Journal of General Virology (2011), **92**, 307–314 DOI 10.1099/vir.0.025973-0

Intra-host diversities of the receptor-binding domain of stork faeces-derived avian H5N1 viruses and its significance as predicted by molecular dynamic simulation

Sukathida Ubol,[1] Ampa Suksatu,[1] Naphak Modhiran,[1] Chak Sangma,[2] Arunee Thitithanyanont,[1] Mark Fukuda[3] and Tada Juthayothin[4]

Correspondence
Sukathida Ubol
scsul@mahidol.ac.th

[1]Department of Microbiology, Faculty of Science, Mahidol University, 272 Rama 6 Road, Ratchatewee, Bangkok 10400, Thailand

[2]Department of Chemistry, Faculty of Science, Kasetsart University, Bangkok, Thailand

[3]Armed Forces Research Institute of Medical Sciences, 315/6 Rajvithi Road, Bangkok 10400, Thailand

[4]National Center for Genetic Engineering and Biotechnology, National Science and Technology Development Agency, Bangkok, Thailand

Virus evolution facilitates the emergence of viruses with unpredictable impacts on human health. This study investigated intra-host variations of the receptor-binding domain (RBD) of the haemagglutinin (HA) gene of the avian H5N1 viruses obtained from the 2004 and 2005 epidemics. The results showed that the mutation frequency of the RBD ranged from 0.3 to 0.6 %. The mutations generated one consensus and several minor populations. The consensus population of the 2004 epidemic was transmitted to the 2005 outbreak with increased frequency (39 and 45 %, respectively). Molecular dynamics simulation was applied to predict the significance of the variants. The results revealed that the consensus sequence (E218K/V248I) interacted unstably with sialic acid (SA) with an α2,6 linkage (SAα2,6Gal). Although the mutated K140R/E218K/V248I and Y191C/E218K/V248I sequences decreased the HA binding capacity to α2,3-linked SA, they were shown to bind α2,6-linked SA with increased affinity. Moreover, the substitutions at aa 140 and 191 were positive-selection sites. These data suggest that the K140R and Y191C mutations may represent a step towards human adaptation of the avian H5N1 virus.

Received 30 July 2010
Accepted 22 October 2010

INTRODUCTION

Although the highly pathogenic avian H5N1 virus has spread globally, only sporadic transmission of this virus to humans has been recorded. This may be because human-to-human transmission of the avian H5N1 virus is inefficient (Cinatl *et al.*, 2007; Li *et al.*, 2004). However, the fact is that H5N1 viruses undergo continuous antigenic drift and infrequent gene reassortment, indicating that avian H5N1 viruses with increasing transmissibility among humans could be emerging. Therefore, humans remain under threat of infection by an avian H5N1 virus.

Like other RNA viruses, avian H5N1 viruses form quasispecies, which arise from rapid genomic mutations due to the absence of proofreading and post-replicative repair mechanisms of their RNA polymerase. This results in a pool of viruses containing similar genome sequences (Wan *et al.*, 2007). Although most sequence variations in this pool are neutral and offer no competitive advantage, having a pool of multiple variants enables the virus to adapt as it encounters changing environments during infection (Drake, 1999; Eigen, 1993), meaning that, under special circumstances, a point mutation may generate jumping clusters, which can move from one environment to another under different ecological pressures. For example, evidence suggests that the 1918 human H1N1 may have evolved from an avian virus that underwent point mutations in the haemagglutinin (HA) gene, yielding variants that could jump from birds to humans (Reid *et al.*, 2004; Taubenberger & Morens, 2006). It is therefore important to monitor carefully the existence of any new wave of variants, as some may emerge as the next pandemic virus.

Influenza genotype analysis reflects the influenza virus evolutionary footprint and thus is critical for preparing a strategy to prevent and control influenza epidemics and pandemics. In the present study, we monitored genotypic

variants of H5N1 viruses present in naturally infected migratory birds, Asian open-bill storks, which migrate from countries that are the epicentres of the H5N1 virus. During their migration, it is possible that storks seed the virus to various countries; therefore, it is interesting to investigate the dynamics of the H5N1 virus in storks. We used the clonal sequencing approach to investigate the H5N1 gene pool in migratory Asian open-bill storks and to determine gene linkage between the 2004 and 2005 epidemics. Furthermore, HA diversities were the focus of the present study because a major health concern is that HA mutations could alter the binding preference of the virus to that of a human receptor, thus allowing the potential infection of humans. Therefore, the significance of HA variants in storks was tested using molecular dynamics (MD) simulation. This simulation was a conformational analysis of glycan receptor binding to HA (Jongkon *et al.*, 2009). The result from this type of simulation is well matched with *in vitro* and *in vivo* HA–glycan binding (Auewarakul *et al.*, 2007; Jongkon *et al.*, 2009), making MD simulation an ideal prediction tool for host-type selectivity in emerging variants. In the present study, the sequences of the major population and some minor variants of the HA gene pools were subjected to MD simulation for prediction of receptor preference. By this approach, we were able to show that some of the epidemic variants of the HA genotypes of the H5N1 virus present in stork faeces bound to both human and avian receptors. This may be the first step towards introducing the H5N1 virus into the human population.

RESULTS

Overall picture of heterogeneities found at the receptor-binding pocket of H5N1 viruses from naturally infected storks

HA plays an important role in the initial step of influenza virus infection by binding to sialic acid (SA)-containing receptors. Thus, the diversity of the HA gene at its receptor-binding domain (RBD) is critical for H5N1 virus transmission and for determination of the host range. In this study, we examined sequence variation of HA at the receptor-binding site, aa 123–285, of H5N1 viruses. A total of seven cloacal swabs from naturally infected storks of the 2004 and 2005 epidemics were subjected to RNA purification. The purified RNAs then served as templates for cDNA synthesis using avian myeloblastosis virus (AMV) reverse transcriptase and were subsequently submitted to PCR amplification using high-fidelity *Taq* DNA polymerase. A total of 90 clones were sequenced. Sequence analysis demonstrated that the percentages of mean diversity of the RBD were approximately 0.3–0.6 at both the nucleotide and the amino acid levels (Table 1). All the cloacal swabs contain an identical major population and several minor populations. Within these 90 sequenced clones, 289 point mutations were found, which could be categorized into 244 synonymous, 43 non-synonymous and two non-coding mutations. Among the non-synonymous mutations, K140R and Y191C were the only two positive selections found. Therefore, the influence of the K140R and Y191C mutations was investigated further. In addition to substitutions, stop codons (*) and non-coding codons were detected. Three stop codons, E268*, R145* and A208*, were detected, resulting in three fatal variants. Two deletions were found at nt 374 and 590 resulting in deletion of the amino acids at positions 125 and 194.

To ensure that the observed heterogeneity was not due to nucleotide misincorporation introduced by the reverse transcriptase or the *Taq* DNA polymerase, a control experiment was carried out. The clonal sequencing of a known sequence, the E gene of dengue virus, was reverse transcribed, PCR amplified, cloned and sequenced under identical conditions. Sequence analysis of 20 independent clones showed absolute identity with the parental sequence. This information indicated that the low error rates of AMV reverse transcriptase and the high-fidelity *Taq* DNA polymerase had no influence on our results.

Genotypic variants at the receptor-binding pocket

Alignment of the deduced amino acid sequences revealed that each stork carried an identical major population of the RBD of the HA gene that was found to contain two point mutations, E218K/V248I, in comparison with A/Thailand/1(KAN-1A)/2004 (H5N1), a human isolate. Moreover, in our screening, viruses that carried this consensus HA gene were transmitted from the 2004 epidemic to the 2005 outbreak at an increased rate: 39 % in the 2004 outbreak and 45 % in the 2005 outbreak. Apart from the consensus HA gene, another four variants, minors 1–4, were of interest (Fig. 1). Both minor 1 and minor 2 contained three point mutations: K140R/E218K/V248I and E218K/V248I/E251G, respectively. The minor 1 mutant was another epidemic marker between the 2004 and 2005 outbreaks, with a rate that increased from 4 % in the 2004 epidemic to 22 % in the 2005 outbreak, suggesting the significance of this variant. The minor variants 3 and 4 contain three and four point mutations: Y191C/E218K/V248I and A214P/E218K/M226I/V248I, respectively. These five variants of the receptor-binding pocket have not been reported for the H5N1 virus elsewhere; some also contained positively selected mutation of the RBD and some served as epidemic markers. In addition, they were the major five variants found in our screening. Hence, they were chosen for further investigation.

Significance of the variants on receptor binding

Identification of mutations that can switch the currently circulating H5N1 HA receptor-binding specificity from birds to humans might provide significant leads that would help us deal with the emergence of pandemic H5N1 viruses. As several genotypic patterns were found in our

Table 1. Comparison of the characteristics of mutations found in the receptor-binding pocket of viruses from the epidemics of 2004 and 2005

Characteristic	2004	2005
No. of sequenced clones	23	67
No. of variant site nucleotides	26	51
No. of synonymous substitutions	15	29
No. of non-synonymous substitutions	14	29
List of synonymous substitutions (frequency)	P136L (1), Y137H (1), K140R (1), N154S (1), V174I (1), Y191C (1), Y197H (1), S199P (1), T204A (1), R208S (1), N244D (1), K266E (1), S267G (1), E268* (1)	Y137H (1), K140R (21), N143D (1), R145* (1), K153R (1), K161M (1), K161R (1), L172P (1), L173S (1), N182H (1), N165Y (2), T195I (1), Y197G (1), I198T (3), T202A (1), A208* (1), A214P (1), M226V (2), K234E (1), D237V (1), N240S (1), F241S (1), E242K (1), E251G (5), Y254H (1), I256Y (3), D261G (1), C274Y (1), K275R (1)
Positive selection site ($\omega \pm$ SEM)	Y191C (1.607 ± 1.283), K140R (1.838 ± 1.352)	K140R (1.917 ± 1.425)
Mean diversity (%)		
Nucleotide level	0.36	0.43
Amino acid level	0.59	0.55

HA gene pool, we thus asked how these genotypic variations might influence the receptor-binding preference. To answer this question, the sequences of five genotypic patterns, shown in Fig. 1, were subjected to MD simulation to reveal the configuration of the glycan receptor analogue in these five HA receptor-binding pockets. The receptor analogue was composed of three sugar motifs: sialic acid (SA), galactose (Gal) and *N*-acetylglucosamine with α2,3 or α2,6 glycosidic linkages. The torsion angle (Φ) – the angle between two planes consisting of O6 and C2 of the sialic acid unit, and O3 (or O6) and C3 (or C6) of the Gal sugar (Fig. 2a) – was used as the criterion to examine binding preference during computer simulations. This torsion drives the glycosides into two different configurations: *cis* (59.5°) and *trans* (−55.1°), as shown in Fig. 2(a). In the H5 binding pocket, SAα2,3Gal seems to have an optimal binding geometry when the Gal and SA units are bound in the *trans* conformation, whilst for SAα2,6Gal they are instead optimally bound in the *cis* configuration (Ha *et al.*, 2001; Li & Wang, 2006). In order to use the relationships

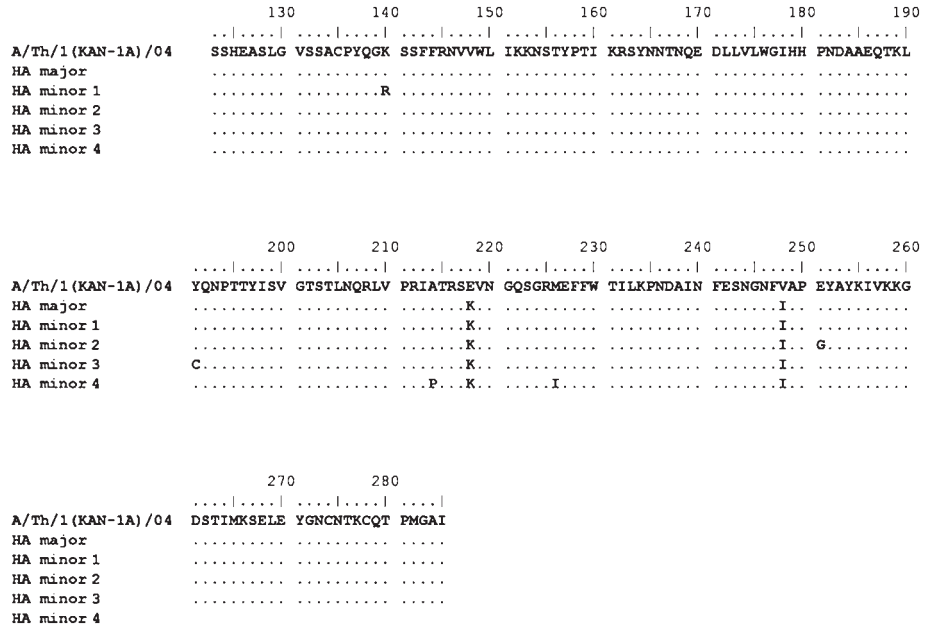

Fig. 1. Amino acid sequence alignment of the HA sequence at the receptor-binding site from aa 123 to 285 (H5 numbering).

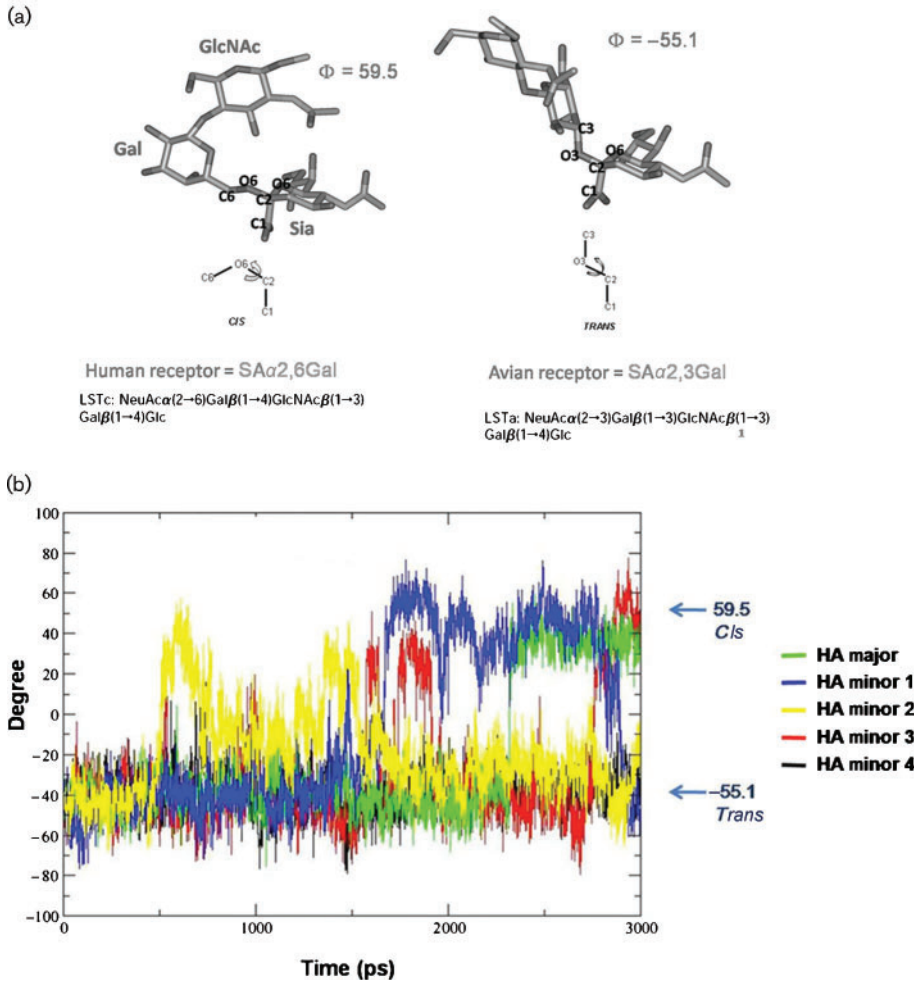

Fig. 2. (a) Structures of SAα2,6Gal and SAα2,3Gal with the atom definitions and defined torsion angle (Φ) used in this experiment. (b) Amount of time that each receptor analogue spent with a particular Φ for SAα2,3Gal in the binding pocket of HA.

between Φ and binding preference to explain host selectivity, homology modelling and MD simulations were employed. During each MD simulation, Φ was monitored and interpreted in terms of binding preference.

In our simulations, all HA genotypes shared similar binding properties to the SAα2,3Gal receptor analogue during the first 1500 ps of simulation. Thus, the SA receptor in the HA pocket spent most of its time in the *trans* conformation (Fig. 2b). However, the torsion angle of this avian receptor analogue during complexing with the HA major, minor 1 and minor 3 sequences was changed from the *trans* conformation in the initial input structure to the *cis* form after 1800 ps of simulation. These data suggested that mutations present in the HA major, minor 1 and minor 3 sequences may allow these populations to bind to both SAα2,3Gal and SAα2,6Gal, whilst the HA minor 2 and minor 4 sequences showed an SAα2,3Gal binding preference. To investigate further whether these three HA genetic variations have acquired dual receptor preference, the HA sequences were subjected to simulation using an SAα2,6Gal receptor analogue. In this experiment, HA from A/Thailand/1(KAN-1A)/2004 (H5N1) and human H1N1 were run as positive controls for SAα2,3Gal and SAα2,6Gal specificity, respectively. As demonstrated in Fig. 3(f), human H1 showed a mean SAα2,6Gal Φ angle of 59.5°, indicating a *cis* conformation. In contrast, the binding of H5 KAN-1A to SAα2,6Gal was unstable in the low-energy *cis* configuration; thus, it was forced into the *trans* conformation throughout the simulation period (Fig. 3a–e). These data indicated that MD simulation and Φ angle can be used as indicators of receptor preference. The tested mutants bound to SAα2,6Gal differently (Fig. 3a–e). The minor 1 (K140R/E218K/V248I) and minor 3 (Y191C/E218K/V248I) sequences bound strongly to SAα2,6Gal in the *cis* configuration, whilst the consensus sequence (E218K/V248I) interacted reversely between the *cis* and *trans* forms (Fig. 3a, b, d). The interaction between SAα2,6Gal and the minor 2 (E218K/V248I/E251G) or minor 4 (A214P/E218K/M226I/V248I) sequences forced a change from their *cis*

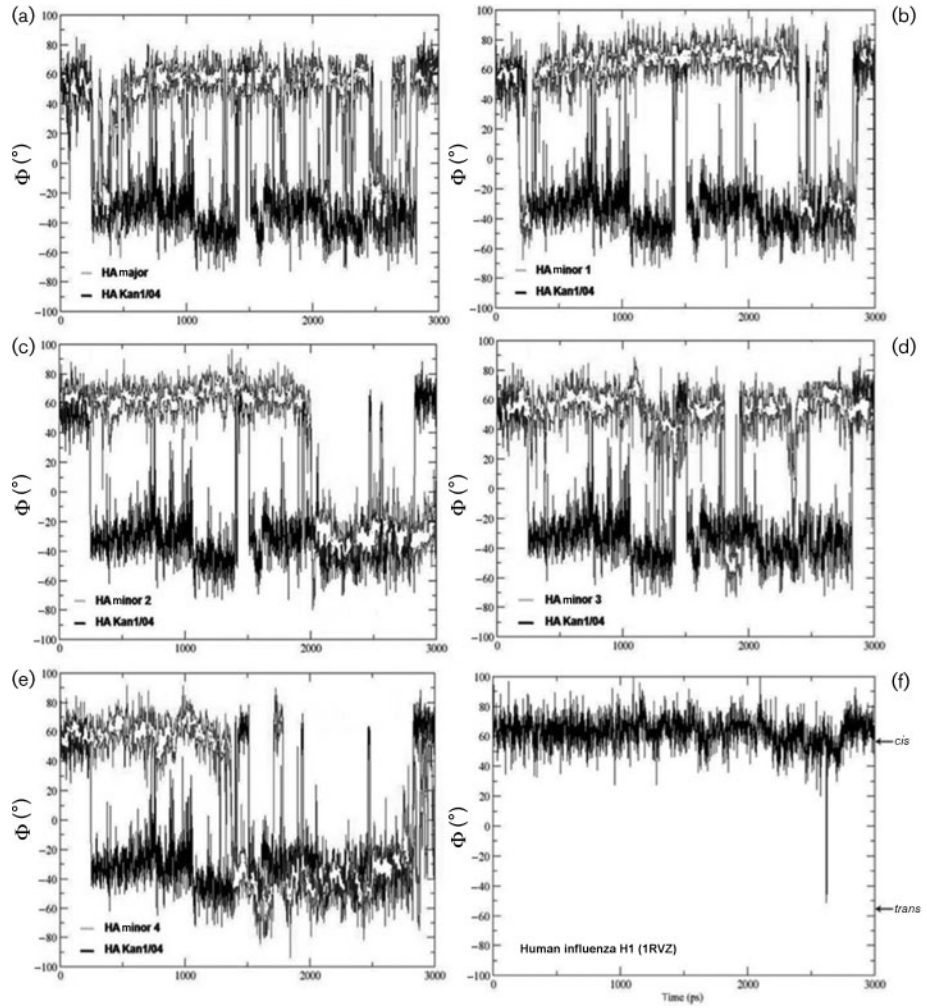

Fig. 3. Amount of time that each receptor analogue spent at a particular torsion angle (Φ) for SAα2,6Gal in the binding pocket of HA variants (grey) compared with wild-type KAN-1 HA (black).

configuration in the initial input structure to a *trans* conformation, suggesting that the minor 2 and minor 4 sequences are unfavourable as SAα2,6 Gal analogues.

In conclusion, this result indicated that there were at least two HA sequences found in infected storks, K140R/E218K/V248I and Y191C/E218K/V248I, that have mixed specificity, with a reduction in binding to the α2,3-linked SA avian receptor, but with increasing affinity for the α2,6 human-type receptor.

DISCUSSION

Outbreaks of the H5N1 avian influenza virus have caused worldwide concern for both animal and human health. Thailand is one of the affected countries that have experienced both avian and human infections. H5N1 outbreaks in this country have occurred primarily in household chickens and free-grazing ducks, and the virus is believed to be imported via migratory birds. Of these migratory species, the Asian open-bill stork is probably the most affected wild-bird species in Thailand. In addition, these wild birds migrate from the epicentre of the occurrence of the H5N1 virus; therefore, they must harbour both parental viruses and their variants. Thus, naturally infected storks should be a good model to study the H5N1 virus footprint.

The breaching of host-range barriers for H5N1 virus is due mainly to modification of the HA and neuraminidase. This modification can be mediated through several mechanisms such as antigenic drift and additional glycosylation (Chakrabarti *et al.*, 2009; Matrosovich *et al.*, 1999). It is unclear whether this modification occurs in the new host or in the original host. Thus, in the present study, we investigated intra-host variations of the receptor-binding pocket where the viral genome was purified, amplified by RT-PCR, cloned and sequenced from cloacal swabs of naturally infected migratory storks. We found that the

mutation frequency in these migratory storks was slightly higher than that reported previously in human influenza virus and in other host systems (Hoelzer *et al.*, 2010; Iqbal *et al.*, 2009). This higher mutation frequency was unlikely to be due to artefacts introduced during the experimental amplification procedure because the error rate of the AMV reverse transcriptase used is 2.7×10^{-5} for base substitution, whilst high-fidelity *Taq* DNA polymerase has a mutation frequency of 0.62 % based on a positive-selection assay (Fujii *et al.*, 1999; Roberts *et al.*, 1989). Although artefactual mutations were unlikely to have been introduced, our data should still be interpreted with caution. The explanation for the higher mutation rate found in our study is unclear, but could be due to several factors. For example, it may be due to the effect of the microenvironment in the intestinal tract of storks or possibly due to the duration of infection. A higher mutation frequency is reported in H5N1 patients with a longer infection period (Kongchanagul *et al.*, 2008). This may be due to the selection pressure from host immune responses, which is known to be a major driving force for influenza virus antigenic drift (Hensley *et al.*, 2009; Wei *et al.*, 2010). However, the best explanation at this moment is the differences in viruses and experimental models used. Iqbal *et al.* (2009) and Murcia *et al.* (2010) investigated the diversity of influenza viruses that have long been adapted in the laboratory and in the models used, whilst our report studied emerging virus from the reservoir that is naturally and initially transmitted into storks. Thus, Iqbal *et al.* (2009) and Murcia *et al.* (2010) reported the dynamics of influenza virus under a mild selective pressure, whilst our work has revealed the mutation rates of emerging H5N1 virus under a strong selective pressure. Therefore, it is not surprising that we found greater dynamics of influenza virus evolution.

The interaction between a virus and the host-cell receptor is an essential step in the complex process of cell infection. For influenza virus, the receptor-binding domain is formed by the 130 loop (residues 134–138), the 220 loop (residues 221–228) and the 190 loop (residues 188–190), based on H3 HA (Das *et al.*, 2009). A few mutations in this binding domain can switch receptor specificity from birds to humans. For example, mutations at aa 129 and 134 (L129V/A134V) can change the receptor-binding preference of the H5N1 virus from SAα2,3Gal to both SAα2,3Gal and SAα2,6Gal (Auewarakul *et al.*, 2007). In addition, N182K and Q192R mutations have been shown to enhance binding of a Vietnam H5N1 virus to the SAα2,6Gal receptor (Ha *et al.*, 2001). In the present study, heterogeneities of HA at the RBD were investigated. The results revealed that more than 70 variants existed in the 2004–2005 epidemics, but only two epidemic markers were found, the E218K/V248I and K140R/E218K/V248I variants, indicating that these two genetic variants were true positively selected populations. The significance of these epidemic markers for bird-to-human transmission was predicted using MD simulation, which confirmed that the major population, E218K/V248I, showed a binding preference for the avian-type receptor. In contrast, the K140R/E218K/V248I variant was shown to bind to the avian-type receptor with reduced affinity while acquired a binding preference for the human-type receptor. In addition to the K140R variant, a Y→C substitution at position 191 gave this variant the ability to bind both SAα2,3Gal and SAα2,6 Gal. In comparison with wild-type HA from the KAN-1A/2004 virus, mutations of K140R and Y191C, respectively allowed these RBD variants to form a stronger complex with the SAα2,6Gal receptors. Our finding is in agreement with previous reports that mutations in or adjacent to one of these loops are significant enough for receptor-type switching. The mechanism of how these substitutions induce conformational change in the RBD requires further study.

In conclusion, the present report investigated genetic pools of H5N1 viruses during the 2004 and 2005 epidemics in migratory storks. These genetic pools contained at least two RBD variants that were able to bind both human-type and avian-type receptors. Importantly, one of these gene variants, the K140R species, was one of the variants that followed the transmission chain. An epidemic marker that acquires dual-receptor preference may be a threat to human health and thus deserves to be monitored further. Finally, our information should be analysed carefully, as only the viral genome was investigated. Hence, some of the mutations found might not support actual virus replication.

METHODS

Avian H5N1 viruses. Seven cloacal swabs from naturally infected Asian open-bill storks were used in our study. These swabs were obtained from dead, infected Asian open-bill storks in Bung Boraphet, NakhonSawan province, Thailand, during the 2004 and 2005 epidemics. The swabs were collected in a virus transport medium containing medium M199 (Invitrogen), 1000 U penicillin G ml^{-1}, 1000 μg streptomycin ml^{-1}, 400 μg gentamicin ml^{-1}, 2.5 μg Fungizone ml^{-1} and BSA and kept at −80 °C. The number of viruses in each cloacal swab was quantified and expressed as H5N1 RNA copy number. These RNAs were subjected directly to clonal sequencing.

Viral RNA extraction and RT-PCR amplification. Viral RNA was isolated from cloacal swab samples with an RNeasy Mini kit (Qiagen) according to the manufacturer's instructions. The viral RNA was reverse transcribed into cDNA using AMV reverse transcriptase (Promega). PCR amplification of the HA gene segment was carried out using a pair of specific primers and the Platinum *Taq* DNA Polymerase High Fidelity (Invitrogen). The HA primer set used in this experiment was 5′-GTTCCAGTCATGAAGCCTCA-3′ (sense) and 5′-TTTATCGCCCCCATTGGAGT-3′ (antisense).

Clonal sequencing. Purified PCR products were cloned into sequencing vector pCR2.1-TOPO as recommended by the manufacturer (TA Cloning kit; Invitrogen). At least 30 well-isolated white colonies were picked and subsequently cultivated in 5 ml Luria–Bertani broth containing 100 mg ampicillin (Invitrogen) ml^{-1} at 37 °C overnight. Plasmid DNAs were extracted from harvested, transformed *Escherichia coli* by the alkaline lysis method and sequenced. The nucleotide sequences and putative amino acid sequences of the H5N1 variants pool were compared using BioEdit

version 7.0.1.2 (Hall, 1999). Pairwise comparison of each nucleotide sequence was performed using MEGA version 4 (Tamura *et al.*, 2007).

Selection at the protein level was measured using the ratio of non-synonymous to synonymous substitution sites (d_N/d_S) in which a value of $\omega>1$ is referred to as being under positive selection, as described previously (Kongchanagul *et al.*, 2008; Yang, 1997; Yang *et al.*, 2005). The codon-based models M7 and M8 of the CODEMEL program in PAML were used. The M7 model contains ten ω categories to describe ω among sites where all are <1. The M8 model was used to estimate ω for an extra class of sites (p10) at which ω can be >1. These two models were compared using a likelihood ratio test, whilst the positive selection of individual codons was estimated using Bayes empirical Bayes method.

MD simulations of HA. The crystal structures of H5 HA from A/Thailand/1(KAN-1A)/2004 (H5N1) (GenBank accession no. EF107522) and from A/Duck/Singapore 3/97 (H1N1) (PDB accession no. 1RVZ) were used as templates for binding to SAα2,3Gal (PDB accession no. 1JSN) and SAα2,6Gal (PDB accession no. 1JSO), respectively. A three-sugar unit consisting of SA, Gal and *N*-acetylglucosamine was used as the input structure for MD simulation, as described previously (Jongkon *et al.*, 2009). The Φ between the two planes of SA and Gal residues was focused. In the homology modelling, the wild-type HA and mutant HA molecules from major and minor populations were three-dimensionally aligned via the SWISS-MODEL server (Schwede *et al.*, 2003). MD simulations were performed at a 3 ns production run and under pressure with 0.002 ps time steps using the SANDER mode in AMBER version 9 simulation software, as described elsewhere (Jongkon *et al.*, 2009). The utility programs Xmgrace and VMD were used to visualize and render all the figures presented in this paper (Das *et al.*, 2009; Tumpey *et al.*, 2007).

ACKNOWLEDGEMENTS

This work was supported by research grants from the National Science and Technology Development Agency (BT-B-01-MG-14-50092) to S. U., from the National Institute of Health and the National Institute of Allergy and Infectious Diseases (YI-AI-5026-01) to M. F. and from the Thailand Graduate Institute of Science and Technology (TGIST) to A. S., who received a scholarship and financial support during graduate study.

REFERENCES

Auewarakul, P., Suptawiwat, O., Kongchanagul, A., Sangma, C., Suzuki, Y., Ungchusak, K., Louisirirotchanakul, S., Lerdsamran, H., Pooruk, P. & other authors (2007). An avian influenza H5N1 virus that binds to a human-type receptor. *J Virol* **81**, 9950–9955.

Chakrabarti, A. K., Pawar, S. D., Cherian, S. S., Koratkar, S. S., Jadhav, S. M., Pal, B., Raut, S., Thite, V., Kode, S. S. & other authors (2009). Characterization of the influenza A H5N1 viruses of the 2008–09 outbreaks in India reveals a third introduction and possible endemicity. *PLoS ONE* **4**, e7846.

Cinatl, J., Jr, Michaelis, M. & Doerr, H. W. (2007). The threat of avian influenza A (H5N1). Part I: epidemiologic concerns and virulence determinants. *Med Microbiol Immunol (Berl)* **196**, 181–190.

Das, P., Li, J., Royyuru, A. K. & Zhou, R. (2009). Free energy simulations reveal a double mutant avian H5N1 virus hemagglutinin with altered receptor binding specificity. *J Comput Chem* **30**, 1654–1663.

Drake, J. W. (1999). The distribution of rates of spontaneous mutation over viruses, prokaryotes, and eukaryotes. *Ann N Y Acad Sci* **870**, 100–107.

Eigen, M. (1993). Viral quasispecies. *Sci Am* **269**, 42–49.

Fujii, S., Akiyama, M., Aoki, K., Sugaya, Y., Higuchi, K., Hiraoka, M., Miki, Y., Saitoh, N., Yoshiyama, K. & other authors (1999). DNA replication errors produced by the replicative apparatus of *Escherichia coli*. *J Mol Biol* **289**, 835–850.

Ha, Y., Stevens, D. J., Skehel, J. J. & Wiley, D. C. (2001). X-ray structures of H5 avian and H9 swine influenza virus hemagglutinins bound to avian and human receptor analogs. *Proc Natl Acad Sci U S A* **98**, 11181–11186.

Hall, T. A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp Ser* **41**, 95–98.

Hensley, S. E., Das, S. R., Bailey, A. L., Schmidt, L. M., Hickman, H. D., Jayaraman, A., Viswanathan, K., Raman, R., Sasisekharan, R. & other authors (2009). Hemagglutinin receptor binding avidity drives influenza A virus antigenic drift. *Science* **326**, 734–736.

Hoelzer, K., Murcia, P. R., Baillie, G. J., Wood, J. L., Metzger, S. M., Osterrieder, N., Dubovi, E. J., Holmes, E. C. & Parrish, C. R. (2010). Intrahost evolutionary dynamics of canine influenza virus in naive and partially immune dogs. *J Virol* **84**, 5329–5335.

Iqbal, M., Xiao, H., Baillie, G., Warry, A., Essen, S. C., Londt, B., Brookes, S. M., Brown, I. H. & McCauley, J. W. (2009). Within-host variation of avian influenza viruses. *Philos Trans R Soc Lond B Biol Sci* **364**, 2739–2747.

Jongkon, N., Mokmak, W., Chuakheaw, D., Shaw, P. J., Tongsima, S. & Sangma, C. (2009). Prediction of avian influenza A binding preference to human receptor using conformational analysis of receptor bound to hemagglutinin. *BMC Genomics* **10**, S24.

Kongchanagul, A., Suptawiwat, O., Kanrai, P., Uiprasertkul, M., Puthavathana, P. & Auewarakul, P. (2008). Positive selection at the receptor-binding site of haemagglutinin H5 in viral sequences derived from human tissues. *J Gen Virol* **89**, 1805–1810.

Li, M. & Wang, B. (2006). Computational studies of H5N1 hemagglutinin binding with SAα-2,3-Gal and SA-α-2,6-Gal. *Biochem Biophys Res Commun* **347**, 662–668.

Li, K. S., Guan, Y., Wang, J., Smith, G. J., Xu, K. M., Duan, L., Rahardjo, A. P., Puthavathana, P., Buranathai, C. & other authors (2004). Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia. *Nature* **430**, 209–213.

Matrosovich, M., Zhou, N., Kawaoka, Y. & Webster, R. (1999). The surface glycoproteins of H5 influenza viruses isolated from humans, chickens, and wild aquatic birds have distinguishable properties. *J Virol* **73**, 1146–1155.

Murcia, P. R., Baillie, G. J., Daly, J., Elton, D., Jervis, C., Mumford, J. A., Newton, R., Parrish, C. R., Hoelzer, K. & other authors (2010). Intra- and interhost evolutionary dynamics of equine influenza virus. *J Virol* **84**, 6943–6954.

Reid, A. H., Taubenberger, J. K. & Fanning, T. G. (2004). Evidence of an absence: the genetic origins of the 1918 pandemic influenza virus. *Nat Rev Microbiol* **2**, 909–914.

Roberts, J. D., Preston, B. D., Johnston, L. A., Soni, A., Loeb, L. A. & Kunkel, T. A. (1989). Fidelity of two retroviral reverse transcriptases during DNA-dependent DNA synthesis in vitro. *Mol Cell Biol* **9**, 469–476.

Schwede, T., Kopp, J., Guex, N. & Peitsch, M. C. (2003). SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res* **31**, 3381–3385.

Tamura, K., Dudley, J., Nei, M. & Kumar, S. (2007). MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol Biol Evol* **24**, 1596–1599.

Taubenberger, J. K. & Morens, D. M. (2006). 1918 influenza: the mother of all pandemics. *Emerg Infect Dis* **12**, 15–22.

Tumpey, T. M., Maines, T. R., Van Hoeven, N., Glaser, L., Solorzano, A., Pappas, C., Cox, N. J., Swayne, D. E., Palese, P. & other authors (2007). A two-amino acid change in the hemagglutinin of the 1918 influenza virus abolishes transmission. *Science* **315**, 655–659.

Wan, X. F., Chen, G., Luo, F., Emch, M. & Donis, R. (2007). A quantitative genotype algorithm reflecting H5N1 avian influenza niches. *Bioinformatics* **23**, 2368–2375.

Wei, C.-J., Boyington, J. C., Dai, K., Houser, K. V., Pearce, M. B., Kong, W.-P., Yang, Z.-Y., Tumpey, T. M. & Nabel, G. J. (2010). Cross-neutralization of 1918 and 2009 influenza viruses: role of glycans in viral evolution and vaccine design. *Sci Transl Med* **2**, 24ra21.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13**, 555–556.

Yang, Z., Wong, W. S. & Nielsen, R. (2005). Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol* **22**, 1107–1118.