VIROLOGICA SINICA, October 2007, 22 (5):405-411
CLC number: Q78 Document code: A Article ID: 1674-0769(2007)05-0405-07

Proteotyping: A New Approach Studying Influenza Virus Evolution at the Protein Level*

Wei-feng SHI[1], Zhong ZHANG[2], Lei PENG[3], Yan-zhou ZHANG[4], Bin LIU[5] and Chao-dong ZHU[4**]

(1. Institute of Life Sciences, Taishan Medical College, Shandong Tai'an, 271000 China; 2. Department of Basic Medicine, Taishan Medical College, Shandong Tai'an, 271000 China; 3. College of Information and Engineering, Taishan Medical College, Shandong Tai'an, 271000 China; 4.Institute of Zoology, Chinese Academy of Sciences, Beijing, 100101 China; 5. Department of Biological Sciences, Taishan Medical College, Shandong Tai'an, 271000 China)

Abstract: Phylogenetic methods have been widely used to detect the evolution of influenza viruses. However, previous phylogenetic studies of influenza viruses do not make full use of the genetic information at the protein level and therefore cannot distinguish the subtle differences among viral genes. Proteotyping is a new approach to study influenza virus evolution. It aimed at mining the potential genetic information of the viral gene at the protein level by visualizing unique amino acid signatures (proteotypes). Neuraminidase gene fragments of some H5N1 avian influenza viruses were used as an example to illustrate how the proteotyping method worked. Bayesian analysis confirmed that the NA gene tree was mainly divided into three lineages. The NA proteotype analysis further suggested there might be multiple proteotypes within these three lineages and even within single genotypes. At the same time, some proteotypes might even involve more than one genotype. In particular, it also discovered some amino acids of viruses of some genotypes might co-reassort. All these results proved this approach could provide additional information in contrast to results from standard phylogenetic tree analysis.

Key words: Proteotyping; Genotype; H5N1; Avian influenza virus; Neuraminidase

Tracing influenza viruses' evolution has been the subject of much research relevant to influenza viruses and frequent reassor- tment events for both avian and human influenza viruses have been detected using phylogenetic methods (6, 13). For instance, several genotypes of H5N1 avian influenza viruses have been detected in the past five years, and these have been designated *A*, *B*, *C*, *D*, *E*, *G*, *V*, *W*, *X*, *Y*, *Z*, Z^+ and so on (4, 5, 11, 13). Viruses of genotype *A*, *B*, *C*, *D*, *E* and *F* were also identified for the H9N2 subtype (1).

Received: 2007-05-11, Accepted: 2007-07-03
* Foundation items: National Nature Science Funds (30670242, 30500056)
** Corresponding author. Tel: +86-10-64807085, E-mail: zhucd@ioz.ac.cn

There is no denying that these studies revealed differences among viruses. However, all of the differences were principally at the DNA level and genetic information at the protein level was not fully utilized. Therefore, phylogenetic trees sometimes could not provide subtle differences among viral genes.

To better make use of the information at the protein level, molecular characterization analyses and reverse genetics techniques have been performed to help find the key sites relevent to pathogenicity, virulence and even host selection of influenza viruses (15, 16) etc. Up to date, some positions playing important roles in viral genomes have been found, such as the connecting peptide sites in HA, Lys-627 in the PB2 fragment (7, 10) and so on. Thus, molecular characterization analysis does have advantages in seeking single amino acid and short peptide mutations. However, it is difficult for it to integrate all these genetic information as a whole to find genes that co-reassort or proteins displaying compensatory mutations. To this end, Obenauer *et al.* introduced a proteotyping method to visualize unique amino acid signatures (proteotypes) (17). This method was able to identify co-reassorting genes, 50+ protein-protein pairs, virus "families" that share specific combination of genes and proteins exhibiting compensatory mutations (8).

Neuraminidase (NA) is a surface protein that cleaves sialic acid from virus and host cell glycoconjugates at the end of the virus life cycle to allow mature virions to be released (25). Phylogenetic studies have revealed that the H5N1 avian influenza viruses of China were divided into three lineages according to the NA gene tree, with one lineage (I) possessing a 19-aa deletion in the stalk of NA, one lineage (II) without deletion, and one lineage (III) with a 20-aa deletion (9,24). Viruses of genotypes *A*, *G*, *X*, *Y*, *Z* and ShanTou3-like (ST3-like) belonged to group III, while *B*, *C*, *D*, *E*, *W*, Z^+, ST1-like and ST2-like isolates belonged to group II and HK/156/97 was placed into group I.

In this paper, we took NA gene fragments of some H5N1 influenza viruses isolated from mainland China, Hong Kong Special Administration Region (SAR) and Southern Asia as an example to illustrate how the proteotyping method worked.

DATASET AND METHODS

Our dataset included the NA gene segments of typical H5N1 avian influenza viruses of the known genotypes isolated from mainland China, Hong Kong SAR and Southeast Asia. In addition, some isolates from human were also included to assist our analysis. Parrot/Ulster/73 was designated as an outgroup to root the tree. All nucleotide sequences were obtained directly from GenBank.

The first step in proteotyping was similar to that of normal phylogenetic analysis. Multiple sequence alignment was performed with ClustalX 1.81 (23) and the alignment parameters were set to default. To estimate the trees accurately, MrBayes, version 3.0b4 was used to construct the NA gene tree (19). Four Markov chains were run for two million generations and sampled every 100 generations to yield a posterior probability distribution of 20 000 trees. After eliminating the first 5 000 trees as burn-in, a 50% majority-rule consensus tree was constructed. Bayesian Posterior Probability (BPP) was used to assess the support for the recovered clades, given the aligned sequence data. A six parameter substitution model (General Time Reversible) was used with a gamma rate parameter

allowing site variation. It should be noted that besides Bayesian, other trees search methods can also be used.

In the second step, DNA data were translated to their protein sequences accordingly by using Mega 3 (12). Alternatively, protein sequences could be downloaded directly from GenBank. After that, protein sequences were aligned using ClustalW included in Mega 3. The protein alignment was then re-sorted according to the sequence order displayed by the tree. Consequently, a so-called "clade-guided" sequence alignment was produced by assigning a unique color to each kind of amino acid. It also should be noted that leading and trailing gaps were generally artifacts of aligning sequences with different 5' and 3' termini (22) and were set to white. The remaining gaps were set to black in order to highlight the real amino acid deletions.

Thirdly, a consensus sequence was calculated for the alignment. All the consensus amino acids were set to white to match the background color so that only non-consensus sites were visible. Obenauer *et al*. proposed a residue occur more than any other residue to be the consensus (17). However, by our method, all the residues would be displayed if no residue occurred more than 50% in the column. The remaining residues were used to define the proteotype according to the numbers of variable amino acids among proteins.

Finally, the proteotypes of NA proteins of the representative H5N1 viruses were determined mainly based on the amino acid differences among protein sequences and position information of the sequences on the tree. After the proteotypes are determined, serial numbers will be assigned starting at the top downwards for each proteotype and these numbers would be summarized into a table 1. At the same time, unique amino acids were sought from the NA proteotypes.

RESULTS

Bayesian Analysis

The NA gene tree was mostly divided into two major lineages with a small branch out of them (Fig. 1). One major lineage involved viruses of genotype *A*, *G*, *X*, *Y*, *Z*, while the other involved genotype *B*, *C*, *D*, *E*, *W*, Z^{+}.

Proteotyping Analysis

Proteotypes of NA proteins supported the phylogeny revealed by the NA gene tree (Fig. 1). However, there were some differences between the results of the phylogenetic and proteotyping analyses. First of all, the proteotyping analysis displayed protein differences within the lineage and even within the genotype. For example, the differences of NA proteins among the viruses of the *X* genotype were observed. Likewise, our results indicated that the *Z* genotype viruses might be further divided into more proteotypes (Fig. 1). Secondly, some proteotypes might involve more than one genotype. For instance, some viruses of genotype *X*, *A*, *Y* and *Z* were defined as the same proteotype - p1.2 (Fig. 1, Table 1). Finally, some co-variable amino acids that might be potentially important to maintain the advanced structures and functions of the proteins were found. Particularly, it was possible that Thr17, Lys64, Asn75, His233 and Ser320 co-evolved in the NA proteins of some Southeast Asia isolates of Genotype *Z* (Fig.1). It also suggested some sites of viruses of genotype *W* might evolve with each other (Fig. 1).

DISCUSSION

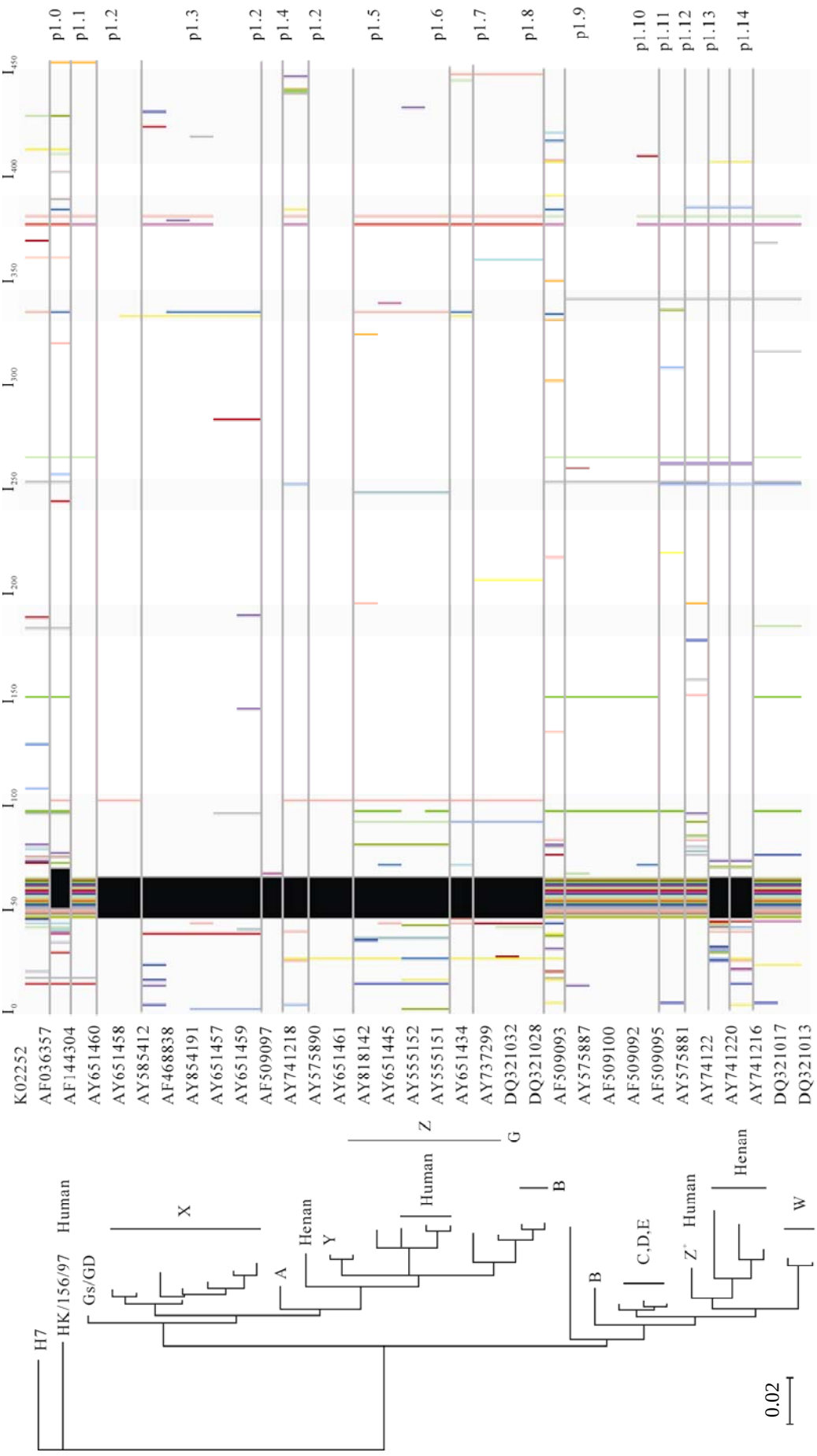

Fig. 1. Proteotypes for NA genes/proteins of some H5N1 avian and human influenza viruses. Phylogenetic analysis was based on nucleotides 20-1426 (1,407 bp) of the NA gene and the tree was rooted to K02252 (A/Parrot/Ulster/73, H7N1). Following the GenBank accession numbers there was the corresponding genotype or host information of these viruses. Scale bar, 0.02 nucleotide change per site. The left column was the GenBank accession numbers of the representative H5N1 avian and human influenza viruses. The protein alignment was adjusted according to the sequence orders of the viruses in the NA gene tree. The right column was the serial numbers designated to the NA proteotypes respectively.

Table 1. Some representative H5N1 avian and human influenza viruses and their corresponding NA proteotypes

GenBank NO.[a]	Virus strains	Proteotype
AF036357	HongKong/156/97	p1.0
AF144304	Goose/Guangdong/1/96	p1.1
AY651460	Sck/HK/YU100/2002	p1.2
AY651458	Ck/HK/31.2/2002	p1.2
AY651412	Duck/Guangxi/53/2002	p1.3
AF468838	Duck/Anyang/AVL-1/2001	p1.3
AY854191	Duck/Shandong/093/2004	p1.3
AY651457	Gf/HK/38/2002	p1.3
AY651459	Ck/HK/37.4/2002	p1.3
AF509097	Silky Chicken/HongKong/SF189/ 01	p1.2
AY741218	Tree sparrow/Henan/2/2004	p1.4
AY575890	Ck/HK/96.1/02	p1.2
AY651461	Ck/HK/YU22/2002	p1.2
AY818142	Chicken/Vietnam/C58/04	p1.5
AY651445	Viet Nam/1194/2004	p1.5
AY555152	Thailand/2(SP-33)/2004	p1.5
AY555151	Thailand/1(KAN-1)/2004	p1.5
AY651434	Dk/Indonesia/MS/2004	p1.6
AY737299	Chicken/Guangdong/178/04	p1.7
DQ321032	Duck/Guangxi/951/2005	p1.7
DQ321028	Goose/Guangxi/345/2005	p1.7
AF509093	Chicken/Hong Kong/YU562/01	p1.8
AY575887	Ck/HK/31.4/02	p1.9
AF509100	Chicken/Hong Kong/715.5/01	p1.9
AF509092	Chicken/Hong Kong/FY77/01	p1.9
AF509095	Chicken/Hong Kong/FY150/01	p1.9
AY575881	HK/212/03	p1.10
AY741222	Tree sparrow/Henan/4/2004	p1.11
AY741220	Tree sparrow/Henan/3/2004	p1.12
AY741216	Tree sparrow/Henan/1/2004	p1.13
DQ321017	Duck/Guangxi/1586/2004	p1.14
DQ321013	Goose/Guangxi/1097/2004	p1.14

[a]The virus strains were listed according to the sequence orders in the NA gene tree.

Proteotyping is a recently introduced method akin to genotyping at the DNA level, but wichit additionally captures the variability of proteins as they occur in populations and change over time (20). Using this method to help find the proteins related to diseases and study the changes of these proteins both in healthy and morbid situations has been reported (2, 3, 26). It has also been used as a tool to study developmental lesions (21). Some researchers even used it to link genotype and phenotype of some diseases (14). In these studies, the proteotyping processes were often fulfilled by the assistance of mass spectrum (MS) (18, 20).

Method has also been reported to be used to study influenza virus evolution at the protein level (17). In this study, it is have modified and has some particular characteristics. First of all, Proteotyping analysis is principally sequence-based, and therefore it can be completed without MS data. Secondly, as mentioned in the method section, the protein sequence alignment has been changed to clade-guided rather than normal multiple sequence alignment. Thirdly, for influenza viruses, genotype is only determined by the whole genome rather than single or few genes of it. In contrast, the proteotypes of the viruses can be determined for both each gene of the virus and the whole genome. In fact, integrating all the eight proteotypes determined for each gene segment, one can ascertain the proteotype of the whole genome like what have done to define a genotype of an influenza virus. Fourthly, the serial numbers designated to the proteotypes of the same viruses may be different because the serial numbers are decided both by the sample size into analysis and by the positions of the viruses in the gene tree. At last, the proteotype can also be linked to genotype. In fact, information at the genotype level is helpful to define the proteotype.

Bayesian analysis in this paper confirmed the previously constructed topology (9, 24). However, differences between the results from phylogenetic and proteotyping analyses proved that the proteotyping

method had a higher resolution and was able to mine more subtle differences among viruses. The specific amino acids found by the proteotyping method could be further analyzed by other bioinformatics techniques or reverse genetic techniques to study their potential biological functions. However, only the proteotypes of NA proteins were determined here (Table 1). If the proteotypes of all the eight proteins of the influenza viruses were identified, proteins co-reassorting or showing compensatory mutations could be detected (17).

Unlike the consensus definition proposed by Obenauer *et al.*(17), we suggest all the residues should be displayed if no residue occurs more than 50% in the column. If none of them are occurring more than 50%, this may be an indication that this site is super variable. Although a super variable site suggests weak selective pressure and absence of biological function, if these variable sites were neglected, it is difficult to find sites that might co-reassort and they would be hidden subjectively. It is likely that these coreassorting sites might be related to the function of the protein. Therefore, hiding the residue taking up less than 50% in the column might lose potentially important information.

It should be also mentioned that there is no general criterion available to guide the definition of a proteotype. If it is defined arbitrarily, potentially, useful information might be hidden by the noise. However, the number of different amino acids among proteins and the positions of the viruses in the gene tree may be helpful to distinguish different proteotypes, but in some cases this may not be sufficient. Additional factors should be also taken into account such as serotype, subtype, genotype, host, collection time and natural selection pressure. In particular, sample size is also an important factor that should not be ignored. However, in this paper, we mainly introduced the proteotyping method and therefore only a sample of small size was used and the proteotypes were designated mostly by the numbers of different amino acids of NA proteins. Consequently, the proteotypes designated here were not strict.

To sum up, proteotyping method is a useful tool for studying virus evolution at the protein level. It also can be applied to other viruses, especially to viruses with segmented genomes.

Acknowledgement

The authors are indebted to John C. Obenauer of the Hartwell Center for Bioinformatics and Biotechnology for technical assistance. This study was partially funded by The Youth Science Funds of Taishan Medical College (2007QNZR056).

References

1. **Choi Y K, Ozaki H, Webby R J, *et al.*** 2004. Continuing Evolution of H9N2 Influenza Viruses in Southern China. **J Virol,** 78: 8609-8614.
2. **Cummings J L.** 2003. Toward a molecular neuropsychiatry of neurodegenerative diseases. **Ann Neurol,** 54 (2): 147-154.
3. **Cummings J L.** 2004. Dementia with Lewy Bodies: Molecular Pathogenesis and Implications for Classification. **J Geriatr Psychiatry Neurol,** 17 (3): 112-119.
4. **Guan Y, Peiris J S M, Lipatov A S, *et al.*** 2002. Emergence of multiple genotypes of H5N1 avian influenza viruses in Hong Kong SAR. **Proc Natl Acad Sci USA,** 99: 8950-8955.
5. **Guan Y, Poon L L M, Cheung C Y, *et al.*** 2004. H5N1 influenza: A protean pandemic threat. **Proc Natl Acad SciUSA,** 101: 8156-8161.
6. **Holmes E C, Ghedin E, Miller N, *et al.*** 2005. Whole-genome analysis of human influenza A virus reveals multiple persistent lineages and Reassortment among

Recent H3N2 Viruses. **Plos Biology,** 3: 1579-1589.

7. **Hatta M, Gao P, Halfmann P, *et al.*** 2001. Molecular basis for high virulence of Hong Kong H5N1 influenza A viruses. **Science,** 293: 1840-1842.
8. http://www.flu.org.cn/upfile/attachment/200693093153272.pdf
9. **Huang K, Fan X H.** 2005. **Molecular Epidemiological Studies on H5N1 Influenza Viruses from Poultry in Nanning** (Mr. thesis).: *Guangxi Medical University*, Guangxi, China. (in Chinese)
10. **Iwatsuki-Horimoto K, Kanazawa R, Sugii S, *et al.*** 2004. The index influenza A virus subtype h5n1 isolated from a human in 1997 differs in its receptor-binding properties from a virulent avian influenza virus. **J Gen Virol,** 85: 1001-1005.
11. **Kou Z, Lei F M, Yu J, *et al.*** 2005. New genotype of avian influenza H5N1 viruses isolated from tree sparrows in China. **J Virol,** 79: 15460-15466.
12. **Kumar S, Tamura K, Nei M.** 2004. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. **Brief Bioinform,** 5: 150-163.
13. **Li K S, Guan Y, Wang J, *et al.*** 2004. Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia. **Nature,** 430: 209-213.
14. **Lutskiy M I, Rosen F S, Remold-O'Donnell E.** 2005. Genotype-Proteotype Linkage in the Wiskott-Aldrich Syndrome. **J Immunol,** 175: 1329-1336.
15. **Matrosovich M N, Krauss S, Webster R G.** 2001. H9N2 influenza A viruses from poultry in Asia have human virus-like receptor specificity. **Virology,** 281: 156-162.
16. **Matrosovich M, Zhou N N, Kawaoka Y, *et al.*** 1999. The surface glycoproteins of H5 influenza viruses isolated from humans, chickens, and wild aquatic birds have distinguishable properties. **J Virol,** 73: 1146-1155.
17. **Obenauer J C, Denson J, Mehta P K, *et al.*** 2006. Large-scale sequence analysis of avian influenza isolates. **Science,** 311 (5767): 1576-1580.
18. **Rodriguez C, Quero C, Dominguez A, *et al.*** 2006. Proteotyping of human haptoglobin by MALDI-TOF profiling: Phenotype distribution in a population of toxic oil syndrome patients. **Proteomics,** 6(Suppl 1): S272-S281.
19. **Ronquist F, Huelsenbeck J P.** 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. **Bioinformatics,** 19: 1572-1574.
20. **Roth M J, Forbes A J, Boyne II M T, *et al.*** 2005. Precise and Parallel Characterization of Coding Polymorphisms, Alternative Splicing and Modifications in Human Proteins by Mass Spectrometry. **Mol Cell Proteomics,** 4 (7): 1002-1008.
21. **Shillingford J M, Miyoshi K, Robinson G W, *et al.*** 2003. Proteotyping of Mammary Tissue from Transgenic and Gene Knockout Mice with Immunohistochemical Markers: a Tool To Define Developmental Lesions. **J Histochem Cytochem,** 51 (5): 555-565.
22. **Simmons M P, Ochoterena H.** 2000. Gaps as Characters in Sequence-Based Phylogenetic Analyses. **Syst Biol,** 49 (2): 369-381.
23. **Thompson J D, Gibson T J, Plewniak F, *et al.*** 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. **Nucl Acids Res,** 25: 4876-4882.
24. **Wang J, Li K S.** 2004. **Genotype Evolution of the H5N1 Influenza Viruses in Aquatic Birds in Southern China** (Mr. thesis). Shantou University, Guangdong, China. (in Chinese)
25. **Webster R G, Bean W J, Gorman O T, *et al.*** 1992. Evolution and ecology of influenza A viruses. **Microbiol Rev,** 56: 152-179.
26. **Zhuang Z P, Huang S, Kowalak J A, *et al.*** 2006. From tissue phenotype to proteotype: Sensitive protein identification in microdissected tumor tissue. **Int J Oncol,** 28 (1): 103-110.