AVIAN DISEASES 51:434–439, 2007

Research Note—

Computational Identification of Reassortments in Avian Influenza Viruses

Xiu-Feng Wan,[AB] Xiaomeng Wu,[ABC] Guohui Lin,[C] Samuel B. Holton,[A] Racheal A. Desmone,[A] Chi-Ren Shyu,[D] Yi Guan,[E] and Michael E. Emch[F]

[A]Systems Biology Laboratory, Department of Microbiology, Miami University, Oxford, OH 45056
[C]Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada T6G 2E8
[D]Department of Computer Science, University of Missouri, Columbia, MO 65211
[E]Department of Microbiology, Hong Kong University, Hong Kong SAR, China
[F]Department of Geography and Carolina Population Center, University of North Carolina, Chapel Hill, NC 27599

Received 27 April 2006; Accepted 2 October 2006

SUMMARY. The avian influenza virus (AIV) has eight genomic segments (hemagglutinin [HA], neuraminidase [NA], RNA polymerase subunit A [PA], RNA polymerase subunit B1 [PB1], RNA polymerase subunit B2 [PB2], nucleoprotein [NP], nonstructural gene [NS], and matrix protein [M]). The genetic reassortments, recombinations, and mutations lead to a rapid emergence of novel genotypes of the AIVs during their evolution. These emerging viruses provide a large reservoir for pandemic strains. Here we describe a novel computational strategy for genetic reassortment identification. In contrast to the traditional phylogenetic approaches, our method views the genotypes through the modules in networks. Genetic segments with short phylogenetic distance are grouped into modules. Our method is not limited to the number of sequences. We applied this method in reassortment identification of NP segments in H_5N_1 AIVs. We identified two new potential reassortments for H_5N_1 AIVs beyond the reported genotypes in literature.

RESUMEN. *Nota de Investigación*—Identificación computacional de recombinaciones en virus de influenza aviar.

El virus de influenza aviar tiene genoma con ocho segmentos: hemaglutinina [HA], neuraminidasa [NA], subunidad A de la polimerasa ARN [PA], subunidad B1 de la polimerasa ARN [B1], subunidad B2 de la polimerasa ARN [B2], nucleoproteína, gen no estructural [NS] y gen M. Las recombinaciones y mutaciones del genoma durante su proceso evolutivo permitieron la aparición rápida de nuevos genotipos del virus de influenza aviar. Estos virus nuevos proporcionaron un amplio reservorio para las cepas observadas en las pandemias. En el presente artículo se describe una novedosa estrategia computacional para la identificación de recombinaciones genéticas. En contraste con el enfoque filogenético tradicional, este método enfoca los genotipos a través de módulos en redes. Los segmentos genéticos con distancias genéticas cortas entre sí, se agrupan en módulos. El presente método no se limita al número de secuencias. El método se aplicó en la identificación de recombinaciones en los segmentos del gen de la nucleoproteína provenientes de virus H5N1 de influenza aviar. Se logró la identificación de dos recombinaciones potenciales para los virus H5N1 de influenza aviar, mas allá de los genotipos reportados en la literatura.

Key words: avian influenza virus, genotyping, H_5N_1, phylogenetic analysis, composition vector, genetic reassortment, influenza reassortment

Abbreviations: AIV = avian influenza virus; Ck = Chicken; Dk = Duck; Gd = Guangdong; Gs = Goose; HA = hemagglutinin; HK = Hong Kong; HN = Hunan; IND = Indonesia; JK = Japan–Korea; M = matrix protein; MB = migration bird; NA = neuraminidase; NP = nucleoprotein; NS = nonstructural gene; PA = RNA polymerase subunit A; PB1 = RNA polymerase subunit B1; PB2 = RNA polymerase subunit B2; VN = Vietnam; VT = Vietnam–Thailand; YN = Yunnan

Avian influenza virus (AIV) is a negative-stranded RNA virus with eight genomic segments: RNA polymerase subunit B2 (PB2), RNA polymerase subunit B1 (PB1), RNA polymerase subunit A (PA), hemagglutinin (HA), nucleoprotein (NP), neuraminidase (NA), matrix protein (M), and nonstructural gene (NS). A total of 16 HA and 9 NA subtypes have been reported (5,12). All of these subtypes came from avian species (8). Genetic reassortment refers to the exchange of the genetic segments between two co-infected viruses, and it is one of the main causes for the rapid emergence of novel influenza viral genotypes, especially for pandemic influenza strains. For instance, the 1957 and 1968 influenza pandemic strains were reported to be reassortments from a previously circulating H_1N_1 strain that existed before 1957 (18). Both HA and NA genes in the 1957 H_2N_2 strain were from avian strains. The 1968 H_3N_2 strain was generated from the 1957 H_2N_2 strain by replacing H_2 with another avian H_3. However, PA, PB2, NP, M, and NS were all from previous H_1N_1 viruses. Genetic reassortment can occur frequently not only among influenza A viruses but also among influenza B viruses and among influenza C viruses (20).

It has been reported that H_5N_1 avian influenza virus could potentially be the next pandemic strain. To date, these viruses have spread to more than 20 countries and areas in three continents, including Asia, Europe, and Africa. In addition to causing a huge economic loss in the poultry industry, H_5N_1 AIVs have caused 105 deaths out of 184 cases, since 2003 in eight countries, including Azerbaijan, Cambodia, China, Indonesia, Iraq, Thailand, Turkey, and Vietnam (http://www.who.int). After we isolated the first two strains of H_5N_1 AIVs (A/Goose/Guangdong/1/1996 and A/Goose/Guangdong/2/1996) (15), these viruses have undergone rapid evolution,

[B]These two authors contributed equally to this paper.

and multiple genotypes have been reported to co-exist even in single epidemics (6). The reported genotypes include at least Goose/Guangdong (Gs/Gd), A, B, C, D, E, X0, Z, Z+, Y, W, X0-X2, and V (7). It is quite interesting that the dominant genotype causing the 2003–04 avian influenza outbreaks in Asia is the Z genotype, although some other genotypes may have coexisted (9,17). The genotype causing human infection in Vietnam and Thailand also belongs to the Z genotype, although the original H_5N_1 AIV causing human cases in Hong Kong in 1997 may belong to the original Gs/Gd genotype (9). However, a systematic analysis of avian influenza genotypes has not been done because of the ineffectiveness of currently available methods.

Phylogenetic tree construction is the general approach used to define the genotypes for AIVs. Briefly, the target sequences are aligned using multiple sequence alignments, such as Clustal W (14). The aligned sequences will be input for tree construction using maximum parsimony or maximum likelihood through PAUP* (13) or Phylip (4). The increasing number of influenza sequences in the database poses a great challenge for influenza genotyping. Since both multiple sequence alignments and tree construction are heuristic approaches, the results will be less reliable as the sequence number increases. To obtain reliable results, a small group of sequences must be selected as prototypes for the phylogenetic analyses. As a result, it is very difficult to reflect the genotype information systematically. Thus, a robust genotyping system is needed for influenza genotype analysis.

Here we present a novel computational genotyping method for AIVs. Our method is based on the distance measurement using the Complete Composition Vector (CCV) (19); this averts heuristic steps of multiple sequence alignments and allows the maximum parsimony or the maximum likelihood phylogenetic tree construction to be avoided. We further treat each gene segment as a node in the module, and each genotype is visualized as a module in the network. The separation of a node is determined by the distance distribution. After validation with known H_5N_1 genotyping analyses, we analyze the genotypes of H_5N_1 AIVs. Several new genotypes have been identified through our studies.

MATERIALS AND METHODS

Genotyping algorithm. The composition vector (11,19) was used to contain the evolutionary information in the sequences. Briefly, we use S to denote one of the eight gene segments, $S \in$ {HA, NA, NP, PB2, PB1, PA, NS, M}. By using window size k ($k \geq 2$), we can scan S to generate $L - k + 1$ strings with the length k; each string can be represented as $\alpha_1\alpha_2 \ldots \alpha_k$, $\alpha \in$ {A, T/U, G, C} for nucleotide, and α will have 20 amino acids for protein. We denote the number of occurrences in S as $f(\alpha_1\alpha_2 \ldots \alpha_k)$. Thus, the probability of string $\alpha_1\alpha_2 \ldots \alpha_k$ is $p(\alpha_1\alpha_2 \ldots \alpha_k) = f(\alpha_1\alpha_2 \ldots \alpha_k)/(L-k+1)$. Similarly, we can calculate the probability of strings $\alpha_1\alpha_2 \ldots \alpha_{k-1}$ and $\alpha_2\alpha_3 \ldots \alpha_k$. Thus, we can calculate the expected probability for string $\alpha_1\alpha_2 \ldots \alpha_k$ through a Markov model:

$$p^e(\alpha_1\alpha_2 \ldots \alpha_k) = \begin{cases} \dfrac{p(\alpha_1\alpha_2 \ \ldots \ \alpha_{k-1}) \times p(\alpha_2\alpha_3 \ \ldots \ \alpha_{k-2})}{p(\alpha_2\alpha_3 \ \ldots \ \alpha_{k-1})}, \\ \\ 0 \end{cases}$$

$$\text{if } p(\alpha_2\alpha_3 \ldots \alpha_{k-1}) \neq 0 \qquad (1)$$

If $k = 1$, we can simplify the above equation as $p^e(\alpha_1) = p(\alpha_1)$. We can calculate the difference $d(\alpha_1\alpha_2 \ldots \alpha_k)$ between the real probability and expected probability as the evolutionary information carried by string $\alpha_1\alpha_2 \ldots \alpha_k$:

$$d(\alpha_1\alpha_2 \ldots \alpha_k) = \begin{cases} \dfrac{p(\alpha_1\alpha_2 \ \ldots \ \alpha_k) - p^e(\alpha_1\alpha_2 \ \ldots \ \alpha_k)}{p^e(\alpha_1\alpha_2 \ \ldots \ \alpha_k)}, & \text{if } p^e(\alpha_1\alpha_2 \ldots \alpha_k) \neq 0 \\ \\ 0 \end{cases} \qquad (2)$$

We denote the composition vector containing string $\alpha_1\alpha_2 \ldots \alpha_k$ as $V^k(S)$, in which k is the length of the string. For a protein sequence that has 20 amino acids, the number of entries in V^k will be 20^k. For a nucleotide sequence, the number of entries in V^k will be 4^k. To contain the information as rich as possible, we can include all of the strings from length m to n, that is, $m \leq k \leq n$. Thus, the CCV can be defined as $V(S) = [V^m(S), V^{m+1}(S), \ldots, V^k(S), \ldots, V^{n-1}(S), V^n(S)]$. We will determine the optimal values of m and n by checking the amount of information contained. In this study, we set $m = 1$ and $n = 3$. The distance between two segments is computed using Euclidean distance (16):

$$D(S, S') = \sqrt{\sum_{i-1}^{N} (d_i - d_i')^2}, \qquad (3)$$

where d and d' are the evolutionary information for string $\alpha_1\alpha_2 \ldots \alpha_k$ in segment S and S' in two AIVs, and $N = 20^n$ for amino acids or 4^n for nucleotides, where n is the maximum length of the string. For instance, S and S' can be HA segments in two AIVs, respectively.

To determine the relationship between the genetic segments, we chose a filter value to visualize the distance between AIVs. By using this filter, we will be able to remove the edges between nodes having remote distances in the network. Thus, the AIVs will result in separate modules, and each separated module will denote a potential genotype. A network visualization package, BioLayout (3), was applied to visualize the resulting genotypes.

Multiple sequence alignments and the maximum parsimony phylogenetic tree construction. To validate our genotype method, the protein sequence alignments were performed using Clustal W (14), and the phylogenetic analyses were based on PAUP* with the tree bisection reconnection branch-swapping option for the heuristic search of the maximum parsimony (13).

Data sets. The influenza data sets were downloaded from the Influenza Virus Resource database at GenBank (http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html), which was updated in December 2005. In this study, we only analyzed the HA protein with at least 500 amino acids and the NP protein with at least 450 amino acids (about 90% sequence coverage of the whole protein). As a result, we utilized 297 HA genes and 1271 NP genes in our analyses. The computational analyses covered all sequences for both HA and NP genes.

RESULTS AND DISCUSSION

Validation of the distance measurement. We calculated the distance between HA gene segments based on CCV. To validate our distance measurement, we constructed the phylogeny tree using the Neighborhood-Joining method of Phylip based on the distance CCV measured. From the resulting phylogenetic tree (Fig. 1), we can surmise that the H_5N_1 AIVs fall into 11 lineages: Gs/Gd/96-like, Hong Kong (HK)/97, ChinaMix, Hunan (HN), migration bird (MB), Japan–Korea (JK), Indonesia (IND), Yunnan (YN), ChinaMix2, HK02/03, and Vietnam–Thailand (VT). A complete phylogenetic tree can be seen at http://www.sysbio.muohio.edu/Flu/AD2006/SFig1.pdf. Most of these lineages have been reported previously (1,2,6,7). The Gs/Gd/96-like lineage includes those AIV decedents of the first goose isolate from Guangdong province. ChinaMix and ChinaMix2 are distinct groups of AIVs, which include those isolates from China in 2001–04. Both ChinaMix and

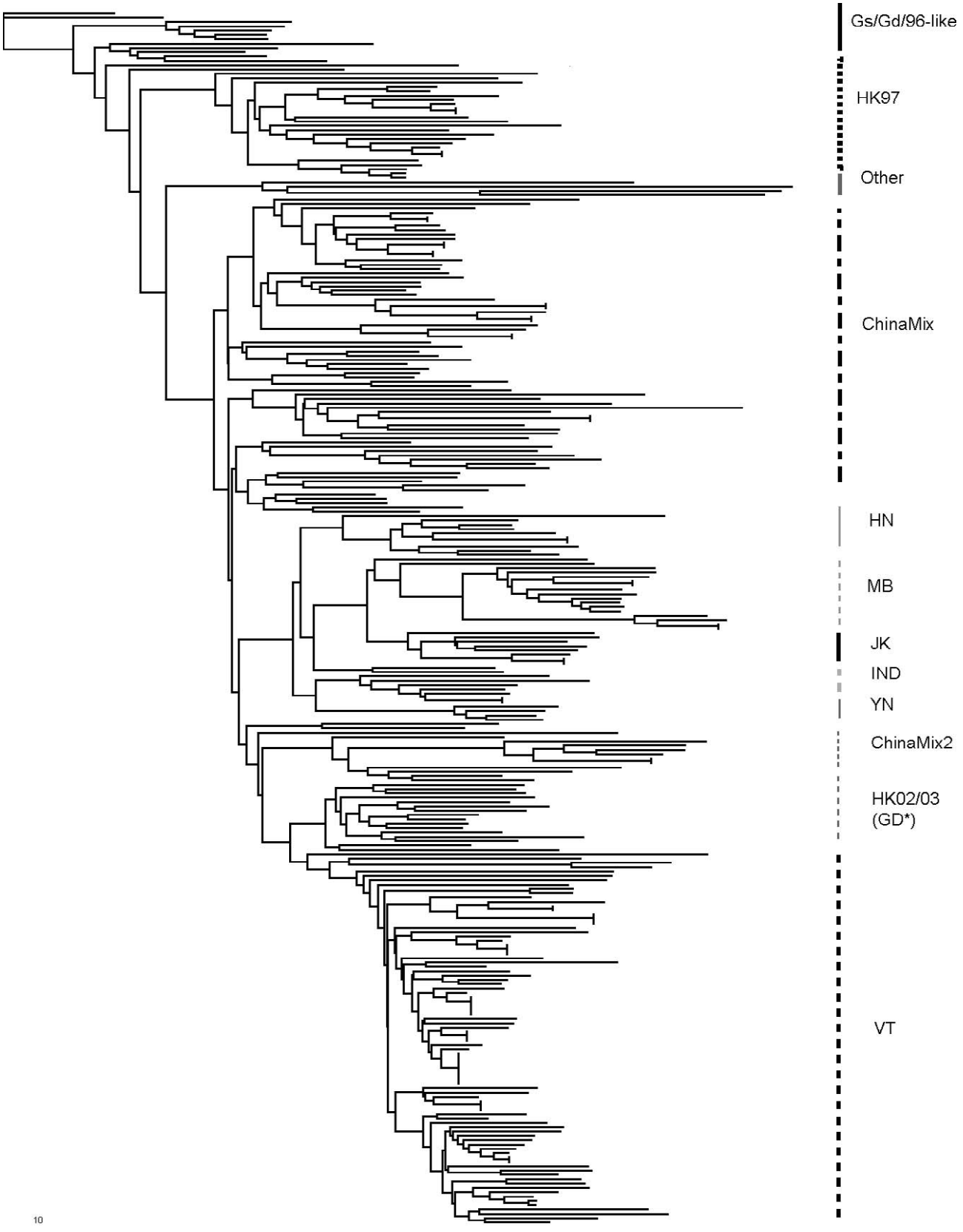

Fig. 1. The phylogenetic tree for H_5N_1 AIVs based on the distance measurement from CCV. Eleven lineages are observed including Gs/Gd/96-like, HK/97, ChinaMix, HN, MB, JK, IND, YN, ChinaMix2, HK02/03 (GD*), and VT. A complete phylogenetic tree is available at http://www.sysbio.muohio.edu/Flu/AD2006/SFig1.pdf.

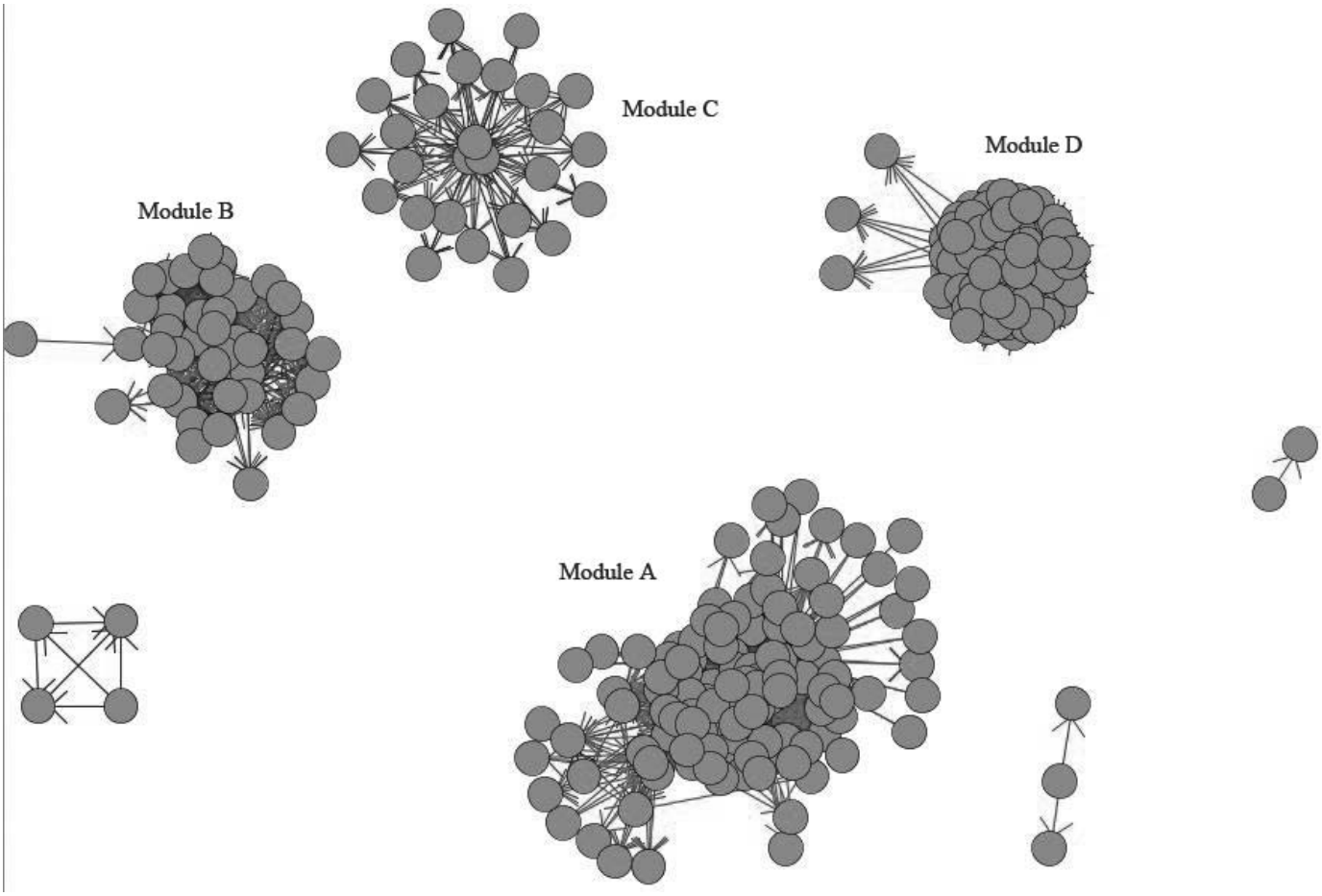

Fig. 2. Module representation of the NP genes. The H_5N_1 AIVs can be grouped into four major modules (A–D). Each module includes nodes (the NP genes in AIVs) with relatively smaller genetic distance.

ChinaMix2 can be separated into several more specific sublineages. HN and YN are two reported alleles of local epidemics of AIV. JK lineage includes the H_5N_1 isolates from Eastern Asia, including Japan and Korea. IND is the lineage for viruses isolated from Indonesia and VT (from Vietnam and Thailand). The lineage of HK02/03 (GD*) are AIVs isolated from Hong Kong during 2002–03. The group denoted as ''others'' contains the H_5N_1 HA control genes, including A/Tukey/England/50-92/91 (H_5N_1). The phylogenetic analyses have supported the complicated evolutionary paths of H_5N_1 AIVs, which indicate the necessity of further viral surveillance.

These results demonstrate that our distance measurement is effective for measuring the distances between AIVs. Our method, however, only requires several minutes to obtain all the pairwise distances for more than 1000 HA sequences, and the distance measurement will not be affected by the number of input gene segments. As shown in our earlier research, CCV is even effective in distance measurement of bacteria using complete genomes (19). It would be impossible for any currently available multiple sequence alignments to generate a robust distance for this type of application. Our method may have a limitation for sequence length since information in a short sequence might not be able to contain enough genetic information for distance measurement. Nevertheless, this limitation may be present with any available method. These advances in genomic technology can result in rapidly produced complete genomes of AIVs.

Module representation of gene segments. The avian influenza gene segments can be input as nodes in a network. To group nodes within a short distance to the HA gene as a module/cluster, we removed the edges above a specified threshold. Fig. 2 shows the module representation of the NP genes of AIVs using the filter threshold value of 50. Specifically, all H_5N_1 AIVs were kept in the flu network, and the edges with a distance larger than 50 were removed from the network. After filtering, 298 NP genes from the 1271 were present in this final network. In the NP genes, four main modules (A–D) have been identified. The phylogenetic trees for individual modules are available at http://www.sysbio.muohio.edu/Flu/AD2006/SFig2-5.pdf. The separation of the NP genes can identify the potential reassortments between different virus strains. Further analyses are required to determine the optimal value for each gene segment based on the distance distribution of the AIV genes, which may result in a better clustering of each module. In addition, the modules might be refined by applying local filter values instead of a global filter value. In Module D, for example, the sublineages from the phylogenetic trees are not shown. Traditional clustering algorithms are used to resolve the module/cluster identification problem. However, we may meet the same NP challenge as the maximum parsimony phylogenetic tree construction.

Reassortments in the NP genes of H_5N_1 AIVs. Fig. 2 shows four major modules for H_5N_1 AIVs. Each of these modules shows mixed serotypes, which indicate potential reassortments

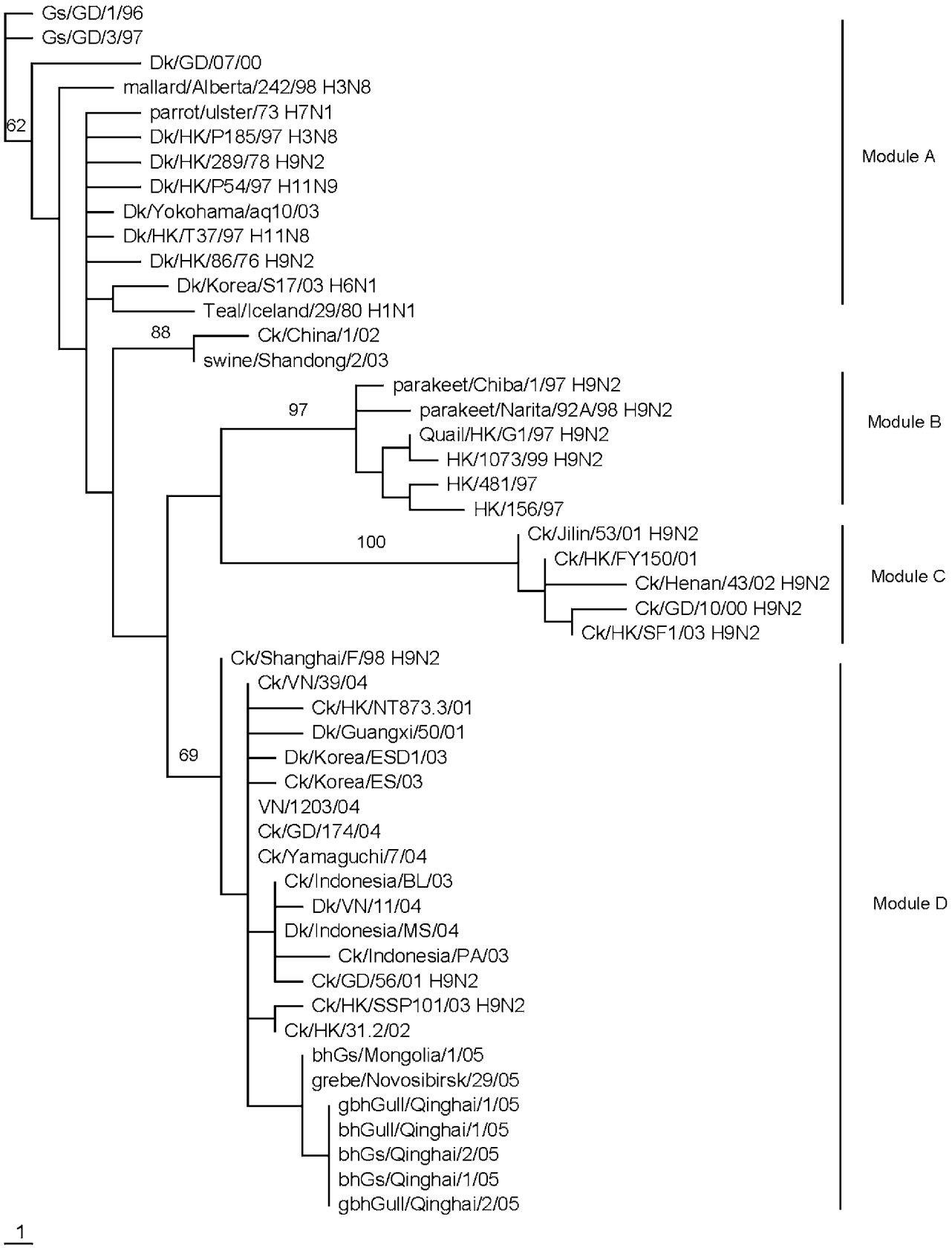

Fig. 3. Reassortments between the H_5N_1 AIVs and other serotypes of AIVs. The genes were selected from the Modules A–D from Fig. 2. The serotypes other than H_5N_1 were shown. The phylogenetic analyses were based on PAUP* with the tree bisection reconnection branch-swapping option for the heuristic search of the maximum parsimony. The critical bootstrapping values of more than 50% from 100 replications were shown. The protein length of the NP genes varied from 450 to 498 amino acids.

between values. To validate these reassortments, we selected a small portion of the genes from each module and constructed the phylogenetic trees using multiple sequence alignments and the maximum parsimony using PAUP* (Fig. 3). Module A showed a mixture between H_5N_1, H_3N_8, H_7N_1, H_9N_2, $H_{11}N_9$, $H_{11}N_8$, H_6N_1, and H_1N_1 (Fig. 2). There might be a potential reassortment for NPs between Duck (Dk)/Yokohama/aq10/03 (H5N1) and Dk/HK/T37/97 (H11N8), which has not yet been reported. Dk/Yokohama/aq10/03 was a virus isolated from ducks but showed a mild pathogenesis to another strain Chicken (Ck)/Yamaguchi/7/04 isolated during the 2003–04 flu outbreaks in Japan (10), which were located in Module D. Module D showed the reassortments between H_9N_2 and H_5N_1 AIVs. For instance, Ck/Gd/56/01 (H_9N_2) may be the potential NP donor for Indonesia sublineages, which has not yet been reported. The NP gene in Module D can be associated with the H_5N_1 AIV genotype Z, which was the main cause for the 2003–04 influenza pandemic (9). The NPs in Module A are phylogenetically closer to each other than other viruses. Module B demonstrated the potential reassortments between H_9N_2 and H_5N_1 1997 isolates in Hong Kong, which have been reported earlier as G1-like H_9N_2 lineage in genotype H_5N_1/97 by Guan *et al.* (6). Module C identified the reassortments between the 2001 isolates and H_9N_2 strains in China, which has been reported as H_9N_2 Y280-like lineage in H_5N_1 genotype D (6).

In summary, here we developed a new method to identify the reassortments from AIVs by integrating the distance measurements using CCV and the concept of a network module. We applied this method to identify the reassortments of the NP genes in AIVs, and we were able to identify the reported reassortments between H_5N_1 and H_9N_2 for NP genes based on this approach. Two new reassortments were identified using this method. Further research will be pursued to refine the module representation for AIV sublineages by studying the distance distribution between influenza viruses.

REFERENCES

1. Chen, H., G. J. Smith, K. S. Li, J. Wang, X. H. Fan, J. M. Rayner, D. Vijaykrishna, J. X. Zhang, L. J. Zhang, C. T. Guo, C. L. Cheung, K. M. Xu, L. Duan, K. Huang, K. Qin, Y. H. Leung, W. L. Wu, H. R. Lu, Y. Chen, N. S. Xia, T. S. Naipospos, K. Y. Yuen, S. S. Hassan, S. Bahri, T. D. Nguyen, R. G. Webster, J. S. Peiris, and Y. Guan. Establishment of multiple sublineages of H5N1 influenza virus in Asia: implications for pandemic control. Proc. Natl. Acad. Sci. U. S. A. 103:2845–2850. 2006.

2. Chen, H., G. J. Smith, S. Y. Zhang, K. Qin, J. Wang, K. S. Li, R. G. Webster, J. S. Peiris, and Y. Guan. Avian flu: H5N1 virus outbreak in migratory waterfowl. Nature 436:191–192. 2005.

3. Enright, A. J., and C. A. Ouzounis. BioLayout—an automatic graph layout algorithm for similarity visualization. Bioinformatics 17:853–854. 2001.

4. Felsenstein, J. PHYLIP—Phylogeny Inference Package (Version 3.2). Cladistics 5:164–166. 1989.

5. Fouchier, R. A., V. Munster, A. Wallensten, T. M. Bestebroer, S. Herfst, D. Smith, G. F. Rimmelzwaan, B. Olsen, and A. D. Osterhaus. Characterization of a novel influenza A virus hemagglutinin subtype (H16) obtained from black-headed gulls. J. Virol. 79:2814–2822. 2005.

6. Guan, Y., J. S. Peiris, A. S. Lipatov, T. M. Ellis, K. C. Dyrting, S. Krauss, L. J. Zhang, R. G. Webster, and K. F. Shortridge. Emergence of multiple genotypes of H5N1 avian influenza viruses in Hong Kong SAR. Proc. Natl. Acad. Sci. U. S. A. 99:8950–8955. 2002.

7. Guan, Y., L. L. Poon, C. Y. Cheung, T. M. Ellis, W. Lim, A. S. Lipatov, K. H. Chan, K. M. Sturm-Ramirez, C. L. Cheung, Y. H. Leung, K. Y. Yuen, R. G. Webster, and J. S. Peiris. H5N1 influenza: a protean pandemic threat. Proc. Natl. Acad. Sci. U. S. A. 101:8156–8161. 2004.

8. Kilbourne, E. D. Perspectives on pandemics: a research agenda. J. Infect. Dis. 176(Suppl. 1):S29–S31. 1997.

9. Li, K. S., Y. Guan, J. Wang, G. J. Smith, K. M. Xu, L. Duan, A. P. Rahardjo, P. Puthavathana, C. Buranathai, T. D. Nguyen, A. T. Estoepangestie, A. Chaisingh, P. Auewarakul, H. T. Long, N. T. Hanh, R. J. Webby, L. L. Poon, H. Chen, K. F. Shortridge, K. Y. Yuen, R. G. Webster, and J. S. Peiris. Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia. Nature 430:209–213. 2004.

10. Mase, M., T. Imada, K. Nakamura, N. Tanimura, K. Imai, K. Tsukamoto, and S. Yamaguchi. Experimental assessment of the pathogenicity of H5N1 influenza A viruses isolated in Japan. Avian Dis. 49:582–584. 2005.

11. Qi, J., B. Wang, and B. I. Hao. Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. J. Mol. Evol. 58:1–11. 2004.

12. Rohm, C., N. Zhou, J. Suss, J. Mackenzie, and R. G. Webster. Characterization of a novel influenza hemagglutinin, H15: criteria for determination of influenza A subtypes. Virology 217:508–516. 1996.

13. Swofford, D. L. PAUP*: Phylogenic Analysis Using Parsimony. Sinauer, Sunderland, MA. 1998.

14. Thompson, J. D., D. G. Higgins, and T. J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680. 1994.

15. Wan, X.-F. Isolation and characterization of avian influenza viruses in China. M.Sc. Thesis. College of Veterinary Medicine, South China Agricultural University, Guangzhou, Guandong Province, China. 1998.

16. Wan, X. F., S. M. Bridges, and J. A. Boyle. Revealing gene transcription and translation initiation patterns in archaea, using an interactive clustering model. Extremophiles 8:291–299. 2004.

17. Wan, X. F., T. Ren, K. J. Luo, M. Liao, G. H. Zhang, J. D. Chen, W. S. Cao, Y. Li, N. Y. Jin, D. Xu, and C. A. Xin. Genetic characterization of H5N1 avian influenza viruses isolated in southern China during the 2003–04 avian influenza outbreaks. Arch. Virol. 150:1257–1266. 2005.

18. Webster, R. G., W. J. Bean, O. T. Gorman, T. M. Chambers, and Y. Kawaoka. Evolution and ecology of influenza A viruses. Microbiol. Rev. 56:152–179. 1992.

19. Wu, X., X.-F. Wan, G. Wu, D. Xu, and G. Lin. Phylogenetic analysis using complete signature information of whole genomes and clustered Neighbour-Joining method. Int. J. Bioinformatics Res. Appl. 2:219–248. 2006.

20. Xu, X., S. E. Lindstrom, M. W. Shaw, C. B. Smith, H. E. Hall, B. A. Mungall, K. Subbarao, N. J. Cox, and A. Klimov. Reassortment and evolution of current human influenza A and B viruses. Virus Res. 103:55–60. 2004.

ACKNOWLEDGMENTS

In this study, X.-F. Wan, X. Wu, S. B. Holton, and R. A. Desmone were supported by Miami University CFR and a Grant to Promote Research. X. Wu and G. Lin were partially supported by NSERC.