC	TG
	Value	0.66	0.68	1.65	1.82
Human Host	Dinucleotide	CG	TA	CC	TG
	Value	0.54	0.78	1.61	1.71
Avian Host	Dinucleotide	CG	TA	CT	TG
	Value	0.59	0.74	1.62	1.72
Swine Host	Dinucleotide	CG	TA	CT	TG
	Value	0.61	0.74	1.63	1.78

The two highest and lowest dinucleotides based on Odds Ratio values.

3.6 RSCU and Odds Ratio

Table 4 is an ordered list of the mean RSCU values for each of our four viral and 3 host sets. The codons for each set of sequences are ordered by the mean RSCU value from highest to lowest. Select codon and value pairs are highlighted based on dinucleotide content. The codons that are highlighted in red contain the least found dinucleotides, and the highlighted green codons contain the second least common dinucleotide, as represented in Table 3. Likewise, most frequent dinucleotides from Table 3 were highlighted in orange, and the second most frequent in purple.

As seen in Table 4, codons that are highlighted as containing the most frequent dinucleotides tend to correspond with higher RSCU values. Contrastingly, the least common dinucleotides are most generally contained in the codons that have the lowest overall RSCU values (the lower portion of Table 4). This reveals a very strong correlation between the dinucleotide content in a gene and the RSCU values of codons within that same gene. Further research on this correlation is warranted.

Table 4. Ordered list (high to low) of 59 non-trivial codons in Influenza and hosts by RSCU values

Human		2009 H1N1		Swine		Avian		Human Host		Swine Host		Avian Host	
Codon	Value	Codon	Value	Codon	Value	Codon	Value	Codon	Value	Codon	Value	Codon	Value
AGA	2.928	AGA	2.916	AGA	2.852	AGA	2.456	AGA	1.947	CTG	1.791	CTG	1.911
GCA	1.840	GCA	1.82	CCA	1.791	GGA	1.926	AGG	1.666	AGA	1.668	AGA	1.7566
GGA	1.718	ACA	1.794	ACA	1.758	GCA	1.813	CTG	1.612	AGG	1.629	GTG	1.581
ACA	1.650	GGA	1.793	GGA	1.615	ACA	1.792	GTG	1.458	GTG	1.557	AGG	1.4835
AGG	1.635	CCA	1.647	GCA	1.608	AGG	1.685	ACA	1.33	GCC	1.339	AGC	1.4313
TCA	1.599	TCA	1.484	TTG	1.422	GTG	1.683	GCC	1.271	AGC	1.294	ACA	1.338
GTG	1.567	CCT	1.421	AGG	1.380	CCA	1.646	AGC	1.258	CAG	1.264	CAG	1.2743
CCA	1.463	AGG	1.367	TCA	1.365	CCT	1.434	GGA	1.213	ACC	1.262	GGA	1.2119
CCT	1.437	GTG	1.310	GTG	1.355	TGC	1.343	CCA	1.206	TCC	1.247	GCT	1.1945
TTG	1.366	CAT	1.31	CTT	1.348	TCA	1.329	CAG	1.205	ACA	1.213	GCA	1.1909
AAA	1.359	CTT	1.256	CCT	1.344	TTG	1.291	ATT	1.200	CTC	1.18	CCA	1.1866
GCT	1.349	AAA	1.208	AAA	1.33	CTG	1.283	TCT	1.181	AAA	1.181	TCT	1.136
CTT	1.343	AGT	1.204	GCT	1.294	AGT	1.27	CCT	1.178	CCC	1.178	CCT	1.1322
GGG	1.285	GTA	1.20	TGC	1.293	CTT	1.23	ACC	1.171	ATC	1.176	ATT	1.1262
TGC	1.266	GAA	1.18	CAT	1.253	AAA	1.216	TCC	1.162	GGA	1.152	ATC	1.1231
AGT	1.248	ATA	1.164	GAT	1.234	CAT	1.176	TTT	1.161	CCT	1.142	TCC	1.1097
ACT	1.242	TTG	1.156	AGC	1.228	TAC	1.162	GCT	1.143	GCT	1.138	CCC	1.1030
GAA	1.215	TGC	1.154	GGG	1.206	GAT	1.151	AAA	1.138	TCT	1.138	TTT	1.0980
TAC	1.197	AAT	1.146	ACT	1.193	GAA	1.147	TCA	1.131	GGC	1.120	GCC	1.0972
TAT	1.195	CTA	1.142	GAA	1.186	GGG	1.142	CCC	1.127	CCA	1.119	TGC	1.0750
GAT	1.188	AGC	1.126	TAC	1.181	TTC	1.1	TAT	1.102	ATT	1.096	ACC	1.0743
ATA	1.16	GCT	1.120	TCT	1.175	ACT	1.132	GCA	1.100	TTT	1.074	GGC	1.0711
GCC	1.154	GGG	1.087	ATT	1.170	AGC	1.121	AAT	1.079	GAC	1.067	CAC	1.0609
CAT	1.146	CAG	1.083	AAT	1.165	GCT	1.117	ACT	1.079	GAG	1.055	ACT	1.0592
CTG	1.135	TAC	1.077	AGT	1.141	ATT	1.10	CTC	1.072	CAC	1.051	TCA	1.0498
AAT	1.128	GAT	1.064	GTT	1.140	CAG	1.101	GGC	1.071	TGC	1.047	GAG	1.0463
CAA	1.080	TTC	1.042	GCC	1.128	ATA	1.095	CTT	1.038	GGG	1.0	AAA	1.0446
TTC	1.078	TAT	1.038	ATA	1.117	CTC	1.094	GGG	1.029	TCA	1.039	TAC	1.0367
ATT	1.073	ATT	1.032	CAA	1.086	AAT	1.069	CAC	1.003	AAC	1.021	GAT	1.034
CAC	1.065	ACT	1.03	GTA	1.054	TCT	1.06	GAA	1.003	TAT	1.019	CTC	1.0277
AGC	1.049	TCT	1.011	CTA	1.050	TAT	1.039	TGT	1.00	ACT	1.008	AAC	1.0244
TCT	1.046	CTG	0.976	TTT	1.030	AAC	1.020	GAC	1.003	GCA	0.992	GGG	1.0112
GTT	1.038	TTT	0.958	CAG	1.01	CAA	0.99	TGC	1.000	CTT	0.989	AAT	0.9756
CAG	1.011	GCC	0.947	TAT	1.003	GTC	0.984	GAT	0.997	TAC	0.981	CTT	0.9677
TTT	1.006	ACC	0.940	CTG	1.001	CAC	0.981	GAG	0.997	AAT	0.979	GAC	0.966
ACC	0.998	GAC	0.937	TTC	0.994	TCC	0.974	CAT	0.997	TGT	0.953	TAT	0.9633
CTA	0.983	CAA	0.921	ACC	0.969	ACC	0.940	GTT	0.974	CAT	0.949	GTT	0.960
TCC	0.981	GTC	0.919	CAC	0.948	GCC	0.936	ATC	0.972	GTC	0.95	AAG	0.9554
GTA	0.956	CAC	0.910	TCC	0.9	TTT	0.92	TTG	0.957	GAA	0.945	GAA	0.9537
CTC	0.911	ATC	0.893	AAC	0.944	CCC	0.923	AGT	0.944	GAT	0.933	TTG	0.9511
AAC	0.909	CTC	0.884	CTC	0.938	ATC	0.914	AAC	0.921	TTC	0.926	CAT	0.9391
CCC	0.904	GTT	0.878	TGT	0.9	CTA	0.908	TAC	0.898	GTT	0.913	AGT	0.9261
ATC	0.874	AAG	0.88	CCC	0.899	TGT	0.899	GTC	0.897	CGG	0.902	TGT	0.9250
TGT	0.859	TGT	0.86	GAG	0.874	GTT	0.891	AAG	0.862	TTG	0.902	TTC	0.9020
GAC	0.835	AAC	0.857	ATC	0.852	GAC	0.860	TTC	0.839	AGT	0.873	CGG	0.8862
GTC	0.834	CCC	0.857	GTC	0.848	AAG	0.860	ATA	0.829	AAG	0.819	CGC	0.8223
GAG	0.828	CGA	0.838	GAC	0.848	GAG	0.857	CAA	0.795	CGC	0.754	GTC	0.8180
TTA	0.736	GAG	0.821	GGT	0.83	GTA	0.814	CGG	0.778	CAA	0.736	ATA	0.7507
CGA	0.718	TCC	0.789	CGA	0.821	CGG	0.690	TTA	0.743	ATA	0.728	CAA	0.7257
CCG	0.67	GGT	0.696	TTA	0.793	CGA	0.658	GGT	0.688	GGT	0.684	GGT	0.7058
AAG	0.661	TTA	0.685	AAG	0.736	GGT	0.653	CGC	0.680	TTA	0.610	TTA	0.6484
GGC	0.638	CCG	0.666	ACG	0.670	GGC	0.615	GTA	0.671	GTA	0.582	GTA	0.6405
CGG	0.588	GGC	0.622	CCG	0.626	ACG	0.602	CTA	0.579	CCG	0.561	CCG	0.5781
GGT	0.563	CGG	0.581	CGG	0.602	CCG	0.596	CGA	0.496	CGA	0.552	CGT	0.5366
ACG	0.531	ACG	0.533	TCG	0.585	TTA	0.553	CCG	0.488	GCG	0.53	ACG	0.5283
CGC	0.521	CGC	0.481	GCG	0.541	TCG	0.551	GCG	0.486	CTA	0.524	GCG	0.5174
TCG	0.426	TCG	0.453	CGC	0.511	CGT	0.540	CGT	0.432	ACG	0.517	CGA	0.5148
CGT	0.368	GCG	0.437	GGC	0.498	CGC	0.459	ACG	0.420	CGT	0.496	CTA	0.4937
GCG	0.368	CGT	0.433	CGT	0.481	GCG	0.405	TCG	0.324	TCG	0.408	TCG	0.3474

The lower RSCU values (those below 1.0) were shaded to partition the results. Codons containing the least abundant dinucleotide (as specified in Table 3) are highlighted in red; codons containing the second least abundant dinucleotides are highlighted in green. Likewise the most abundant and second most abundant dinucleotide containing codons are highlighted in orange and purple respectively.

4. Conclusion

Our analysis revealed several new and noteworthy findings relating to codon bias in Influenza A and their hosts. We found that the GC content in the genes of Influenza A viruses varies in a predictable manner dependent on the species of the infected host. In particular, the standard deviation of the three single-position GC contents across the 11 genes of avian, human, or swine viruses will uniquely identify its host. Also, as with the Flaviviridae virus [15], there is a strong correlation between a codon's RSCU value of Influenza viruses and the Odds Ratio of the dinucleotides contained within that codon. Relatedly, the avian, human, 2009 H1N1, and swine viruses exhibit a similar trend to that of their hosts in both the RSCU and Odds Ratio measurements. Specifically, the Odds Ratio values in all seven datasets (avian, human, pandemic, and swine viruses as well as the avian, human, and swine hosts) show both CG and TA as being severely under-represented whereas TG is most over-represented. This phenomenon was previously known within the avian and human viruses [14] but has not previously been seen within the swine virus. Interestingly, the swine and avian viruses both follow a distinctly different pattern than that of human viruses; the avian and swine viruses are most severely TA under-represented whereas the five other sets are CG under-represented. Finally, out of the six considered codon bias measurements, only the Odds Ratio and RSCU values of the avian, human, and swine host mRNA transcripts are distinct. However, the GC content, ENC, SCUO, and Codon Volatilities of these hosts are in close ranges.

In summary, we verified the positive correlation between the Odds-Ratio and RSCU measurements in a much larger dataset, including new species, than in previous research. Surprisingly, a new pattern was found in the deviation of GC content across hosts. Furthermore, a positive correlation was identified between the Odds-Ratio and RSCU measurements of the viruses a correlation that was previously observed only in Flaviviridae viruses.

Works Cited

- [1] Smith, G., & et al. (2009). Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* , 459, 1122-1125.
- [2] Dawood, F., & et al. (2009). Emergence of a novel swine-origin influenza A (H1N1) virus in humans. *N Engl J Med* , 360 (25), 2605-2615.
- [3] Zhou, T., Gu, W., Ma, J., Sun, X., & Lu, Z. (2005). Analysis of synonymous codon usage in H5N1 virus and other influenza A viruses. *BioSystems* , 81, 77-86.
- [4] Suzuki, H., Brown, C. J., Forney, L. J., & Top, E. M. (2008). Comparison of Correspondence Analysis Methods for Synonymous Codon Usage in Bacteria. *DNA Research* , 15, 357-365.
- [5] Zeeberg, B. (2002). Shannon Information Theoretic Computation of Synonymous Codon Usage Biases in Coding Regions of Human and Mouse Genomes. *Genome Res.* , 12, 944-955.
- [6] Wan, X., Xu, D., & Zhou, J. (2003). A New Informatics Method for Measuring Synonymous Codon Usage Bias. *Intelligent engineering systems through artificial neural networks* , 13, 1101-1108.
- [7] Wan, X., Xu, D., & Zhou, J. (2006). CodonO: a new informatics method for measuring synonymous codon usage bias within and across genomes. *International Journal of General Systems* , 35 (1), 109-125.
- [8] Wan, X.-F., Xu, D., Kleinhofs, A., & Zhou, J. (2004). Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes. *BMC Evolutionary Biology* , 4, 19.
- [9] Plotkin, J. B., & Dushoff, J. (2003). Codon bias and frequency-dependent selection on the hemagglutinin epitopes of influenza A virus. *PNAS* , 100 (12), 7152-7157.
- [10] Plotkin, J. B., Dushoff, J., & Fraser, H. B. (2004). Detecting selection using a single genome sequence of *M. tuberculosis* and *P. falciparum*. *NATURE* , 428, 942-945.
- [11] Zhang, J. (2004). On the evolution of codon volatility. *Genetics* , 169, 495-501
- [12] Burge, C., Campbell, A. M., & Karlin, S. (1992). Over- and under-representation of short oligonucleotides in DNA sequences. *Proc. Natl. Acad. Sci. USA* , 89, 1358-1362.

- [13] Arndt, P. F., Burge, C. B., & HWA, T. (2003). DNA Sequence Evolution with Neighbor-Dependent Mutation. *Journal of Computational Biology* , 10 (3-4), 313-322.
- [14] Greenbaum, B. D., Levine, A. J., Bhanot, G., & Rabadan, R. (2008). Patterns of Evolution and Host Gene Mimicry in Influenza and Other RNA Viruses. *PLoS Pathogens* , 4 (6).
- [15] Lobo, F. P., Mota, B. E., Pena, S. D., Azevedo, V., Macedo, A. M., Tauch, A., et al. (2009). Virus-Host Coevolution: Common Patterns of Nucleotide Motif Usage in Flaviviridae and Their Hosts. *PLoS ONE* , 4 (7).
- [16] Bao, Y., Bolotov, P., Dernovoy, D., Kiryutin, B., Zaslavsky, L., & et al. (2008). The Influenza Virus Resource at the National Center for Biotechnology Information. *J Virol* , 82 (2), 596-601.
- [17] Katoh, K., Misawa, K., Kuma, K.-i., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* , 30 (14), 3059-3066.
- [18] Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* , 33, D501-D504.
- [19] Qiu, S., Bergero, R., Zeng, K., Charlesworth, D. (2011). Patterns of codon usage bias in *Silene latifolia*. *Molecular Biological Evolution*, Jan 28(1) , 771-780.
- [20] Wright, F. (1990). The 'effective number of codons' used in a gene. *Gene* , 87, 23-29.
- [21] Karlin, S., Doerfler, W., & Cardon, L. (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, 2889-2897.
- [22] Alexander Solovyov, Benjamin Greenbaum, Gustavo Palacios, W. Ian Lipkin, and Raul Rabadan. Host Dependent Evolutionary Patterns and the Origin of 2009 H1N1 Pandemic Influenza. Version 3. *PLoS Curr.* 2010 January 30 [revised 2010 August 4]; 2: RRN1147.doi:10.1371/currents.