Graphical Representation and Numerical Characterization of H5N1 Avian Flu Neuraminidase Gene Sequence†

Ashesh Nandy,‡ Subhash C. Basak,* and Brian D. Gute

University of Minnesota Duluth, Natural Resources Research Institute, 5013 Miller Trunk Highway, Duluth, Minnesota 55811

Received December 15, 2006

The high degree of virulence and potential for development of a pandemic strain of the H5N1 avian flu has resulted in wide interest in characterization of the various genes of the H5N1 virus genome. We have considered for our analysis all 173 available complete sequences, as of February 2006, of the neuraminidase gene, which is the target of the most effective treatment regimen comprising the inhibitors oseltamivir and zanamivir. We have used a 2D graphical representation of the neuraminidase RNA sequences of H5N1 strains to identify a few distinct structural motifs. The H5N1 strains were split into two main classes: strains that were benign to human beings in the years up to 1996 and the period 1999−2002 and strains that were highly pathogenic to humans in the periods 1997 and 2003 to present. Comparisons with earlier H1N1 pandemic and epidemic strains have also been made to understand the current status of the gene. Our findings indicate that the base composition and distribution patterns are significantly different in the two periods, and this may be of interest in studying mutational changes in such viral genes.

1. INTRODUCTION

The recent epidemic of the highly pathogenic H5N1 avian flu that has also caused the death of over 130 humans has raised concerns of a potential pandemic among human populations through possible mutations and implications for their control.[1,2] Numerous studies have shown how rapid mutations in the H5N1 strains and reassortments among the various subtypes of the flu viruses cocirculating in the avian population have led to the growth of many genotypes of the H5N1 subtype of avian flu.[3] While the evidence to date suggests that the H5N1 flu is still spread through close contact with diseased birds, and one case in Indonesia of six deaths in a family of seven remains under investigation by the World Health Organization,[4] many scientists believe that it is only a matter of time before a mutation or reassortment with a human flu virus will result in a strain that can be passed from human to human, leading to a pandemic situation.

Such a possibility has rendered it important to study and characterize the various genes of the H5N1 virus genome. Viral genomes are known to mutate very rapidly, so that different strains evolve over relatively short periods of time through genetic drift.[3] While many authors have studied the evolutionary characteristics of the different strains of the H5N1 avian flu virus from their gene and protein sequences by constructing phylogenetic trees and focusing on the role of specific amino acids in the sequences,[5] we approach the issue from the viewpoint of the base composition and distribution characteristics in the family of the avian flu genes to determine any systematic differences that may arise from the mutational changes. Studies of systematics in gene sequences have been attempted before, albeit in different contexts. Peng et al.[6] considered the arrangements of purine and pyrimidine bases in selected sequences and demonstrated the possible existence of long-range correlations, and chaos game representation diagrams proposed by Jeffrey[7] had shown the fractal nature inherent in gene sequences, while Voss[8] determined characteristic lengths from consideration of the 1/f noise spectrum in families of DNA sequences.

Our choice of the H5N1 gene sequence and the tools for analysis is based on the following considerations. While the hemagglutinin of the H5N1 virus is known to be the agent that cleaves the cell barrier to enable the infection, neuraminidase (NA) is required for the spread of the infection. Since the only effective treatment regimen to date against the H5N1 bird flu is provided by inhibitors of the NA enzyme,[9] studies of this gene are of particular importance. To characterize the effects of mutational changes on base composition and distribution of the neuraminidase gene sequence, we use the recent theoretical tool of numerical characterization of gene sequences.[10] From a study of all the complete sequences of the neuraminidase genes available in the GENBANK database, we report our finding of significant differences between the neuraminidase genes that are part of the H5N1 strains that are apparently highly pathogenic to human beings and those that are not. We have also studied the earlier H1N1 pandemic and epidemic strains to compare the status of the current type of neuraminidase genes with the earlier strains and conclude that the neuraminidase sequences of the currently prevalent H5N1 strains have significant differences with the earlier neuraminidase strains. We report also a small segment of the neuraminidase which we have found to be strongly conserved among most of the neuraminidase N1 sequences irrespective of the hemagglutinin type.

† Dedicated to Professor Nenad Trinajstić on the occasion of his 70th birthday.

* Corresponding author phone: +1 218 720 4230; fax: +1 218 720 4328; e-mail: sbasak@nrri.umn.edu.

‡ On leave from the Programme in Environmental Science, Jadavpur University, Jadavpur, Kolkata 700032, India.

10.1021/ci600558w CCC: $37.00
Published on Web 04/07/2007

2. MATERIALS AND METHODS

2.1. Materials. We selected from GENBANK all sequences of the H5N1 neuraminidase genes which were listed as having a complete coding sequence as of February 7, 2006. This resulted in a database of 173 sequences comprising 29 strains with full-length stalks of the neuraminidase and 144 with short-length stalks [eight strains with 19 amino acid (aa) deletions and 136 strains with 20 aa deletions in the stalk region]. There was one human isolate in the 29 full-length strains (A/Hong Kong/213/03) and 21 human isolates in the short-length strains. The difference in sample sizes of the long and short stalk strains is reflective of the fact that the genotypes of the H5N1 with a short-length stalk in the neuraminidase have dominated over the varieties with the long-length stalks.[3] The selections covered the year 1959 (one sample from the United Kingdom—A/chicken/Scotland/59) and the period 1996 to 2005, when severe outbreaks of the H5N1 subtype of influenza were observed among avian populations, and also infections in humans in 1997, and again from 2003 to 2005.[11]

For comparative purposes, we also selected sample sequences of neuraminidase genes for the influenza A H1N1 subtype corresponding to human isolates from earlier pandemic and severe influenza years. These comprised the 1918 Brevig_Mission strain and the Weiss/43, Leningrad/54/1, Denver/57, India/80, Chile/1/83, and USSR/90/77 strains.

2.2. Methods. To compare the mutational changes among the various strains of the H5N1 neuraminidase gene, we adopt the graphical representation method for visual clues and associated numerical characterization descriptors to get quantitative comparisons. Several authors have developed different techniques for visual representation of DNA sequences, for example, Gates,[12] Nandy,[13] and Leong and Morganthaler[14] with their 2D Cartesian coordinates approach; Yau et al.[15] with a nondegenerate 2D representation; Randic et al.[16] with a 3D prescription and models with four and higher dimensions.[17,18] Many of these prescriptions for graphical representation have been further analyzed to yield numerical characterization of DNA/RNA sequences, whose utility, especially in estimating similarities/dissimilarities between sequences, have recently been elaborated through several applications.[10,16−20] On the basis of the recent review we have conducted on the merits and demerits of these techniques,[10] and because of the large size of the sequences we are interested in here, we use Nandy's 2D graphical representation method[13] for visual cues and the geometrical descriptors derived therefrom[21] for the comparative analysis.

In this method, we generate a graphical representation of a DNA/RNA sequence by plotting one point for each base on a 2D Cartesian axes system by the algorithm: Move one step in the negative x direction for an adenine (a), one step in the positive y direction for a cytosine (c), one step in the positive x direction for a guanine (g), and one step in the negative y direction for a thymine (t). A succession of such points generates a walk whose trace on the graph provides information on the local and global base distribution patterns in the sequence.

For comparative analysis, use numerical characterization methods to calculate indexes for each segment of the gene.

(a) We define a base composition index n_R as $n_R \equiv (a + t)/(c + g)$ where the a, t, c, and g are the numbers of the individual bases comprising the segment under consideration.

(b) We next define an index to provide a numerical estimate of how the four bases are distributed in the sequence. From the graphical representation, we calculate the first-order moments, μ_x and μ_y, and graph radius, g_R,[21]

$$\mu_x = \frac{\sum x_i}{N}, \quad \mu_y = \frac{\sum y_i}{N}, \quad \text{and } g_R = \sqrt{\mu_x^2 + \mu_y^2}$$

where x_i and y_i represent the coordinates of each point on the plot and N is the total number of the bases in the segment. The g_R represents the base distribution index.

Graph radius is a sensitive measure of sequence composition and distribution,[21−23] and its values change with the type of mutations and their location in the sequence. g_R is especially useful in comparing sequences of equal length.[24] This is the case for our NA genes: eight of our short-stalk NAs each have 1353 bases; 136 other short-stalk NAs each have 1350 bases, and all 29 of the long-stalk length strains each have 1410 bases.

(c) g_R can be used to compare the similarity/dissimilarity of two or more DNA or RNA sequences. We use the moments defined above to define a distance metric between two DNA/RNA sequences, arbitrarily referred to as 1 and 2, as

$$\Delta g_R = \sqrt{(\mu_{x1} - \mu_{x2})^2 + (\mu_{y1} - \mu_{y2})^2}$$

The smaller the value of Δg_R, the more similar the two sequences; conversely, greater values of Δg_R indicate that the sequences are more dissimilar. Such a description has been shown to approximately correlate to evolutionary distances between various species.[21] Several authors have proposed different methods for measuring the degree of similarity/dissimilarity between DNA sequences based upon graphical approaches,[16−20] but for our purposes, the simple metric mentioned above will suffice. For comparison, we note that Suarez et al.[5] determined from phylogenetic analysis that A/Chicken/Hong Kong/220/97(H5N1) and A/Hong Kong/156/97 (H5N1) had >99% nucleotide sequence homology in pairwise comparisons for all eight H5N1 genes. Our similarity score for the neuraminidase gene sequence distance between the two is 2.54, implying that the two sequences are very similar. In contrast, the distance is 17.2 when A/Chicken/Hong Kong/220/97(H5N1) is compared with a strain from 5 years later, A/chicken/China/1/02-(H5N1), implying a large degree of difference in base arrangements. We note in passing that, since the internal chemical structure of the sequences is not used in any way, we plot the graphs and calculate the moments in this 2D method reading the sequence from the 5′ end as a matter of convention.

3. RESULTS AND DISCUSSION

Figure 1 parts a and b show plots of the H5N1 neuraminidase gene sequence for strains A/duck/Guangdong/07/2000 (GENBANK accession number AY585404), which has a long stalk, and A/Hanoi/30408/2005 (GENBANK accession number AB239126), the latter having a stalk that is shorter by 60 bases. Sections of the RNA sequences in each case can be cross-identified with four aspects of the neuraminidase

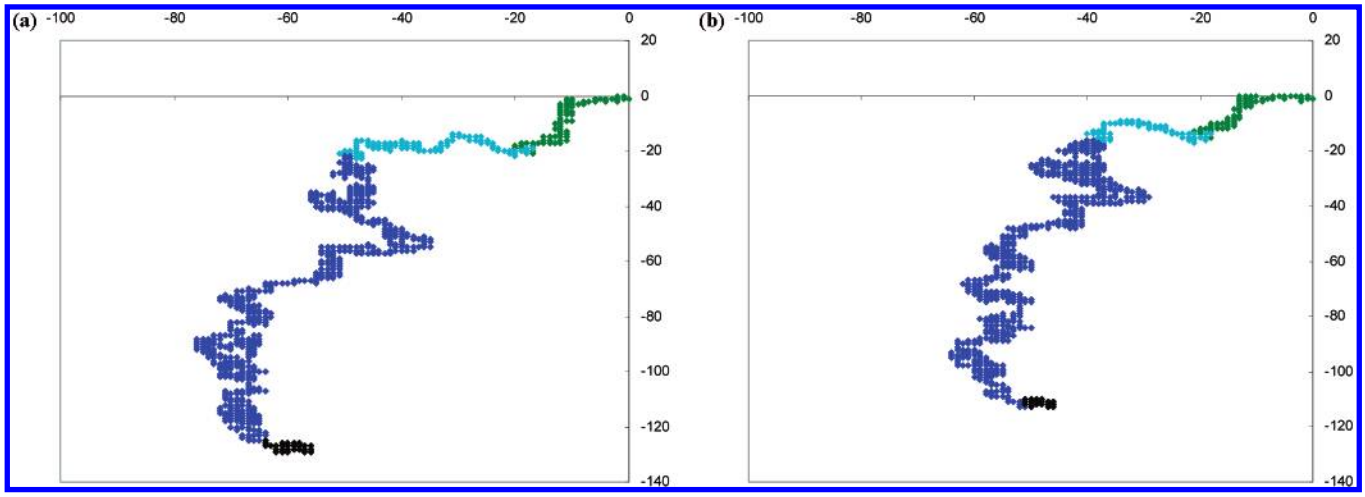

Figure 1. (a) 2D graphical representation of the neuraminidase gene sequence of the H5N1 virus strain A/duck/Guangdong/07/2000 (GENBANK accession number AY585404). (b) 2D graphical representation of the neuraminidase gene sequence of the H5N1 virus strain A/Hanoi/30408/2005 (GENBANK accession number AB239126). The axes for the associated walk are as shown. The different parts that correspond with the protein structure of the gene are marked as A, transmembrane (olive green); B, stalk (light blue); C, head (mauve); and D, head (black). In this paper, region C is referred to as the body and region D as the tail of the neuraminidase sequence.

protein: the first 105 bases correspond with the transmembrane region as identified in the Universal Protein Resource (UNIPROT) protein database;[25] the next section corresponds with the stalk region which can range from 105 bases for short-stalk genes to 165 bases for full-length genes. The remaining 1140 bases relate to the head of the neuraminidase. For our purposes, we have subdivided the head region into two parts: a 1090-base segment we refer to as the body and a 50-base segment at the 3′ end designated as the tail. The four segments are identified by colors in Figure 1a and b as regions A, B, C, and D, respectively.

Computation of the base composition ratio, n_R, and base distribution index (or the graph radius), g_R, for all regions was carried out on all 173 strains of the virus. Because the differences in the total length of the genes depends upon the length of the stalk segments only, we classified the 173 viral strains into 144 short-stalk strains, comprising eight neuraminidase genes with 108 nucleotide stalk lengths and 136 neuraminidases with 105 nucleotide stalk lengths, and 29 long-stalk genes where the stalks all contained 165 bases. Since the H5N1 infections in humans were first observed in 1997 in Hong Kong and then there were no further cases reported until the virus crossed the species barrier again in 2003 to infect humans,[11] we have used the isolates from these blocks of years to demarcate strains which are pathogenic and nonpathogenic to humans. Thus, for purposes of our analysis, we subdivided the selected neuraminidase strains into the following groups: Group 1 included the 18 complete short-stalk strains available for the years 1999−2002 when humans were not affected. Group 2 comprised 126 complete short-stalk strains from the other years, and group 3 comprised all 29 complete long-stalk strains.

Group 2 included 21 strains of the H5N1 virus isolated from human hosts in 1997 and 2003−2005 in Hong Kong, Viet Nam, and Thailand and 105 strains from other mammalian and avian hosts from the same years. The data for the long-stalk neuraminidase strains for humans are sparse. There is only one complete neuraminidase isolate available from a human host (A/Hong Kong/213/03 [H5N1]) for the long-stalk variety; with the remaining 28 long-stalk strains showing very close similarities in base composition and distribution numbers, it was considered best to include all 29 strains into one group. (Until May 17, 2006, there were no other complete sequences for long-stalk H5N1 neuraminidase genes from human hosts.)

One of the reasons for defining the groups as above is that viruses are known to mutate very rapidly, and sample strains from one period of time to another may differ markedly. In the current example, we also have evidence that in some time periods the H5N1 strains appear to have been especially pathogenic to humans. While calendar years may not be the best device to demarcate the different groups as we have done above, it is a first approximation in the absence of any other reliable yardstick and can be justified if the strains within these time periods do indeed display measurable differences. To check on this aspect, we have taken a sample strain from group 1 (A/chicken/China/1/02 [H5N1]) and measured the distances of all other strains from it using our distance metric, Δg_R, defined earlier. Figure 2 shows the number of strains plotted as column graphs against distance intervals. While there is some overlap, as is to be expected in our approximation (also demonstrated specifically later in this section), it is clear that all strains of group 1 are more similar to each other, group 2 strains are generally well-separated from the group 1 strains, and the group 3 strains, characterized by the longer stalk length, are farthest from, and therefore most dissimilar to, those of group 1 and are also well-separated from those of group 2. Thus, we can proceed to analyze the H5N1 strains on the basis of this classification.

The results of the analyses for n_R and g_R are shown in Tables 1 and 2. We find that two segments of the neuraminidase show a similar range of mutations irrespective of groups: the segment identified with the transmembrane region of the neuraminidase protein, region A in Figure 1, shows only small variations among the three groups. A single-factor ANOVA determines that, at the 95% confidence level, significant differences exist in the n_R values (the a + t/c + g ratios) between the groups (P value <0.0001); in the case of the g_R values, there is also a significant difference at the same confidence level, but with a P value ~ 0.0004

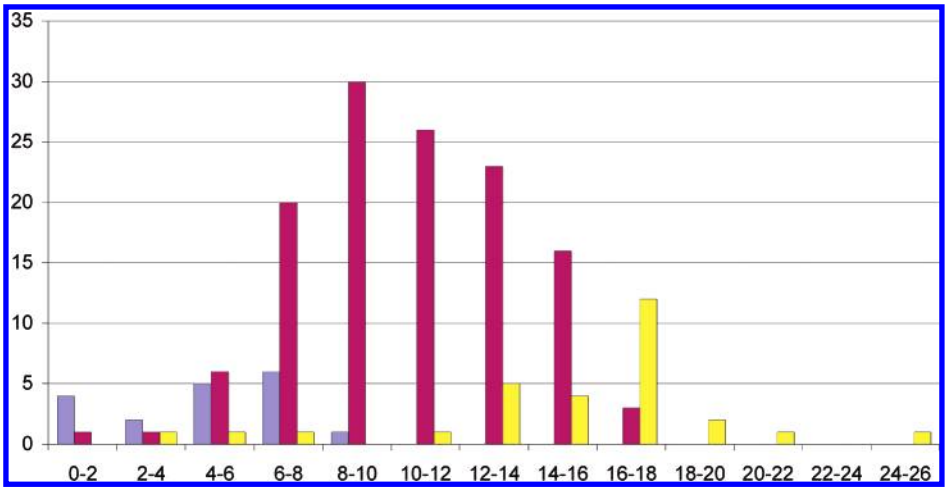

Figure 2. Column chart of number of H5N1 strains plotted against distance intervals where the distance between two strains is determined in terms of Δg_R. The blue columns refer to group 1 H5N1 strains, the brown columns to group 2 H5N1 strains, and the yellow columns to group 3 H5N1 strains. The sequences are from region C of the neuraminidase as described in the text.

Table 1. Base Composition Index (n_R) Values for H5N1 Neuraminidase Genes

	sequences	group 1	group 2	group 3
region	no. of strains	18	126	29
A	transmembrane	2.00 ± 0.05	1.92. ± 0.02	2.10 ± 0.04
B	stalk	1.47 ± 0.04	1.50 ± 0.02	1.62 ± 0.04
C	body	1.19 ± 0.01	1.23 ± 0.00	1.24 ± 0.01
D	tail	0.92 ± 0.01	0.92 ± 0.00	0.92 ± 0.01

Table 2. Base Distribution Index (g_R) Values for H5N1 Neuraminidase Genes

	sequences	group 1	group 2	group 3
region	no. of strains	18	126	29
A	transmembrane	14.41 ± 0.40	13.72 ± 0.15	14.24 ± 0.32
B	stalk	12.70 ± 0.62	12.21 ± 0.24	21.29 ± 0.49
C	body	33.18 ± 1.76	42.63 ± 0.67	48.31 ± 1.38
D	tail	6.04 ± 0.13	6.05 ± 0.05	5.97 ± 0.10

and an F statistic of 8 compared to an F critical value of 3.05.

The stalk region of the neuraminidase gene sequence (region B in Figure 1a and b), being of shorter length in groups 1 and 2 and around 60% longer for the strains collected under group 3, is expected to have different numbers for the two varieties. The g_R and n_R values of groups 1 and 2 comprising the short-stalk strains have a lower ratio of a + t to c + g compared with the long-stalk samples; n_R values are 1.48 for the short-stalk segments and 1.62 for the longer varieties. The base distribution index reflects a more significant change: ~12.5−21.2 from the short- to the long-stalk segments, due primarily to a longer run of adenines in the sequence of the long-stalk varieties (see Figure 1a). A single-factor ANOVA confirms the observations that at the 95% confidence level the short- and long-stalk segments are indeed significantly different. The difference between the groups in terms of the g_R values is significantly higher, indicating that the base distribution patterns in the case of the long-stalk segments are very different from the distribution patterns in the short-stalk segments, even though the a + t/c + g ratios are not as markedly different.

On the other hand, the 50-base-long region D at the 3′ end of the H5N1 neuraminidase gene sequences presents a completely different picture. The base composition and distribution pattern in this segment was found to be very strongly conserved across all host species and all groups; no differences could be determined through ANOVA. The purine/pyrimidine ratio was maintained at 25:25 with a variation of one base across all strains of the virus, and the a + t/c + g ratio was also conserved at 0.92 ± 0.01. The base distribution index, g_R, of the 50-base tail section was found to be constant at 6.032 346 for 150 of the 173 strains of H5N1 neuraminidase genes which we report here. A BLAST analysis of the tail segment of A/Ck/HK/2133.1/2003 (H5N1) (GENBANK: AY651466) which had a g_R of 6.032 346 showed that this RNA sequence segment was conserved across a wide range of the avian flu strains. Of the first 471 matches with significant homology, 301 were of the H5N1 subtype and 108 were of the H1N1 subtype; the balance included influenza A subtypes H3N1, H4N1, H6N1, H7N1, H9N1, H10N1, H11N1, and H12N1. Thus, this segment seems to be a well-conserved feature of the N1 gene and could be useful in devising H5N1 influenza antidotes or vaccines that would remain effective over many mutations.

Region C, comprising 1090 bases for all strains, shows a significant difference in both the a + t/c + g base composition index (n_R) and the base distribution index (g_R) values between group 1 and groups 2 and 3, that is, between the viral strains that were prevalent during the years when humans were unaffected and the years when human cases of H5N1 were identified. A single-factor ANOVA shows that the n_R ratios (a + t/c + g) as well as the base distribution index, g_R, values are significantly different at the 95% confidence level between group 1 and groups 2 and 3 with an F statistic of 58.7 and 88.0, respectively, for the two analyses performed. The ANOVA results are presented in graphical form in Figure 3, showing plots of the three group means with 95% confidence intervals. The two indices n_R and g_R together provide a novel result showing that the mutations within a segment of 1090 bases in the neuraminidase genes of the H5N1 circulating in the period 1999−2002 have been significantly different in their effect on base composition and distribution compared to the mutations in the strains of the other years. While genetic drift causes small differences in strains within each group, the observed differences between groups signify rather large genetic drifts

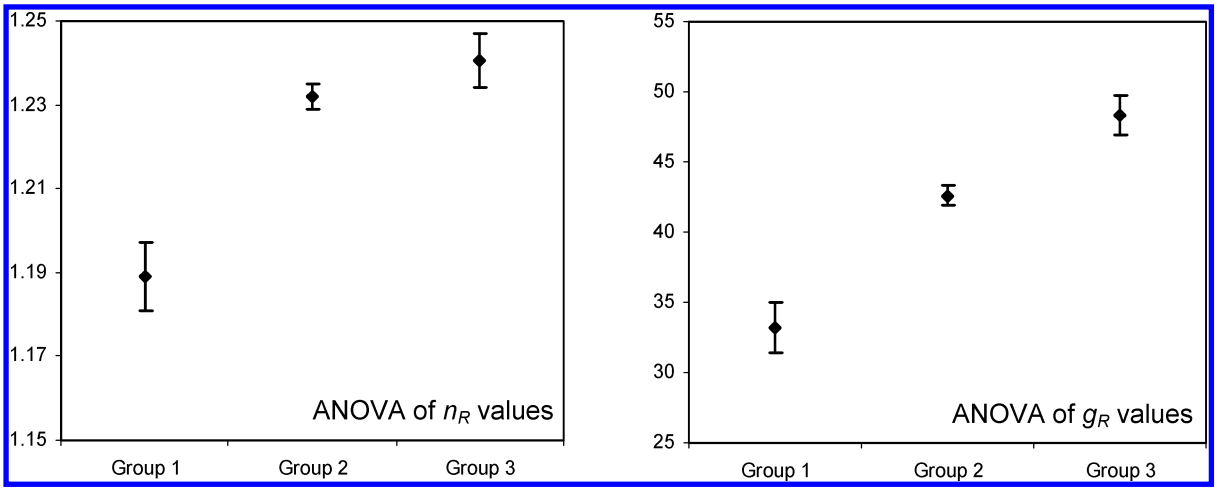

Figure 3. Single-factor ANOVA analysis of the n_R (left) and g_R (right) values of the C segment of three groups of H5N1 neuraminidase genes showing the mean and the 95% confidence interval for the mean for each group.

taking place within short time periods.

To gain a better understanding of the mutations taking place within the body region (region C as shown in Figure 1), a detailed comparison was conducted to examine base distribution and amino acid composition within a selected sample of H5N1 sequences. For a comparative standard, the group 1 sequence, A/chicken/China/1/02 [H5N1] (GENBANK DQ023147), was selected; this same sequence was used as the standard for the comparative frequency distribution chart (Figure 2). Focusing on isolates from the same species, namely, chicken, the three sequences chosen for comparative analyses were one of the earliest samples of the virus from group 2, A/Chicken/Hong Kong/220/97 [H5N1] (GENBANK AF046081); a representative sample from the median range of group 2, A/Ck/HK/YU324/2003 [H5N1] (GENBANK AY651478); and a representative sample from the median range of group 3, A/Chicken/Hong Kong/FY150/01 [H5N1] (GENBANK AY509095). In terms of the base mutations in these RNA sequences, it was found that the synonymous to asynonymous changes were close to the ratio 3:1. In terms of the amino acid changes compared to our standard sample, mutations in the other three selected sequences decreased the count of nonpolar (hydrophobic) amino acids, while the count of glycines remained constant or decreased. For the two group 2 samples (A/Ck/HK/YU324/2003 and A/Chicken/Hong Kong/220/97), the mutations decreased the count of the nonpolar amino acids by two and the glycine count decreased by three; in the group 3 sample, the charged group and nonpolar group counts were both reduced by 1. Thus, structural conformations of the neuraminidase body section tend toward lowering hydrophobic effects for the samples in the two groups we have considered, as compared to the sample from non-human-afflicted years. For the record, comparison with the only long-stalk human isolate available (A/Hong Kong/213/03 [H5N1]) again shows that the hydrophobic group count decreased and the charge group count increased while the glycine count remained unaltered as compared to the standard sample.

One striking difference between the two clusters of the short-stalk neuraminidase genes arises from our initial grouping, namely, group 1, which collects together viral strains from the period 1999−2002 when the H5N1 was benign to humans, and group 2, which is a collection of genes from viral strains of other periods during which human infections and deaths resulted from the H5N1 influenza. In region C, the two clusters have widely different values of the a + t/c + g ratio as well as the g_R values, showing that the mutation trends affected differently not only the base compositions but also the base distribution patterns.

Group 3, which includes strains of long-stalk neuraminidase genes from all years, shows little variation between the two periods; in fact, whereas the g_R values of the same 1090 segments for the group including all strains is 48.31 ± 1.38 (Table 2), the g_R for only the 1999−2002 period is 48.53 ± 1.53. The base composition index, n_R, shows a similar trend, implying that the base distribution and composition patterns in this segment of the long-stalk neuraminidase gene sequences do not change very much. Part of the reason that there are no apparent differences likely arises from the extreme paucity of data: of the 29 samples in total in group 3, there is only one from a human isolate and four from nonhuman isolates from periods outside 1999−2002. The fact that this genotype of the H5N1 viral strain seems to have greatly diminished in the wild as reported by Li et al.,[3] and that whatever data are available are for the most part incomplete in terms of the neuraminidase sequences, makes our analyses even more difficult.

In the case of group 2, which includes short-stalk neuraminidase sequences from human and nonhuman isolates for the periods 1997 and 2003−2005, that is, periods when there were cases of human beings infected by the H5N1 virus, a similar analysis shows that at the 95% confidence level the intervals marking out the nonhuman and human variants of the neuraminidase genes of this group overlap: the nonhuman range is from 1.22 to 1.23 for n_R values, and the interval for the human samples is from 1.23 to 1.24. Thus, our decision to combine all these samples into one group can be justified; however, the original decision to separate the H5N1 viral strains into periods when humans were affected and periods they were not still remains a crude first approximation, and there are samples included in the list by this criterion such as A/tree sparrow/Henan/1/2004 (GENBANK: AY41216), which has been found to be nonpathogenic to mice,[26] and perhaps therefore also to humans. We note that the n_R and g_R values for this sample are 1.21 and

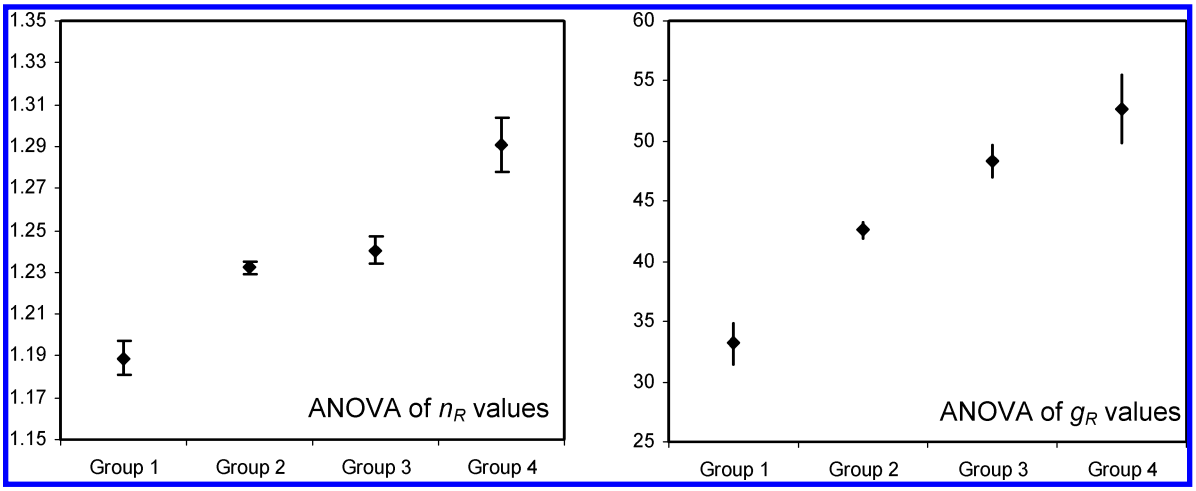

Figure 4. Single-factor ANOVA analysis of the n_R (left) and g_R (right) values of C segments of three groups of H5N1 neuraminidase genes and one group (group 4) of H1N1 neuraminidase genes showing the mean and the 95% confidence interval for the mean for each group.

37.4, respectively, placing it closer to group 1 (in fact, all three of the tree sparrow samples have similar values), but in the absence of complete information on all the viral strains, we are constrained to the kind of groupings used in this analysis.

To further determine the characteristics of the sequences of long-stalk neuraminidase genes to compare with the short-stalk varieties, we considered some of the H1N1 sequences from human hosts of previous years when influenza epidemics or pandemics were observed; the selections are given in the Materials and Methods section of this paper. We found that the human isolate H1N1 neuraminidase sequences are significantly different from all three previous groups in both n_R and g_R values for the transmembrane, stalk, and body regions. In the case of the 50-base region at the 3′ end of the sequences, the variation in the base composition is again minimal, showing only a one to two base difference with our H5N1 standard sequence mentioned earlier; however, the position of the mutations affects the g_R values, the average working out to 3.8 compared with 6.0 for the other three groups.

The 1090-base segment comprising the body of the neuraminidase gene sequence, region C as defined in this paper, of the H1N1 subtype viral genes reveals large, significant differences with the H5N1 strains, groups 1−3, in n_R and g_R values. Figure 4 shows plots of the four group means with 95% confidence intervals for both indexes where group 4 relates to the H1N1 neuraminidase. If we assume that the group 3 index values relating to the long-stalk H5N1 neuraminidases are basically from viruses benign to human beings (note: our data set of 29 sequences contains only one from a human isolate), then Figure 4 shows an interesting feature in relation to effects on human beings: the base composition and distribution indexes for region C of the neuraminidase sequences have a significant difference between the benign and pathogenic forms of the viruses for the short-stalk and long-stalk varieties individually. The mean n_R values of the short-stalk varieties with 95% confidence intervals for the mean are 1.19 ± 0.01 and 1.23 ± 0.003, and those of the long-stalk varieties are 1.24 ± 0.01 and 1.29 ± 0.01, implying comparable differences between the two periods for the short- and long-stalk varieties individually. In the case of the g_R values, the two sets are 33.2 ± 1.8 and 42.6 ± 0.7 for the short-stalk varieties and 48.3 ± 1.4 and 52.7 ± 2.8 for the long-stalk varieties.

4. CONCLUSION

We conclude from our analysis using the mathematical descriptors, n_R and g_R, that different regions of the neuraminidase gene sequence show different mutational behavior. On the basis of our broad classification of the different strains of the H5N1 virus into short- and long-stalk neuraminidase genes and periods when humans were affected versus when they were not, we have found that the neuraminidase sequence segments corresponding to the transmembrane and stalk regions of the protein appear to show significant differences in base composition and distribution patterns, but a small 50-base segment at the 3′ end is very strongly conserved. Such a segment, strongly conserved over a large variety of these influenza strains, presents the possibility of using this part of the neuraminidase as an anchor point for a function-debilitating molecule. The conserved nature of this section could enable such a molecule to continue to attach to the neuraminidase irrespective of the various mutations taking place in the overall sequence, thus mitigating one of the main problems of devising strong viral antidotes and vaccines.

We found especially that the large, 1090-base-long C segments of the neuraminidase sequences that are part of the viral genomes which are highly pathogenic to humans have higher n_R and g_R values compared to the relatively more benign forms; this is also the case for the H1N1 human isolates we have studied. Such differences imply that mutations that enhance the adenine and thymine content in these genes have a higher probability of modifying the genome into a form more pathogenic to human beings. A detailed analysis of the amino acid composition of the C segments of selected sequences from our H5N1 set showed that the mutations led to small reductions in the number of hydrophobic amino acids in the group 2 and 3 samples, as compared to a selected group 1 sample, indicating that there could be some conformational adjustments in the neuraminidase structure in the more virulent years.

Further, we have observed distinct differences in the base composition and distribution indexes in the C segment of

the short-stalk and long-stalk varieties of the neuraminidase gene separately. Whether these differences can be related to issues of transmissibility to humans remains open to question, but we note that the only strain of the H5N1 virus found by Shinya et al.[27] to bind to the upper and lower respiratory tracts was the lone sample of the H5N1 human isolate with the long-stalk neuraminidase (A/Hong Kong/213/03 [H5N1]; GENBANK: AB212056). All previously identified transmissible forms of the avian flu virus with N1 neuraminidase have been the long-stalk H1N1 subtype, where we have found that on average the n_R and g_R values are the largest among all the groups studied. However, we note for the record that our computation on the 1918 H1N1 flu virus (A/Brevig_Mission/1/18 [H1N1]; GENBANK: AF250356) sequence yields $n_R = 1.27$ and $g_R = 47.7$, whereas the long-stalk H5N1 human isolate strain referred to above has $n_R = 1.25$ and $g_R = 46.8$.

ACKNOWLEDGMENT

This manuscript is dedicated to Professor Nenad Trinajstic on his 70th birthday. Assistance from the Consortium for Bioinformatics and Computational Biology, University Minnesota, and discussions with Dr. R. Natarajan and D. Mills are gratefully acknowledged. A.N. would also like to thank Suman Bhandari of Jadavpur University for helpful discussions. This is paper number 447 from the Center for Water and the Environment of the Natural Resources Research Institute, University of Minnesota Duluth.

REFERENCES AND NOTES

(1) Chen, H.; Smith, G. J. D.; Li, K. S.; Wang, J.; Fan, X. H.; Rayner, J. M.; Vijaykrishna, D.; Zhang, J. X.; Zhang, L. J.; Guo, C. T.; Cheung, C. L.; Xu, K. M.; Duan, L.; Huang, K.; Qin, K.; Leung, Y. H. C.; Wu, W. L.; Lu, H. R.; Chen, Y.; Xia, N. S.; Naipospos, T. S. P.; Yuen, K. Y.; Hassan, S. S.; Bahri, S.; Nguyen, T. D.; Webster, R. G.; Peiris, J. S. M.; Guan, Y. Establishment of Multiple Sublineages of H5N1 Influenza Virus in Asia: Implications for Pandemic Control. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 2845−2850.

(2) Mills, C. E.; Robins, J. M.; Bergstrom, C. T.; Lipsitch, M. Pandemic Influenza: Risk of Multiple Introductions and the Need to Prepare for Them. *PLoS Med.* **2006**, *3*, 1−5.

(3) Li, K. S.; Guan, Y.; Wang, J.; Smith, G. J. D.; Xu, K. M.; Duan, L.; Rahardjo, A. P.; Puthavathana, P.; Buranathai, C.; Nguyen, T. D.; Estoepangestie, A. T. S.; Chalsingh, A.; Aueswarakul, P.; Long, H. T.; Hanh, N. T. H.; Webby, R. J.; Poon, L. L. M.; Chen, H.; Shortridge, K. F.; Yuen, K. Y.; Webster, R. G.; Peiris, J. S. M. Genesis of a Highly Pathogenic and Potentially Pandemic H5N1 Influenza Virus in Eastern Asia. *Nature (London, U.K.)*

(4) World Health Organization (WHO). http://www.who.int/csr/don/2006_05_23/en/ (accessed Feb 2007).

(5) Suarez, D. L.; Perdue, M. L.; Cox, N.; Rowe, T.; Bender, C.; Huang, J.; Swayne, D. E. Comparisons of Highly Virulent H5N1 Influenza A Virus Isolated from Humans and Chickens from Hong Kong. *J. Virol.* **1998**, *72*, 6678−6688.

(6) Peng, C.-K.; Buldyrev, S. V.; Goldberger, A. L.; Havlin, S.; Sciortino, F.; Simons, M.; Stanley, H. E. Long Range Correlations in Nucleotide Sequences. *Nature (London, U.K.)* **1992**, *356*, 168−170.

(7) Jeffrey, H. J. Chaos Game Representation of Gene Structures. *Nucleic Acids Res.* **1990**, *18*, 2163−2170.

(8) Voss, R. Evolution of Long-Range Fractal Correlations and 1/f Noise in DNA Base Sequences. *Phys. Rev. Lett.* **1992**, *68*, 3805−3808.

(9) Moscona, A. Neuraminidase Inhibitors for Influenza. *New Engl. J. Med.* **2005**, *353*, 1363−1373.

(10) Nandy, A.; Harle, M.; Basak, S. C. Mathematical Descriptors of DNA Sequences: Development and Applications. *ARKIVOC (Gainesville, FL, U.S.)* **2006**, *9*, 211−238.

(11) Peiris, J. S. M.; Yu, W. C.; Leung, C. W.; Cheung, C. Y.; Ng, W. F.; Nicholis, J. M.; Ng, T. K.; Chan, K. H.; Lai, S. T.; Lim, W. L.; Yuen, K. Y.; Guan, Y. Re-emergence of Fatal Human Influenza A Subtype H5N1 Disease. *Lancet* **2004**, *363*, 617−619.

(12) Gates, M. A. A Simple Way to Look at DNA. *J. Theor. Biol.* **1986**, *119*, 319−328.

(13) Nandy, A. A New Graphical Representation and Analysis of DNA Sequence Structure: I. Methodology and Application to Globin Genes, *Curr. Sci.* **1994**, *66*, 309−314.

(14) Leong, P. M.; Morgenthaler, S. Random Walk and Gap Plots of DNA Sequences. *Comput. Appl. Biosci.* **1995**, *11*, 503−507.

(15) Yau, S. S. T.; Wang, J.; Niknejad, A.; Lu, C.; Jin, N.; Ho, Y.-K. DNA Sequence Representation Without Degeneracy. *Nucleic Acids Res.* **2003**, *31*, 3078−3080.

(16) Randic, M.; Vracko, M.; Nandy, A.; Basak, S. C. On 3-D Graphical Representation of DNA Primary Sequences and Their Numerical Characterization. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1235−1244.

(17) Chi, R.; Ding, K. Novel 4D Numerical Representation of DNA Sequences. *Chem. Phys. Lett.* **2005**, *407*, 63−67.

(18) Randic, M.; Lers, N.; Plavsic, D.; Basak, S. C.; Balaban, A. T. Four-Color Map Representation of DNA or RNA Sequences and Their Numerical Characterization. *Chem. Phys. Lett.* **2005**, *407*, 205−208.

(19) Randic, M.; Guo, X.; Basak, S. C. On the Characterization of DNA Primary Sequences by Triplets of Nucleic Acid Bases. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 619−626.

(20) Wang, J.; Zhang, Y. Characterization and Similarity Analysis of DNA Sequences Grounded on a 2D Graphical Representation. *Chem. Phys. Lett.* **2006**, *423*, 50−53.

(21) Raychaudhury, C.; Nandy, A. Indexing Scheme and Similarity Measures for Macromolecular Sequences. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 243−247.

(22) Nandy, A.; Basak, S. C. Simple Numerical Descriptor for Quantifying Effect of Toxic Substances on DNA Sequences. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 915−919.

(23) Nandy, A.; Nandy, P.; Basak, S. C. Quantitative Descriptor for SNP Related Gene Sequences. *Internet Electron. J. Mol. Des.* **2002**, *1*, 367−373. http://www.biochempress.com/ (accessed Feb 2007).

(24) Nandy, A.; Nandy, P. On the Uniqueness of Quantitative DNA Difference Descriptors in 2D Graphical Representation Models. *Chem. Phys. Lett.* **2003**, *368*, 102−107.

(25) Universal Protein Resource (UNIPROT). http://www.uniprot.org (accessed Feb 2007).

(26) Kou, Z.; Lei, F. M.; Yu, J.; Fan, Z. J.; Yin, Z. H.; Jia, C. X.; Xiong, K. J.; Sun, Y. H.; Zhang, X. W.; Wu, X. M.; Gao, X. B.; Li, T. X. New Genotype of Avian Influenza H5N1 Viruses Isolated from Tree Sparrows in China. *J. Virol.* **2005**, *79*, 15460−15466.

(27) Shinya, K.; Ebina, M.; Yamada, S.; Ono, M.; Kasai, N.; Kawaoka, Y. Avian Flu: Influenza Virus Receptors in the Human Airway. *Nature (London, U.K.)* **2006**, *440*, 435−436.

CI600558W