Computational Study of Dispersion and Extent of Mutated and Duplicated Sequences of the H5N1 Influenza Neuraminidase over the Period 1997−2008

Ambarnil Ghosh,[†] Ashesh Nandy,[‡] Papiya Nandy,[†] Brian D. Gute,[§] and Subhash C. Basak*[,§]

Physics Department, Jadavpur University, and School of Environmental Studies, Jadavpur University, 188 Raja S.C. Mallick Road, Jadavpur, Kolkata, 700032 West Bengal, India, and Natural Resources Research Institute, University of Minnesota Duluth, 5013 Miller Trunk Highway, Duluth, Minnesota 55811

Received May 6, 2009

Study of mutational changes in neuraminidase (NA) gene sequences is important to track the effectiveness of the inhibitors to the H5N1 avian flu virus that targets this component of the viral apparatus. Our analysis based on numerical characterization studies of 682 complete neuraminidase gene and protein sequences available in the database, updated to March 2009, and which extends our previous work based on a sample of 173 sequences has revealed several interesting features. We have noticed that identical sequences have appeared over significant distances in space and time, raising the need for a deeper understanding of the longevity of such viral strains in the environment. Structural sections like transmembrane, stalk, body, and C-terminal tail regions have shown independent recombinations between strains from various species including human and avian hosts highlighting influenza's flexibility in host selection and recombination. Our analysis confirmed a biased nature in mutational accumulation in structural segments: a highly conserved 50-base C-terminal tail section identified in our earlier paper seems to accumulate mutational changes at a rate of about a fifth to an eighth of transmembrane and stalk regions, although the length is about half of these. Parallel study of the equivalent section to the C-terminal region in protein sequences reveals only 13 separate varieties, and all the other 669 sequences are duplicates to three of these varieties showing the highly conserved nature of this segment. Our analysis of active site related bases and amino acids showed highly conserved characteristic of those constructs, whereas the rest of the segments demonstrated rather large mutational changes. These kinds of high level of mutation in major part of the H5N1 NA sequences and recombinations within structural segments coupled with strong conservation of a few select segments show that the potential of rapid mutations to more virulent forms of this variety of avian flu continue to remain of concern, especially with the possibility of long duration dormancy of some of these viral strains, whereas islands of highly conserved segments could signify potential regions for inhibitor designs.

1. INTRODUCTION

The H5N1 avian flu virus has been a potent killer of domestic and wild fowl since its strong emergence and identification in 1997. Since then the virus has spread from its place of origin in South Central China to the rest of Asia and been carried by migratory birds to Europe and Africa. Large scale culling has mitigated the spread of the virus to some extent, but the existence of the virus gene pool in China and continuous mutations among the virus strains have led to the contagion spreading worldwide by migratory birds and other carriers,[1−3] with sudden conflagrations erupting at various locations, the latest being the outbreak in eastern India and Bangladesh in 2007 and again in 2008−2009[4] as well as in Egypt and elsewhere.

The primitive poultry farming techniques in countries such as Indonesia have led to infections among humans also. The H5N1 flu has already resulted in over 256 deaths out of 412 confirmed cases of infections among humans (see Table - Human Cases of Avian Influenza, and Situation Updates, ref 5) and is believed to have the potential to mutate to a highly contagious form that may cause severe pandemic among human populations.[3] Although the number of fatalities in humans from this virus appears small, the rapid mutations that viruses can undergo, and the possibility of whole gene or gene fragments shuffling between avian and mammalian hosts, carries the potential to cause a pandemic challenge and also render current H5N1 inhibitors ineffective through antigenic shift and drift changes.[6,7] It is believed that a human pandemic from the H5N1 virus could cause the deaths of millions of people;[8] the 1918 H1N1 avian flu pandemic had resulted in fatalities of 20 to 50 million people worldwide.[9] Since the inhibitors of this influenza virus, principally oseltamivir and zanamivir, act on the neuraminidase component of the H5N1 protein, continuous monitoring of the mutational changes in this gene assumes significance.[10,11]

Currently prevention and treatment of influenza rely on inactivated vaccines and antiviral drugs. Impact of mutational changes in amino acid residues on the stability, activity, and sensitivity of the target protein is a widely studied topic in antiviral drug design and for adequate remedy.[10,12,13] The general causes involved in generation of antiviral-resistant variety of the strains have been the main target of the major researches (see e.g., ref 14.). Several investigations have

* Corresponding author phone: 218-720-4230; fax: 218-720-4328; e-mail: sbasak@nrri.umn.edu.
[†] Physics Department, Jadavpur University.
[‡] School of Environmental Studies, Jadavpur University.
[§] University of Minnesota Duluth.

10.1021/ci9001662 CCC: $40.75
Published on Web 09/24/2009

focused upon phylogenetic relationships in evolutions of virulence,[15] and some other researchers are trying to correlate the evolution of varieties that affect humans and those that infect avian populations. Evolutionary and transmission dynamics study of Vijaykrishna et al.[16] and Lam et al.[17] relate viral evolution with reassortment of H5N1 influenza strains. Lam et al. showed phylogenetic evidence for inter-lineage reassortment among H5N1 HPAI viruses isolated from different sources in Indonesia and identified potential genetic parents of reassorted genome segments. They also discussed the origin of reassortants from genetic, temporal, and geographical viewpoints. Recently, in a study done by Owoade et al.,[18] it had been shown that the reassortment events drive the local emergence of HPAI viruses in Nigeria. Related studies on this same area lead to a field of recombinomics study in influenza virus genome. Evolutionary genetics studies on model RNA viruses already proved an important role of homologous recombination in generation of virulent varieties,[19] but the occurrence of homologous recombination within avian influenza segments still remains an issue of controversy. Recent works of He et al.[20,21] showed results demonstrating the role of intragenic recombination in avian influenza virulence and host tropism changes.

These studies, among others, of the H5N1 virus have been focused on the evolutionary and reassortant characteristics of the different strains of the H5N1 gene and protein sequences by constructing phylogenetic trees and focusing on the role of specific amino acids in the sequences.[15−18] In our previous paper on the H5N1 neuraminidase gene sequences,[22] we had used graphical representation and numerical characterization techniques for the analysis of global characteristics of the sequences. Our current work on the family of the avian flu neuraminidase genes is also based on graphical representation and numerical characterization of base composition and distribution characteristics, coupled with a new extension of these techniques to the amino acid sequences, to determine any systematic and exceptional behavior that may have arisen from mutational changes. This has led to several important results on mutational variability of the structural segments as well as the complete sequences of the neuraminidase genes and proteins which we report in this paper.

The neuraminidase protein is commonly viewed as consisting of three main parts:[23,24] the transmembrane composed of 35 amino acids, the stalk, which is variable consisting of 35, 36, or 55 amino acids, and the body composed of a sequence of 379 amino acids which we had analyzed individually in our previous paper.[22] In this paper we extend the previous work further by making a systematic review of all 682 complete H5N1 neuraminidase gene sequences available in the GenBank database,[25] updated to March 10, 2009. We find from this data, inter alia, that (a) identical gene and protein sequences have appeared over significant distances in space and time; (b) a short 50-base segment at the 3′-end of the neuraminidase gene sequence identified in the previous paper[22] as being highly conserved is seen to be highly conserved in this larger universe of neuraminidase sequences; (c) structural sections like transmembrane, stalk, body, and C-terminal tail regions have shown independent recombinations between strains from various species including human and avian hosts; (d) comparison of mutational

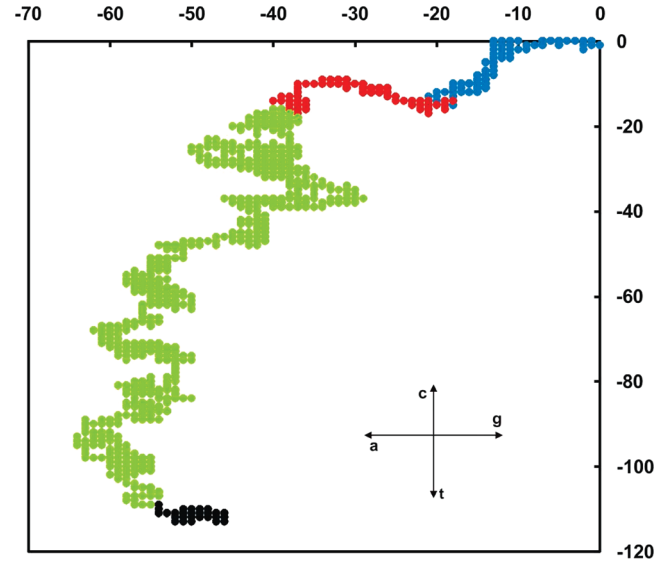

Figure 1. A 2D graphical representation of A/Hanoi/30408/2005(H5N1) neuraminidase gene sequence. The colored regions indicate: blue - transmembrane; reddish brown - stalk; green and black: body. The C-terminal tail region used in the analysis is shown in black.

accumulation in structural segments shows different degrees of changes with the 50-base C-terminal tail section accumulating mutational changes at a rate of about a fifth to an eighth of transmembrane and stalk regions, although the length is about half of these; and (e) active site related nucleoside bases and amino acid residues showed highly conserved characteristics, whereas the rest of the segments demonstrated rather large mutational changes. These kinds of high levels of mutation in a major part of the H5N1 neuraminidase sequences and recombinations within structural segments coupled with strong conservation of a few select segments show that the potential of rapid mutations to more virulent forms of this variety of avian flu continue to remain of concern. However, islands of highly conserved segments could signify potential regions for inhibitor designs.

2. METHODS

For our analysis we have collected and numerically characterized 682 complete RNA sequences and protein products of the neuraminidase genes of the H5N1 avian flu strains available in the NCBI-Flu[25] database with recent updates to March 10, 2009, encompassing strains up to 2008. These comprised 639 sequences with the 35 aa stalk lengths, i.e. total 1350 bases in the nucleotide sequence, 8 sequences with 36 aa stalk length, i.e. 1353 base nucleotide sequences, and 35 sequences with 55 aa stalk lengths, i.e. 1410 base nucleotide sequences.

For the analyses we have chosen graphical representation and numerical characterization methods since these are easy to use for global and local comparisons of the sequences and have been widely applied.[26−30] For the gene sequences of the different strains of the H5N1 avian flu, we first generate 2D graphical plots by plotting one point for each base of a sequence by the algorithm: move one step in the negative x-direction for an adenine, one step in the positive y-direction for a cytosine, one step in the positive x-direction for a guanine, and one step in the negative y-direction for a thymine. Figure 1 shows the representative graph of the

H5N1 strain A/Hanoi/30408/2005 NA sequence. Sections of the RNA sequence can be cross-identified with four structural segments of the NA (Figure 1) mentioned earlier.

For numerical characterization of the RNA sequences, we calculate the first order moments and graph radius based on the graphical representation as follows

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

where (x_i, y_i) represent the coordinates of each point on the plot, and N is the total number of the bases in the segment. The graph radius, g_R, represents the Base Distribution index. The g_R is very sensitive to changes in the sequence composition and distribution, the values depending on the type of mutations and where in the sequence they are. g_R is specially useful in comparing equal length sequences, which is the case for the NA genes considered in this paper. Equal values of g_R imply that the compared sequences are identical,[31] which gives us a strong handle to identify duplicates in the complete sequences and the structural segments and thus provide an indication of mutational changes and whether any structural segment has preferential advantages over another.

We use an equivalent method to compare the neuraminidase protein sequences.[32] Here we use an abstract 20D Cartesian coordinate system to generate a protein sequence walk by plotting one point for each amino acid in the sequence along a designated axis and measure the resultant first order moments and protein graph radius through

$$\mu_1 = \sum \frac{x_1}{N}, \mu_2 = \sum \frac{x_2}{N}, \mu_{20} = \sum \frac{x_{20}}{N}$$

$$\text{and } p_R = \sqrt{\mu_1^2 + \mu_2^2 + \ldots + \mu_{20}^2}$$

In our method of representation and characterization of protein sequences, we associate each amino acid with one axis of a 20D Cartesian coordinate system; the choice of association is equivalent for all residues and can be arbitrarily assigned but once assigned will be fixed for the duration of the computation. Our choice of axes and associated amino acid unit vectors for the current exercise are given in Table 1. Further details of the method can be found in ref 32. As in the case of the g_R, the p_R values also are found to be sensitive to changes in the amino acid sequences, and equal values of the p_R imply exact duplication of the amino acid composition and distribution along the sequences. This numerical characterization method, as developed to date, refers strictly to the identities of the amino acids and is transparent to their chemical properties, i.e., no distinctions are made between residues that are mutationally conservative or not, between polar and nonpolar, between basic and acidic, etc., and all residues are treated at par.

We used a Pentium-4 processor Dual-Core machine as computational platform. Necessary applications were built by C/C++-Programming language and Turbo-C++ compiler. Other numerical analyses are done by Microsoft Office 2007 software.

3. RESULTS AND DISCUSSION

Based on the graphical representations of the neuraminidase gene and protein sequences, the numerical characterization data of the sequences and their structural segments are determined. The tables below list a summary of the main results.

Table 1. Assignment of Axes to Individual Amino Acids

axis no.	amino acid	3-letter code	single letter code
1	alanine	Ala	A
2	cysteine	Cys	C
3	aspartic acid	Asp	D
4	glutamic acid	Glu	E
5	phenylalanine	Phe	F
6	glycine	Gly	G
7	histidine	His	H
8	isoleucine	Ile	I
9	lysine	Lys	K
10	leucine	Leu	L
11	methionine	Met	M
12	asparagine	Asn	N
13	proline	Pro	P
14	glutamine	Gln	Q
15	arginine	Arg	R
16	serine	Ser	S
17	threonine	Thr	T
18	valine	Val	V
19	tryptophan	Trp	W
20	tyrosine	Tyr	Y

A brief explanation of the organization of data in Tables 2 and 3 is as follows: Considering the transmembrane segment, arranging the protein numerical descriptors pR in order shows that there are 58 strains whose p_R values are not duplicated, whereas there are e.g., 29 sequences with $p_R = 7.032581$, 3 sequences with $p_R = 7.015072$, etc. comprising 34 such groups of duplicates in all. Thus there are then 58 single occurrence p_R's and 34 multiple occurrences, each group of duplicates being different from any other group or singles, making a total of $58 + 34 = 92$ unique sequences in the total universe of 682 H5N1 neuraminidase strains; accordingly total number of duplicates amount to $682 - 92 = 590$ sequences as shown in Table 2 (protein sequence - TM). We can arrange the nucleotide sequences also in a similar fashion. For the nucleotide transmembrane segment sequences we find that there are likewise a total of 161 unique sequences and 521 duplicates in 63 groups (Table 2, nucleotide sequence - TM). However, several of these nucleotide sequences code for the same protein sequences, i.e. they are synonymous. E.g., in the case of the 3 sequences with $p_R = 7.015072$ quoted above, the three nucleotide sequences all have the same $g_R = 12.38302$; we count these as comprising two sequences synonymous to one of them. Similarly, in the case of the 29 sequences with $p_R = 7.032581$, the nucleotide sequences comprise 5 groups with 2, 4, 5, and 6 (twice) duplicates and 6 single occurrences, all coding for the same protein sequence. We count these as comprising 28 synonymous sequences for the purpose of Table 3 where we have indicated that there are a total of 77 synonymous sequences counted in this manner in the 682 samples of the transmembrane segment.

Table 2 shows that the total number of duplicates in the nucleotide sequences of the neuraminidase gene is the largest in the C-terminal tail sequence followed by the transmembrane, stalk, body, and the complete sequence, in that order. Several mutated sequences are replicated in different host organisms, and several such groups of identical sequences are recognized in each structural segment as also in the

Table 2. Summary of the Number of Mutated and Duplicate Sequences[a]

	nucelotide sequence					protein sequence				
	TM	ST	BD	TL	FULL	TM	ST	BD	TL	FULL
total number of H5N1 strains	682	682	682	682	682	682	682	682	682	682
no. of duplicate sequences in the above	521	451	158	652	135	590	522	329	669	258
no. of groups containing duplicates	63	87	78	15	71	34	65	86	3	94
no. of unique sequences in total	161	231	524	30	547	92	160	353	13	424
percentage of unique sequences to total	23.61	33.87	76.83	4.40	80.21	13.49	23.46	51.76	1.91	62.17

[a] TM - transmembrane; ST - stalk; BD - body; TL - tail; FULL - complete sequence.

Table 3. Statistics of Synonymous Sequences from Numerical Analysis

	TM	ST	BD	TL	FULL
no. of synonymous sequences	77	130	171	25	123
percentage of synonymous sequences to total	11.29	19.062	25.073	3.6657	18.035
percentage of synonymous sequences to uniques	47.826	56.277	30.7	83.333	22.486

complete gene. The total number represented above in the first row is the sum of all such duplicates; the remainders are classified as being unique sequences, arising from accumulations of different mutations. Thus, for example, 547 out of the 682 complete sequences of the neuraminidase gene are found to be different from one another; of these, 71 sequences are found to have one or more duplicates making a total of 135 duplicate strains. In terms of the protein sequence, the corresponding numbers are 424, 94, and 258, respectively. Among the nucleotide sequences, 123 are found to have had synonymous mutations (Table 3), i.e. they code for one or the other of the protein sequences already existing in the H5N1 NA world. As can be expected because of stalk length differences, no identical duplicates were found between the groups of 1410, 1353, and 1350 base neuraminidases, although there are two groups of identical strains in the 1410 base neuraminidase protein sequences, none for the 1353 base strains, and 51 groups in the 1350 base neuraminidases.

Our analysis of sequence duplicates also shows some interesting features. Identical neuraminidase gene sequences appear over distances in time and space, and there are several instances of the exact same sequence showing up in avians and humans also. Duplicate sequences appearing in localized clusters can be understood to have sprung from one strain of the virus circulating at the time, but when duplicate strains appear thousands of kilometers apart then it is intriguing to consider whether that particular strain has remained stable over the long journey, or whether there has been a case of reverse mutations. Especially intriguing is the appearance of the same identical strain in two cases which are geographically and temporally far apart. We consider these through comparison of the full sequences as well as the sequences comprising the transmembrane, stalk, body, and the C-terminal tail segments of the neuraminidase.

We compare the different strains in our database in three categories to determine the geographical and temporal distribution of sequence duplicates. We compare the neuraminidases in terms of the full sequences first, then we compare the neuraminidases by their structural segments, i.e. the transmembrane, stalk, body, and the C-terminal tail segments, separately, and last we compare a construct of the bases and amino acids in the active site region to understand its degree of conservation. For a geographical perspective, we show in Figure 2a,b identical sequences occurring in different locations connected by lines; identical sequences found in Hong Kong over different time periods are indicated by a loop (not done for other locations to avoid cluttering the map), and identical sequence segments occurring over many locations are indicated by symbols.

(a) Complete Sequence. Some of the overall statistics have already been mentioned above. There are no duplicates in the neuraminidase gene sequences of length 1353 bases and 1410 bases. The protein sequences are also unique, except for two groups of two duplicates each from the 1410 base length neuraminidase proteins (implying 2 synonymous mutations in the nucleotide sequences). The 1999 strain A/Environment/HongKong/437-8/99(H5N1) and the 2001 strain A/Chicken/HongKong /317.5/2001(H5N1) have identical amino acid sequences (marked as line 1 in Figure 2a). Two other strains from Hongkong from 2003, A/egret/HongKong/757.2/03(H5N1) an avian isolate and A/HongKong/213/03(H5N1) from an infected human, have identical protein sequences; it is interesting to note that this human isolate was considered as a highly potent strain within the different strains investigated by Shinya et al.[33]

Among the 639 strains of the 1350 length neuraminidase gene sequences in our sample, there are 135 duplicates in 71 groups. Comparing with the protein sequences, we find that among the balance 504 unique gene sequences of this length, there are 121 sequences that lead to protein sequences that are among the duplicates already counted, implying these gene sequences have had synonymous mutations. A detailed study of the exact duplicates show that repetitions occur in widely separated locations and a wide variety of hosts. We find that strains from Qinghai, China, 2005, e.g., A/Bar-headedGoose/Qinghai/59/05(H5N1) (line 2 in Figure 2a), are duplicated in strains collected from swans in Astrakhan, southern Russia, e.g., A/Cygnusolor/Astrakhan/Ast05-2-3/2005(H5N1), about 5000 km distance away from Qinghai Lake in Central China. In the same year, 2005, a turkey and a chicken from Egypt are found to have the exact same sequence as a chicken in Israel (line 3 in Figure 2a), sourced probably through poultry trade between the two countries. In another instance, A/chicken/Egypt/1889N3-SM26/2007-(H5N1) (line 4 in Figure 2a) and A/chicken/Ghana/3159-NAMRU3/2007(H5N1) from two countries separated by the wide swathe of the Sahara are seen to have identical sequences. Two samples collected from ducks in Hunan in

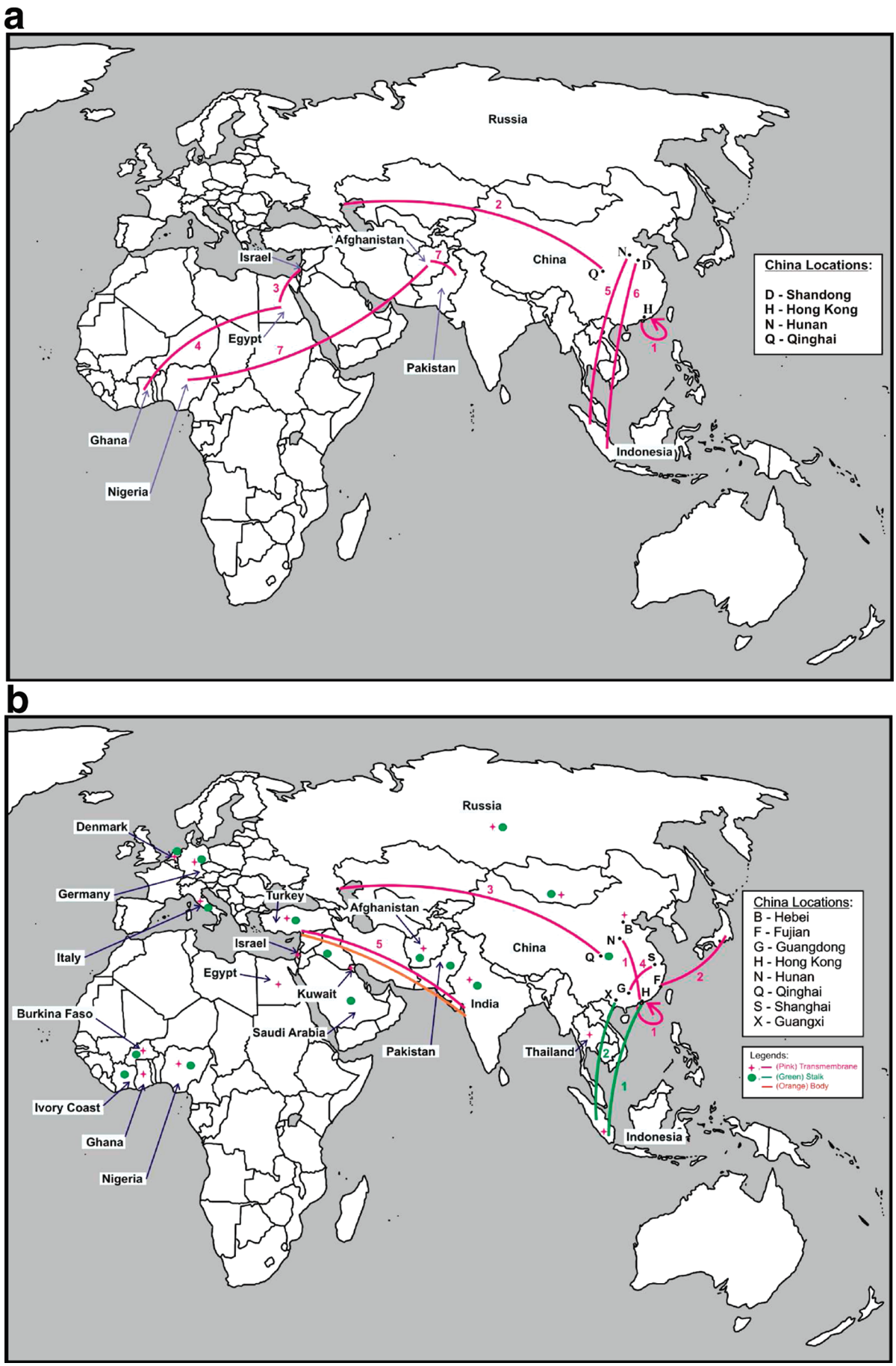

Figure 2. a. Location of identical complete NA sequences found widely separated in space and time. Numbers on connecting lines are referred to in the text. b. Location of identical NA sequence segments found widely separated in space and time (transmembrane - pink; stalk - green; body - orange). Numbers on connecting lines and symbols are referred to in the text.

2006, e.g., A/duck/Hunan/988/2006(H5N1), show the exact same sequence in human isolates from Indonesia (A/Indonesia/CDC938/2006(H5N1) and A/Indonesia/CDC938/2006(H5N1)) (line 5 in Figure 2a), several thousand kilometers away from where the duck sample was gathered.

In our collection of H5N1 flu viral strains, there are 57 strains from infected Indonesians, some from different sources from the same individuals. While strains from e.g., nose, throat, etc., of an individual collected on the same day show identical sequences, there are two reported instances where collections were taken a day or two apart where we find mutated, albeit synonymous, sequences. A/Indonesia/CDC624/2006(H5N1) ETT-wash taken on May 19, 2006 and A/Indonesia/CDC624E/2006(H5N1) throat swab taken the same day, both from a male human aged 39 years, which we presume are one and the same person, show nucleotide sequence of the neuraminidase that are different but have the same translated product. Similarly, A/Indonesia/CDC625/2006(H5N1) throat swab taken on May 21, 2006 and A/Indonesia/CDC625L/2006(H5N1) lung aspirate taken on May 22, 2006, both from a male aged 32 years, which again we assume are one and the same person, show differences in the viral nucleotide sequence but have the same protein sequences. While other Indonesian instances of presumed same persons with samples taken from different body regions show duplicate neuraminidases, the two examples quoted above imply that mutations when they take place can be quite rapid. It is all the more surprising therefore that the exact same sequences appear in locations geographically far apart implying significant time delays between infections as mentioned hereinbefore.

Comparing complete protein sequences of the short stalk neuraminidases, we find that the strain from a mammalian host, A/swine/Shandong/2/03(H5N1), recurs as an exact duplicate 2 years later in an avian isolate, A/chicken/Indonesia/R60/05(H5N1) (line 6 in Figure 2a), in a distant country. Circulation of the same strain in local clusters can be expected: several isolates from avian species and humans in the Egypt-Gaza-Israel area collected in 2006−2007 have exactly the same protein sequences, though different from the above. However, specimens from countries as far apart as the Sudan, Iraq, Turkey, Mongolia, and the Ivory Coast in 2006 alone have also shown exact duplicates of the same protein sequence, albeit different from the previous two instances. Another such clustering of identical protein sequences is observed between 2005 and 2007 in samples collected from Qinghai and Astrakhan in 2005, mentioned for the nucleotide case earlier, but also from Nigeria and Afghanistan in 2006 (A/guineafowl/Nigeria/957-12/2006-(H5N1) and A/chicken/Afghanistan/ 1573-65/2006(H5N1)) and Pakistan in 2007 (A/crow/Peshawar/ NARC7914/2007-(H5N1)) (line 7 in Figure 2a). Such bias toward the exact same sequence structure perhaps indicates that mutations in the gene sequences, at least for these strains, were constrained to be within some strict limits.

(b) Structural Sections. Since the neuraminidase protein has identifiable distinct structural segments, it is instructive to consider whether these structural segments have identical mutational traits or whether their variations are demonstrably independent. We consider the neuraminidase gene in terms of the three main segments we have identified earlier, i.e. transmembrane, stalk, and body and also, separately, the C-terminal tail region.

Table 4. Presentation of Recombination Data from Major Structural Gene Segments[a]

transmembrane	stalk	body	total seqs.
S	S	S	135
S	S	V	279
S	V	S	27
V	S	S	12
S	V	V	443
V	S	V	393
V	V	S	41
V	V	V	157

[a] S represents "similar" (i.e., identical within each segment), and V represents "variable" in relation to sequence identity.

While the complete nucleotide sequences show that only 19.79% of them are duplicates to one or another sequence (Table 2), the structural segments show wide disparities in sequence conservation. The transmembrane region with only 23.61% unique sequences, i.e. 76.39% duplicates, appears to be highly stable, followed closely by the variable length stalk region with 33.87% of the sequences being unique (66.13% duplicates). The corresponding protein sequences have lesser quantum of unique sequences as could be expected: 13.49% protein sequences for the transmembrane region are unique, while 23.46% are unique for the stalk region. The body region has a large percentage of unique sequences, 76.83% for the nucleotide sequences and 51.76% for the protein sequences. These numbers for the protein sequences are expected to be less than for the nucleotide sequences because of the synonymous mutations. Less than a third of the unique gene sequences in the body region are synonymous sequences (Table 3), whereas this ratio is about half for the transmembrane and stalk segments but only about a fourth for the full sequences. When we consider the 50-base tail region at the 3′-end of the neuraminidase gene, we find 652 sequences out of the total sample size of 682 strains are duplicates in 15 varieties; the corresponding numbers for the protein sequences are 669/682 and 3. Of the total 30 unique types of nucleotide sequences that are found in this region, 25 are synonymous, coding for one or the other of the 13 protein varieties. This C-terminal segment had already been identified as being exceptionally stable in our earlier analysis;[22] it continues to show remarkable conservation in the current larger collection of sequence data. Indeed, the C-terminal tail section seems to accumulate mutational changes at a rate at about a fifth to an eighth of the transmembrane or stalk region, although the length is about half of these, implying that the C-terminal tail region must be under comparatively higher evolutionary pressures to filter out mutational changes.

The study of duplicate sequences in the structural segments too reveals interesting facts. Except for the 135 duplicates in the complete neuraminidase gene, in other cases the neuraminidase that shows duplicate sequences in one segment may have the other segments with sequences that are not duplicates of one another. Table 4 lists the number of sequences where two segments are duplicated while one is

not, one segment is duplicated while two others are not, and where all three are nonduplicates. These numbers are indicative perhaps of the possibility of preferential exchange of structural segments between different H5N1 neuraminidase strains. While gene exchanges between different viral genomes have been recorded,[18] exchanges, or homologous recombinations, between structural segments of the same gene from different strains of the H5N1 virus do not seem to have been noted so far. Some segmentwise details are mentioned below; components of the 135 duplicates of the complete sequences covered earlier are omitted from these discussions.

When comparing the segmentwise sequences of the neuraminidase genes and proteins, it is understandable that the variable stalk segments of the 1350-base, 1353-base, and 1410-base strains will yield different sets of results with no overlaps. What is interesting to note is that the 105-base (35 aa) transmembrane region as well as the 1140-base (380 aa) body region show no identity with the g_R and the p_R values among the three sets of strains comprising the variable stalks. This implies that although the base and amino acid numbers are the same for all the neuraminidase sequences, the three varieties of the neuraminidase must be having distinctive features in the transmembrane and body regions that are preserved even with all the mutations that have accumulated so far as per the available data.

Transmembrane Segment. The eight 1353-base sequences (36 amino acids in the stalk region) form three separate transmembrane sequences: one sequence appears in the database twice (Hongkong/482 and 486 of 1997), one sequence appears 5 times, and one is a solitary individual, i.e. there are 5 duplicates altogether. The protein sequences follow the same pattern, i.e. there are no synonymous mutations in this segment. Among the 35 1410-base gene sequences, there are 9 duplicates in 4 groups, while in the corresponding protein sequences there are 15 duplicates in 2 groups, from 6 unique nucleotide sequences, the balance being synonymous. Seven of the duplicates in the transmembrane region are not accompanied by duplicates in the body and stalk regions (Table 4). Although the number of strains of the long stalk variety neuraminidase are small and are expected to have died out in the wild,[3] the transmembrane sequences are found duplicated across time and space. Thus, the same duplicate strain is found in sequences from Hong Kong in 1999 and 2001 (A/Environment/HongKong/437-6/99(H5N1), A/Goose/HongKong/76.1/01(H5N1)) and in Henan in 2004 (A/treesparrow/Henan/4/2004(H5N1)), identified in Figure 2b by line number 1 in pink. On the other hand, a transmembrane sequence found in China in 2001 (A/duck/Fujian/17/2001(H5N1)) is found in Japan in 2003 (A/duck/Yokohama/aq10/2003(H5N1)) (line 2 in pink in Figure 2b) and more duplicates found in the same year in Hong Kong in bird and human (A/egret/HongKong/757.2/03(H5N1), A/HongKong/213/03 (H5N1)).

The 1350-base strains also show very large number of duplications in the transmembrane section - 372 duplicates in a total of 504 strains, not counting the ones already accounted for in the full sequence duplicates. It is understandable that the same strain is found locally in many hosts; e.g. thirteen strains from Bavaria, France, and Switzerland in 2006 show duplicate transmembrane sequences. Another duplicate is found in geese from Qinghai Lake in China and swans from Astrakhan in southern Russia in samples collected in 2005 (line 3 in pink in Figure 2b). Another transmembrane sequence, exemplified by A/chicken/Hebei/326/2005(H5N1), is found to be very strongly conserved, circulating from 2005 to 2008 in domestic and wild birds from China, Mongolia, Russia, Croatia, Italy, Denmark, Germany, Nigeria, Ghana, Burkina-Faso, Egypt, Kuwait, Israel, Turkey, Afghanistan, Pakistan, India, and Thailand and infecting humans in Indonesia, Egypt, and Nigeria (A/Nigeria/6e/07(H5N1)) (countries identified by a "+" symbol in pink in Figure 2b). Such sharing of the same sequence is found locally also as expected; e.g., birds of Thailand, A/quail/Thailand/NakhonPathom/QA-161/2005(H5N1) and A/golden-pheasant/Thailand/VSMU-21-SPB/2005(H5N1) have the same sequence as with humans, e.g., A/Thailand/676/2005(H5N1). Similarly, A/Ck/Thailand/73/2004(H5N1) and human isolate A/Thailand/LFPN-2004/2004(H5N1) are duplicates along with quail and chicken from Vietnam: A/quail/Vietnam/15/2005(H5N1) and A/chicken/Vietnam/TY25/2005(H5N1).

While we may expect the same sequence segment in infected populations from the same locality and same time cluster, there are several instances of the same sequence appearing with a gap of two years or more. Transmembrane sequences from, e.g. A/duck/Fujian/01/2002(H5N1) and A/duck/Shanghai/35/2002(H5N1) show the exact same strain as A/duck/Zhejiang/11/2000 (H5N1) and A/duck/Guangdong/12/2000(H5N1) (line 4 in pink in Figure 2b) but are not found in any other time period. More surprisingly, the transmembrane strain of A/Hatay/2004/(H5N1) from 2004 was found duplicated in India in 2006 in A/chicken/Navapur/Nandurbar/India/7966/2006(H5N1) (line 5 in pink in Figure 2b), two instances that are vastly separated in space and time.

Stalk Segment. Stalk segment of the neuraminidase have been linked with the degree of virulence of the H5N1;[34] mutation studies of this segment thus hold special interest. On a global basis, we find that the stalk sequences show more variation than the transmembrane region. While the stalks of the 1410 and 1353 base neuraminidases have less duplicates (6 and 3, respectively, for the nucleotide sequences, 12 and 5 for the protein sequences) than the corresponding transmembrane segments, the 1350 base neuraminidases also have lesser number of duplications (308 vs 372 out of total 504). Some clusters are localized in time and space, e.g., 10 samples from 2006 like A/swan/Bavaria/14/2006(H5N1) from south Germany, Switzerland, and France (17 more duplicates when the complete sequences considered). Exceptional instances where duplicates appear over time and space are as follows: Stalk sequence from 2003 in A/Ck/HK/WF157/ 2003(H5N1) duplicated three years later in a series of human infections in Indonesia in 2006, e.g., A/Indonesia/292H/2006(H5N1), indicated in Figure 2b by line 1 in green; interestingly, there is no evidence of this strain in 2004, but there are two instances of it in 2005, e.g. in A/chicken/Magetan/BBVW/2005(H5N1). An even greater time-separated case is that of A/duck/Guangxi/50/2001(H5N1) in China in 2001 and human infections in Indonesia in 2007, e.g. A/Indonesia/CDC1046 /2007(H5N1) (line 2 in green in Figure 2b), six years apart with no evidence of propagation of this sequence in the intervening years. An example of duplication over a wide region is provided by the stalk sequence in A/Bar-headed Goose/Qinghai/62/05(H5N1) which

is found to be duplicated in the years 2005−2006 across a wide variety of avians from China, Mongolia, Russia, Denmark, Italy, Germany, Ivory Coast, Nigeria, Burkina Faso, Saudi Arabia, Afghanistan, Pakistan, and India and human isolates from Turkey, Azerbaijan, and Iraq (countries identified by a "•" symbol in green in Figure 2b).

Body Region. Among the thirty-five 1410-base strains of the viral neuraminidase, only two A/Hongkong/Environment strains from 1999 share the same sequence; the rest all are all unique. There are no duplicates in the 1353 base nucleotide sequences which appeared in one burst of infections in 1997. In contrast, the transmembrane and stalk segments of these sequences appear to have had a considerably lesser number of mutated different strains.

For the 504 1350-base sequences for the segmentwise comparison, the body region has only 22 strains that share the same sequence with some other strains, i.e., 482 strains have accumulated mutations resulting in nonidentical sequences. The duplicate strains are for the most part restricted to localized clusters in time and space, but examples are also found across wide distances in time and space. The specially noted duplicates for the whole gene sequence will naturally also have the same sequence in the body region, e.g. the Qinghai geese and the Astrakhan swans or the samples from Egypt and Ghana mentioned earlier. While there are several instances of the same sequence found in poultry and humans, e.g., A/duck/Vietnam/N-TB/2005(H5N1) and A/Hanoi/30408/2005(H5N1), there is one instance of a cat and swan sharing the same body region sequence: A/cat/Germany/606/2006(H5N1) and A/swan/Germany/R65/2006(H5N1); while the stalk sequences of these two are different, their complete sequences are synonymous. Considering the time difference, as mentioned before in the case of the transmembrane region, A/HataY/2004/(H5N1) of 2004 is found duplicated in India in 2006 A/chicken/Navapur/Nandurbar/India/7966/2006-(H5N1) (marked by an orange line in Figure 2b); the stalk region sequences, however, are different for the Hatay and the Indian strains in only one asynonymous mutation resulting in a difference in the protein sequences for the complete genes. If we consider this to be a spontaneous mutation in the Indian isolate, since Hatay is about 2000 km away from the Indian region of Nandurbar, this shows how over time and space sequence duplication can persist.

C-Terminal Tail Segment. As mentioned in our last paper,[22] the 50-base segment at the 5′-end of the neuraminidase gene sequence is found to be very stable with only 30 unique strains out of the total 547 strains that comprise our H5N1 neuraminidase database exclusive of the 135 full sequence duplicates. It is also the only segment where several strains of the 1410 and 1353 base neuraminidase share the exact same sequence with many strains of the 1350-base strains. In case of the protein sequences, there are 534 duplicates of 3 different strains, again irrespective of the stalk lengths of the neuraminidases. Given the strong stability characteristics of this C-terminal segment, it may be worth considering this part of the protein as possible docking point for new neuraminidase inhibitors.

(c) Active Site Construct. In view of the large mutations observed in the H5N1 neuraminidase gene sequences, it is important to determine which parts remain conserved. One would expect that the active site of the protein that establishes contact with the cells of the host species would remain very well conserved. Our interest here is purely from the point of view of study of mutations in the gene and protein sequences; a proper assessment of the functions of the different parts of the protein would require of course wetlab considerations that lie outside the scope of the present work. But, to analyze theoretically the amino acids of the active site, we have considered these residues in the sequences as being of two major types as identified by Ferraris and Lina:[35]The first type consists of eight amino acids which participate in catalysis by directly contacting with substrate, and the second group consists of ten others which are framework residues and consequently contribute in the stabilization of the active site structure. To analyze and characterize these residues we have constructed a peptide by placing them in tandem sequentially and then generated the protein numerical descriptors for each of these peptides. The result showed a very conserved characteristic of the active site residues. Only 10 unique numbers are found, indicating only ten types of active site construct are present in the whole data set of 682 sequences and among them we find only one type dominant while the other nine are present in one or two copies. Further analysis on these constructs showed that only the residues R136, I203, E257, and N275 have some variety in their positions, whereas all other 14 residues are strictly conserved (Table 5). Among these four residues, one is substrate contacting and others are framework builders. The results are given in Table 5. Note that 8 of the 10 NA strains analyzed here are of the 1350 base variety, one (AF084271) is the 1353 base variety, and one (AY741222) is of 1410 base length. The residue numbers given in the top row of the table relate to residue numbers of the 1350 base NA; for the other two strains, the residue numbers have to be augmented by 1 and 20 counts, respectively, as can be seen by performing multiple alignment using ClustalX (vs 1.8).

The results of the analyses done for the present work indicate that several different pathways could have led to the distribution of the virus and composition of its sequence as we find it today. The spread of the HPAI (H5N1) virus by migratory birds is well accepted, and in at least one instance in 2007 a female garganey (*Anas querquedula*), among the most numerous of migrating duck species between Eurasia and Africa, was tracked and recorded traversing intercontinental distances from Nigeria to Russia and Turkey, sometimes covering >2000 km in <2 days, at other times stopping at some locations for several months at a time.[36] The migration path of this garganey had a spatial correlation with areas of major outbreaks of HPAI (H5N1) from 2005 to 2007. Such large scale migrations coupled with experimental evidence of viral shedding from infected birds over 4 days and the persistence of infectivity of the virus in aquatic habitats[37] indicate a possible means for the spread of the HPAI (H5N1) virus over long distances through migratory birds although conclusive proof of this is still to be established.[36]

Such a mechanism for the spread of the infection could be one explanation for the wide geographical dissemination of the H5N1 virus we have seen from our database, especially for those which are cosynchronous in terms of the years of detection. However, since mutations are known to accumulate rapidly in viral strains, estimated to be at a rate of 10^{-4} per nucleotide per replication,[38,39] it is puzzling to find identical

Table 5. Ten Varieties Active Site Construct of Proteins[a]

LOCUS ID	*98*	99	*131*	*132*	136	159	160	179	203	*205*	208	*257*	258	*273*	275	*348*	*382*	405
DQ137874	*R*	E	*D*	*R*	R	W	S	D	I	*R*	E	*E*	E	*R*	N	*R*	*Y*	E
AY651456	*R*	E	*D*	*R*	R	W	S	D	T	*R*	E	*E*	E	*R*	N	*R*	*Y*	E
DQ094292	*R*	E	*D*	*R*	R	W	S	D	I	*R*	E	*K*	E	*R*	N	*R*	*Y*	E
DQ676836	*R*	E	*D*	*R*	R	W	S	D	V	*R*	E	*E*	E	*R*	N	*R*	*Y*	E
EF222323	*R*	E	*D*	*R*	R	W	S	D	I	*R*	E	*E*	E	*R*	S	*R*	*Y*	E
CY034706	*R*	E	*D*	*R*	R	W	S	D	L	*R*	E	*E*	E	*R*	N	*R*	*Y*	E
CY034714	*R*	E	*D*	*R*	K	W	S	D	I	*R*	E	*E*	E	*R*	N	*R*	*Y*	E
CY036215	*R*	E	*D*	*R*	R	W	S	D	M	*R*	E	*E*	E	*R*	N	*R*	*Y*	E
AF084271	*R*	E	*D*	*R*	R	W	S	D	I	*R*	E	*Q*	E	*R*	N	*R*	*Y*	E
AY741222	*R*	E	*D*	*R*	S	W	S	D	I	*R*	E	*E*	E	*R*	N	*R*	*Y*	E

[a] The top row indicates residue numbers of the protein sequence of the short stalk NAs; the leftmost column lists the locus IDs. Columns in italics represent the eight catalytic residues which participate by directly contacting with substrate. Columns shaded blue show some variability in the amino acids in this position, and the mutated residues are highlighted in yellow; all other amino acids in this table are strictly conserved. Note that the column for aa 257 is a catalytic residue; unlike the other three, it shows variability and also contacts with the substrate.

H5N1 strains at places widely separated in space and time. One way this could arise is if a viral strain shed by an infected bird remained in the habitat for a considerable period before being picked up by another bird as a carrier at a later date. This would explain the occurrence of the exact same strain in birds at one or close location at two different periods, e.g., A/duck/kaifeng/1/01/(H5N1) and A/chicken/luohuo/3/03/(H5N1), both in China. (Note however that at least one report[37] suggests a potency limit of the virus of about 12 days in wetland habitats.) The occurrence of the same strain in geographically widely separated regions, e.g., A/BarheadedGoose/Qinghai/62/05(H5N1) and A/Cygnusolor/Astrakhan/Ast05-2-2/2005(H5N1) mentioned earlier, going by this hypothesis, could arise only if the virus was carried by migratory birds without infecting the carrier thus avoiding replication, e.g., infected soil samples carried on the wings or feet of the carriers. For such a scheme to explain almost identical strains spanning considerable time and space, as for example in the case of A/HataY/2004/(H5N1) of 2004 from Turkey duplicated in India in 2006 in A/chicken/Navapur/Nandurbar/ India/7966/2006(H5N1) mentioned earlier (identical except for the stalk region), would require a combination of all these mechanisms, viz. transport and assumption of long latency period of the virus enabling eventual infectivity, whereas an alternative explanation of reverse mutations leading to exact duplication would appear to be highly improbable.

The statistics of segmentwise differential mutation rates (Table 2) implies that some segments preferentially remain constant, while others either undergo mutation or an exchange with a similar segment from another strain of the virus (Table 4). Our analysis is based on neuraminidase structural segments of the H5N1 strain which is a RNA virus. Such viruses are found to utilize recombination processes in three different ways: homologous, nonhomologous, and aberrant homologous.[40] Segmented genome is a unique feature of some RNA viruses. The well established and most

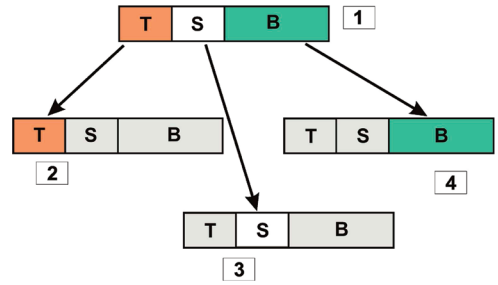

Figure 3. Schematic of RNA swapping model for the neuraminidase genes as explained in the text. Sequence no. 1 is A/chicken/Afghanistan/1573-65/2006(H5N1), sequence no. 2 is A/turkey/Islamabad/NARC7873/2007(H5N1), sequence no. 3 is A/great-crestedgrebe/Denmark/ 7498/06(H5N1), and sequence no. 4 is A/turkey/Islamabad-Pakistan/NARC-7871/02/2007 (H5N1). The three segments of the gene sequences are depicted in different colors: T (orange) represents the transmembrane segment, S (white) denotes the stalk segment, and B (green) denotes the body segment. The nonidentical segments are depicted in gray.

common finding is that this type of viral genomes can undergo genetic evolution by reassortment of the RNA segments. This mechanism accounts for antigenic shift and the selection of certain phenotypes in influenza virus, rotavirus, bluetongue virus, and others.[41,42] One evidence of interspecific recombination is found between coronavirus and influenza C virus. Though, intragenic recombination is not a well-known incident in influenza viruses, recombination between viral RNA and cellular gene has been found to increase the virulence in influenza virus.[40] Table 4 lists the frequencies of occurrence of segment exchanges between different strains; we find that such recombinations could form an important part of the H5N1 neuraminidase evolutionary processes. Figure 3 depicts a schematic of RNA swapping model from g_R similarity data. Using A/chicken/Afghanistan/1573-65/2006(H5N1) (identified as sequence number 1 in the figure) as the base model, we find that its transmembrane segment (indicated by T in the figure) is duplicated in sequence no. 2 [A/turkey/Islamabad/NARC7873/2007(H5N1)],

stalk segment (S) in sequence no. 3 [A/greatcrestedgrebe/Denmark/7498/06(H5N1)], and body segment (B) in sequence no. 4 [A/turkey/Islamabad-Pakistan/NARC-7871/02/2007(H5N1)]. Identical segments are shown in color, and variable regions are shown in gray.

In view of these possibilities of recombinations in the H5N1 RNA virus, its wide spatial and temporal dissemination and high virulence in wild birds and some animals, there has been a fear of development of human-to-human transmission leading to widespread pandemic and large scale fatalities. The methodology adopted in our previous and current paper can indicate two directions to this problem. First, we had shown in our previous work[22] that the base composition and distribution in the H5N1 gene sequences show a distinctly separate clustering compared to the older H1N1 pandemics. This can be attributed in part to the prevalence of the short stalk variety of the H5N1. The biologic importance of the neuraminidase stalk length in avian influenza virulence has been variously reported: Matsuoka et al.[43] had shown an increased virulence and decreased elution property of short stalk varieties over long stalk in H5N1 subtypes and that hemagglutinin glycosylation correlates with stalk length variety, which was also observed by Baignet and McCauley.[44] Earlier, Castrucii and Kawaoka[34] had reported from experiments in embryonating eggs that the longer the stalk the more efficient the replication, whereas stalkless NA was highly attenuated. Our g_R analysis[22] tends to support this view of the importance of the stalk length and indicates that while the H5N1 infections picked up from contact with infected avians have proved fatal for a large number of humans implying high virulence individually, transformation to a form facilitating human-to-human transmission of the old H1N1 type may not occur as long as the short stalk neuraminidase remains the dominant strain. In this connection it may not be coincidental that Shinya et al.[33] found the long stalk A/Hong Kong/213/03(H5N1) to be the most pathological to humans among the different H5N1 strains they had tested to determine the mechanism of action of the virus, but this variety has almost died out in the wild. In this connection it is also significant to note that the new Swine flu circulating in the world since April this year, 2009, through human-to-human transmission, and declared as a phase 6 pandemic by the WHO on June 11, 2009,[45] is of the long stalk H1N1 type.

Second, if a particular motif or groups of nucleotide bases or amino acid residues is determined at some time to be potentially liable to cause human-to-human infection, the techniques used in this paper for detection of duplicate sequences and sequence segments can be used to determine rapidly the presence of any such oligonucleotide or peptide stretch in the relevant sequences. A similar approach had been advocated by one of the authors in connection with genetic disease recognition in the human genome.[46] Thus with some additional inputs it should be possible using the geographical spread of identical H5N1 strains as a template to determine the potential sources of a possible H5N1 pandemic.

4. CONCLUSION

In summary, our study at the global level of mutations in H5N1 neuraminidase gene sequences available in the database shows that about 80% of the genes mutate to new strains, whereas the rest are duplicates of one or the other of these strains. At the protein level, however, only about 62% unique strains are observed, implying that about 22% of the purportedly new strains of the neuraminidase gene have synonymous mutations. Detailed study at the structural level displays wide differences in unique mutation ratios where the body region appears proportionately less stable than the transmembrane or the stalk regions. As indicated in our previous paper,[22] here also we find that the 50-base segment at the 5′-end of the gene is highly stable, mutations there being observed in less than 4.5% of the sequences, of which again only 1.2% are asynonymous. This could be potentially useful for designing novel neuraminidase inhibitors.

Among the duplicate strains we found that none of the 1353-base neuraminidase or the 1410-base neumerinidase shared the same sequence with any of the 1350-base strains except for the C-terminal tail segment. While the stalk segments are different for the three groups in length, the other segments have, in total, exactly the same number of bases but apparently have distinctive characteristics in their base composition and distributions. The 1353 base strains appeared briefly in 1997, and the 1410 base strains appeared mainly in 1999 and 2000 and have since almost died out in the wild. However, given that all three varieties had similar activity in infecting different life forms, and comprise the family of HPAI (H5N1) virus, it is interesting to note that the three groups have nevertheless such distinct identity in base composition and distribution.

We found numerous cases of sequence duplication across species and distributed over substantial distances in space and time. While localized or cosynchronous distributions can be expected to occur due to rapid dissemination of specific strains through viral shedding as one mechanism, the appearance of identical strains in geographically widely separated locations several thousand kilometers apart, or after a lapse of 2 years or more, is puzzling since viral genes mutate so rapidly in replication. We hypothesize that this may arise out of viral shedding in aquatic and nonaquatic habitats that are subsequently spread across wide regions by the migratory or local birds who themselves might be infected or act merely as carrier agents.

One important process highlighted by our analyses is recombination. We have seen that identifiable structural segments in the neuraminidase probably undergo recombination events between different strains of the virus that may have been present in the environment at the time and caused infection in the host species. While intergenic reassortments have been accepted as part of the viral evolutionary processes, evidence for intragenic homologous recombinations have not been forthcoming to any large extent to date. The H5N1 neuraminidase gene provides a possible crucible for further investigations into these aspects.

The potential for transformation of the H5N1 into a form that can be transmitted from human to human leading to a pandemic has been of serious concern. We have remarked that stalk lengths of the neuraminidase may have a direct bearing on their capacity for human-to-human transmission thus rendering certain influenza varieties highly pathogenic but that the current strains of the H5N1 are predominantly of the shorter stalk length variety that may not enable such transmission. However, development of new reassortants or

identification of specific motifs that may cause such transmission can be readily checked in our methodology with the geographical distribution helping to identify possible distribution patterns of such strains.

ACKNOWLEDGMENT

The authors gratefully acknowledge many helpful comments and suggestions from the anonymous referees that have helped to improve the presentation of this paper. One of us (A.G.) gratefully acknowledges financial support from the CSIR, India, Scheme No 37(1288)/07/EMR-II. This is publication number 491 from the Center for Water and the Environment, Natural Resources Research Institute, University of Minnesota Duluth.

REFERENCES AND NOTES

(1) Peiris, J. S. M.; Yu, W. C.; Leung, C. W.; Cheung, C. Y.; Ng, W. F.; Nicholls, 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.

(2) 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.; Auewarakul, 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.)* **2004**, *430*, 209–213.

(3) 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.

(4) Bagchi, S. Bird flu spreading in Asia. *Can. Med. Assoc. J.* **2008**, *178*, 1415.

(5) World Health Organization (WHO). http://www.who.int/csr/disease/avian_influenza/en/ (accessed March 23, 2009).

(6) Wu, W. L.; Chen, Y.; Wang, P.; Song, W.; Lau, S. Y.; Rayner, J. M.; Smith, G. J. D.; Webster, R. G.; Peiris, J. S. M.; Lin, T.; Xia, N.; Guan, Y.; Chen, H. Antigenic profile of avian H5N1 viruses in Asia from 2002 to 2007. *J. Virol.* **2008**, *82*, 1798–1807.

(7) Johansson, B. E.; Brett, I. C. Changing perspective on immunization against influenza. *Vaccine* **2007**, *25*, 3062–3065.

(8) Lal, S. K.; Chow, V. T. K. Avian Influenza H5N1 Virus: An emerging global pandemic. *Issues Infect. Dis.* **2007**, *4*, 59–77.

(9) Tumpey, T. M.; Basler, C. F.; Aguilar, P. V.; Zeng, H.; Solórzano, A.; Swayne, D. E.; Cox, N. J.; Katz, J. M.; Taubenberger, J. K.; Palese, P.; García-Sastre, A. Characterization of the reconstructed 1918 Spanish influenza pandemic virus. *Science* **2005**, *310*, 77.

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

(11) 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* (6), 769–773.

(12) Hoa, H. T.; Hurt, A. C.; Mosse, J.; Barr, I. Neuraminidase inhibitor drug susceptibility differs between influenza N1and N2 neuraminidase following mutagenesis of two conserved residues. *Antiviral Res.* **2007**, *76*, 263–266.

(13) Colman, P. M. Design and antiviral properties of influenza virus neuraminidase inhibitors. *Pure Appl. Chem.* **1995**, *67*, 1683–1688.

(14) Breschkin, J. L. M.; Sahasrabudhe, A.; Blick, T. J.; McDonald, M.; Colman, P. M.; Hart, G. J.; Bethell, R. C.; Varghese, J. N. Mutations in a conserved residue in the influenza virus neuraminidase active site decreases sensitivity to Neu5Ac2en-derived inhibitors. *J. Virol.* **1998**, *72*, 2456–2462.

(15) The World Health Organization Global Influenza Program Surveillance Network, Evolution of H5N1 Avian Influenza Viruses in Asia. *Emerging Infect. Dis.* **2005**, *11*, 1515−1521.

(16) Vijaykrishna, D.; Bahl, J.; Riley, S.; Duan, L.; Zhang, J. X.; Chen, H.; Peiris, J. S. M.; Smith, G. J. D.; Guan, Y. Evolutionary Dynamics and Emergence of Panzootic H5N1 Influenza Viruses. *PLoS Pathog. [Online] 2008* **2009**, *4* (9), e1000161(accessed March 23rd, 10.1371/journal.ppat.1000161).

(17) Lam, T. T.-Y.; Hon, C.-C.; Pybus, O. G.; Pond, S. L. K.; Wong, R. T.-Y.; Yip, C.-W.; Zeng, F.; Leung, F. C.-C. Evolutionary and Transmission Dynamics of Reassortant H5N1 Influenza Virus in Indonesia. *PLoS Pathog. [Online], 2008* **2009**, *4* (8), e1000130(accessed March 23rd, 10.1371/journal.ppat.1000130).

(18) Owoade, A. A.; Gerloff, N. A.; Ducatez, M. F.; Taiwo, J. O.; Kremer, J. R.; Muller, C. P. Replacement of Sublineages of Avian Influenza (H5N1) by Reassortments, Sub-Saharan Africa. *Emerging Infect. Dis.* **2008**, *14*, 1731–1735.

(19) Bruyere, A.; Wantroba, M.; Flasinski, S.; Dzianott, A.; Bujarski, J. J. Frequent Homologous Recombination Events between Molecules of One RNA Component in a Multipartite RNA Virus. *J. Virol.* **2000**, *74*, 4214–4219.

(20) He, C.-Q.; Xie, Z.-X.; Han, G.-Z.; Dong, J.-B.; Wang, D.; Liu, J.-B.; Ma, L.-Y.; Tang, X.-F.; Liu, X.-P.; Pang, Y.-S.; Li, G.-R. Homologous Recombination as an Evolutionary Force in the Avian Influenza A Virus. *Mol. Biol. Evol.* **2009**, *26*, 177–187.

(21) He, C.-Q.; Han, G.-Z.; Wang, D.; Liu, W.; Li, G.-R.; Liu, X.-P.; Ding, N.-Z. Homologous recombination evidence in human and swine influenza-A viruses. *Virology* **2008**, *380*, 12–20.

(22) Nandy, A.; Gute, B.; Basak, S. C. Graphical representation and numerical characterization of H5N1 avain flu neuraminidase gene sequence. *J. Chem. Inf. Model.* **2007**, *47*, 945–951.

(23) Varghese, J. N.; Laver, W. G.; Colman, P. M. Structure of the influenza virus glycoprotein antigen neuraminidase at 2.9 Å resolution. *Nature (London, U.K.)* **1983**, *303*, 35–40.

(24) Li, X.; Fang, F.; Song, Y.; Yan, H.; Chang, H.; Sun, S.; Chen, Z. Essential Sequence of Influenza Neuraminidase DNA to Provide Protection Against Lethal Viral Infection. *DNA Cell Biol.* **2006**, *25*, 197–205,. 10.1089/dna.2006.25.197.

(25) NCBI Influenza Virus Resource. http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html (accessed March 10, 2009).

(26) Nandy, A.; Harle, M.; Basak, S. C. Mathematical descriptors of DNA sequences: development and applications. *ARKIVOC* **2006**, *9*, 211–238.

(27) Liao, B.; Xiang, V.; Zhu, W. Coronavirus phylogeny based on 2D graphical representation of DNA sequence. *J. Comput. Chem.* **2006**, *27*, 1196–1202.

(28) 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.

(29) Ghosh, S.; Roy, A.; Adhya, S.; Nandy, A. Identification of new genes in human chromosome 3 contig 7 by graphical representation technique. *Curr. Sci.* **2003**, *84*, 1534–1543.

(30) Larionov, S.; Loskutov, A.; Ryadchenko, E. Chromosome evolution with naked eye: Palindromic context of the life origin. *CHAOS* **2008**, *18*, 013105.

(31) Nandy, A.; Nandy, P. On the uniqueness of quantitative DNA difference descriptors in 2D graphical representation models. *Chem. Phys. Lett.* **2003**, *368*, 102–107.

(32) Nandy, A.; Ghosh, A.; Nandy, P. Numerical characterization of protein sequences and application to voltage-gated sodium channel α subunit phylogeny. *In Silico Biol. [online] 2009* **2009**, *9*, 0008. http://www.bioinfo.de/isb/toc_vol_09.html (accessed April 25, 2009).

(33) 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.

(34) Castrucii, M. R.; Kawaoka, Y. Biologic importance of neuraminidase stalk length in influenza-A virus. *J. Virol.* **1993**, *67*, 759–764.

(35) Ferraris, O.; Lina, B. Mutations of neuraminidase implicated in neuraminidase inhibitors resistance. *J. Clin. Virol.* **2008**, *41*, 13–19.

(36) Gaidet, N.; Newman, S. H.; Hagemeijer, W.; Dodman, T.; Cappelle, J.; Hammoumi, S.; Simone, L. D.; Takekawa, J. Y. Duck Migration and Past Influenza A (H5N1) Outbreak Areas. *Emerging Infect. Dis.* **2008**, *14*, 1164–1166.

(37) Vong, S.; Ly, S.; Mardy, S.; Holl, D.; Buchy, P. Environmental contamination during influenza A virus (H5N1) outbreaks, Cambodia. *Emerging Infect. Dis. 2008* **2006**, *14*, 1303–1305.

(38) Drake, J. W.; Charlesworth, B.; Charlesworth, D.; Crow, J. F. Rates of spontaneous mutation. *Genetics* **1998**, *148*, 1667–1686.

(39) Klug, W. S.; Cummings, M. R. 14. Gene Mutation, DNA Repair, and Transposable Elements. In *Concepts of Genetics*, 7th ed. (First Indian reprints, 2003); Pearson Education (Singapore) Pte. Ltd.: Indian Branch, 482 F.I.E. Patparganj, Delhi 110092, India, 2003; Chapter 14, pp 362−364.

(40) Lai, M. M. C. RNA Recombination in animal and plant viruses. *Microbiol. Rev.* **1992**, *56*, 61–79.

(41) Fields, B. N. Genetics of reovirus. *Curr. Top. Microbiol. Immunol.* **1981**, *91*, 1–24.

(42) Palese, P. The genes of influenza virus. *Cell* **1977**, *10*, 1–10.

(43) Matsuoka, Y.; Swayne, D. E.; Thomas, C.; Welti, M. A. R.; Naffakh, N.; Warnes, C.; Altholtz, M.; Donis, R.; Subbarao, K. Neuraminidase stalk length and additional glycosylation of the hemagglutinin influence

the virulence of influenza H5N1 viruses for mice. *J. Virol.* **2009**, *83* (9), 4704–4708.

(44) Baigent, S. J.; McCauley, J. W. Glycosylation of haemagglutinin and stalk-length of neuraminidase combine to regulate the growth of avian influenza viruses in tissue culture. *Virus Res.* **2001**, *79* (1−2), 177–185.

(45) Chan, M. World now at the start of 2009 influenza pandemic; WHO website [online] http://www.who.int/mediacentre/news/statements/2009/h1n1_pandemic_phase6_20090611/en/index.html> (accessed June 30, 2009).

(46) Nandy, A. Mathematical Characterization of DNA Sequences - Towards New Directions. Computational Methods in Science and Engineering, Theory and Computation: Old Problems and New Challenge, Corfu, Greece, 2007; Maroulis, G., Simos, T., Eds.; American Institute of Physics: 2007; CP963, Vol. 1, pp 596−602.

CI9001662