2 Smith GJD, Vijaykrishna D, Bahl J *et al.* Origin and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature 2009; 459:1122–1125.

3 Garten RJ, Davis T, Russell CA. Antigenic and genetic characteristics of swine-origin 2009 A (H1N1) influenza viruses circulating in humans. Science 2009; 325:197–201.

4 Burr T, Gattiker JR, LaBerge GS. Genetic subtyping using cluster analysis. SIGKDD Explor, 2001; 3:33–42.

5 Liu S, Ji K, Chen J, Tai D *et al.* Panorama phylogenetic diversity and distribution of type A influenza virus. PLoS ONE 2009; 4:1–20.

6 WHO/OIE/FAO H5N1 Evolution Working Group. Continuing progress towards a unified nomenclature for the highly pathogenic H5N1 avian influenza viruses: divergence of clade 2.2 viruses. Influenza Other Respi Viruses 2009; 3:59–62.

7 Bao Y, Bolotov P, Dernovoy D *et al.* The influenza virus resource at the National Center for Biotechnology Information. J Virol 2008; 82:596–601.

8 Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004; 32:1792–1797.

9 The R Foundation. Available at http://www.r-project.org/.

10 Katoh K, Asimenos G, Toh H. Multiple alignment of DNA sequences with MAFFT. Methods Mol Biol 2009; 537:39–64.

11 Stamatakis A, Ludwig T, Meier H. RAxML-III: A fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 2005; 21:456–463.

12 Lu G, Rowley T, Garten R, Doris R. FluGenome: a web tool for genotyping influenza A virus. Nucleic Acids Res 2007:W275–279.

13 Xu J, Lu G. Evolution of influenza viral neuraminidase (NA) genes revealed by large-scale sequence analysis. Proceedings of the Options for the Control of Influenza VII. In Press

14 Wan XF, Chen G, Luo F *et al.* A quantitative genotype algorithm reflecting H5N1 avian influenza niches. Bioinformatics 2007; 23:2368–2375.

IPMiner: a progenitor gene identifier for influenza A virus

Zhipeng Cai,[a] Yueming Duan,[a,b] Yingshu Li,[b] Guohui Lin,[c] Mufit Ozden[d], Xiu-Feng Wan[a]

[a]Department of Basic Sciences, College of Veterinary Medicine, Mississippi State University, Mississippi State, MS, USA. [b]Department of Computer Science, Georgia State University, Atlanta, GA, USA. [c]Department of Computing Science, University of Alberta, Edmonton, AB, Canada. [d]Department of Computer Science and Systems Analysis, Miami University, Oxford, OH, USA.

Keywords CCV-MST, complete composition vector, influenza A virus, minimum spanning tree, phylogeny, progenitor.

Please cite this paper as: Cai *et al.* (2011) IPMiner: a progenitor gene identifier for influenza A virus. Influenza and Other Respiratory Viruses 5 (Suppl. 1), 395–415.

Abstract

Identification of the genetic origin of influenza A viruses will facilitate understanding of the genomic dynamics, evolutionary pathway, and viral fitness of influenza A viruses. The exponential increases of influenza sequences have expanded the coverage of influenza genetic pool, thus potentially reducing the biases for influenza progenitor identification. However, these large amounts of data generate a great challenge in progenitor identification. To increase computational efficiency, IPMiner is developed by integrating complete composition vector for genetic distance calculation and minimum spanning tree algorithm for progenitor identification. IPMiner is available to at http://sysbio.cvm.msstate.edu/IPMiner.

Introduction

Influenza A virus is a negative-stranded RNA virus that belongs to the *Orthomyxoviridae* family. Influenza A virus has eight genomic segments (segment 1–8) with varying lengths from about 890 to 2341 nucleotides. The subtypes of influenza A viruses are named by combining the serotypes of their surface protein hemagglutinin (HA) and neuraminidase (NA). To date, 16 HA (H1 through H16) and 9 NA (N1 through N9) serotypes have been identified. Influenza A virus causes zoonotic diseases in various hosts, such as human, pig, bird, horse, seal, whale, and dog. As a segmented, negative-stranded RNA virus, influenza A virus is characterized by its rapid mutation and frequent reassortment. A reassortment event refers to the exchange of gene segments between co-infected influenza viruses, and it has facilitated the emergence of 1957 H2N2, 1968 H3N2, and the 2009 H1N1 pandemic strains.[1,2] Identification of the genetic origins of influenza A viruses will enhance our understanding the evolution and adaptation mechanisms of influenza viruses.

The phylogenetic analysis is the traditional approach to identify the influenza progenitor. First, the nucleotide sequences are aligned using multiple sequence alignment methods, such as ClustalW,[3] MUSCLE,[4] and T-COFFEE.[5] Second, phylogenetic analysis is performed on these aligned sequences to infer their evolutionary relationship using Neighbor-Joining (NJ),[6] Maximum Parsimony,[7] Maximum

Likelihood, or Bayesian inference.[8] Bootstrap analyses or computation of posterior probability are usually applied to estimate the phylogenetic uncertainty. However, this phylogenetic analysis is time consuming due to intensive computations in multiple sequence alignments and phylogenetic inferences. It is difficult to perform an analysis using this method on a large dataset, for instance, with more than 1000 taxa, as is the common case for influenza studies.

Alternatively, BLAST [9] is applied to identify the prototype genes in the database. BLAST determines a similarity by identifying initial short matches and starting local alignments. Since influenza viral sequences have very high similarities, especially for most conserved regions, BLAST usually generates a large number of outputs, which will not be helpful for progenitor identification. Since BLAST is a local sequence alignment, the results from BLAST may not reflect the global evolutionary information between the sequences. The BLAST scores cannot be used to define the evolutionary relations between viruses, especially in the context of the entire genetic pool.

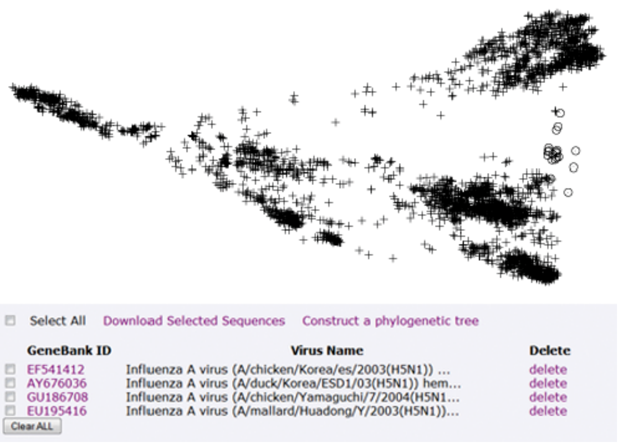
Recently, we have developed a distance measurement method, complete composition vector (CCV), that can calculate genetic distance between influenza A viruses without performing multiple sequence alignments.[10,11] We also adapted the minimum spanning tree (MST) clustering algorithm for influenza reassortment identification.[12] The application of this approach in the analyses of PB2 genes of influenza A virus showed that the integration of CCV and MST allows us to identify the potential progenitor genes rapidly and effectively. Based on these results, here we develop a webserver called IPMiner for influenza progenitor identification. IPMiner can identify potential progenitors for a query sequence against all public influenza datasets within a few minutes.

Precomputed data matrices

In order to improve the computing efficiency, 31 distance matrices were pre-computed by CCV, and they include 16 for HA (H1 to 16), 9 for NA (N1 to N9), and one for each of the internal gene segments (PB2, PB1, PA, NP, NS, and MP). These 31 pre-computed matrices will be updated weekly. IPMiner just needs to compute the query matrices for a query sequence and sequences in the database. The standalone CCV program is also available at http://sysbio.cvm.msstate.edu/IPMiner.

Identification and visualization of influenza progenitor genes

In order to identify the influenza progenitor genes, IPMiner first integrates the query matrix and a corresponding pre-computed matrix into a full distance matrix, which is then clustered by MST clustering algorithm. We adapted the threshold we measured previously in MST, $u + n\sigma$, where u is the average distance and σ is the standard deviation of a cluster.[12] As a result, MST will generate a hierarchical structure for the clusters. In each cluster, we will randomly select 20 viruses or 10% of the cluster size if this cluster has more than 200 viruses. IPMiner will return the viruses with the smallest distances when the search reaches to the lowest level (the largest n) in this hierarchical structure. Our analyses have shown that the level 5 has generally yielded good results for influenza A viruses.



□ Select All Download Selected Sequences Construct a phylogenetic tree

	GeneBank ID	Virus Name	Delete
□	EF541412	Influenza A virus (A/chicken/Korea/es/2003(H5N1)) ...	delete
□	AY676036	Influenza A virus (A/duck/Korea/ESD1/03(H5N1)) hem...	delete
□	GU186708	Influenza A virus (A/chicken/Yamaguchi/7/2004(H5N1...	delete
□	EU195416	Influenza A virus (A/mallard/Huadong/Y/2003(H5N1))...	delete

Figure 1. Visualization of the identified progenitor viruses in the influenza genetic pool. The progenitor viruses are displayed in circles.

To visualize the overall MST structure, IPMiner applies multi-dimensional scaling (MDS) method to project all the viruses in the genetic pool onto a two dimensional graph, and the precursor viruses are marked in different shapes (Figure 1). The users can select other prototype viruses from the graph for further phylogenetic analyses.

A single job with one query sequence takes <2 min. The GenBank identifiers and associated genetic distances and sequence identities are displayed. The users can download the sequences for the identified precursor viruses as well as those from the prototypes viruses. In addition, for the users' convenience, IPMiner generates a phylogenetic tree using NJ method implemented in PHYLIP [13] to illustrate the phylogenetic relationship among the query sequence(s), the identified progenitors, and the selected prototypes viruses.

Implementation and availability

The programs in this solution package are written in Java. The shell scripts are written in korn shell script in order to achieve high performance. Cascading style sheets (CSS) are used for a consistent look across the pages. This also enables to change the overall design just by replacing the CSS definition file. PHP has been used as server side scripting and is written in Java. In order to achieve high performance for computing in a genomic scale, we apply hash function or a binary tree, which enables that the precursor identification has a time complexity of *O*(n). For single queries, the users can visualize the results online. For batch queries of multiple sequences, the results will be sent to the users by e-mail.

IPMiner has been tested on Microsoft Internet Explorer, Mozilla Firefox, and Safari. The users need JavaScript to obtain full function of IPMiner server. The webserver is available at http://sysbio.cvm.msstate.edu/IPMiner.

Conclusions

In summary, IPMiner webserver has three major computational features for influenza progenitor identification: (i) it calculates the genetic distances through CCV and identifies the viruses with the shortest CCV distances against the query virus to be the progenitor genes; (ii) it projects influenza viruses onto a two dimensional map, which illustrates the global relationship between the progenitor genes and other viruses in the genetic pool; and (iii) it performs phylogenetic analyses between the query virus, the identified progenitor genes, and other selected prototype viruses. IPMiner provides a user friendly web service for influenza progenitor identification in real time.

Acknowledgements

CZ is supported partially by Canadian NSERC postdoc fellowship. This project was supported by NSF-EPS-0903787 subaward 012156-007 and NIH NIAID 1RC1AI086830 to XFW.

References

1 Kawaoka Y, Krauss S, Webster RG. Avian-to-human transmission of the PB1 gene of influenza A viruses in the 1957 and 1968 pandemics. J Virol 1989; 63:4603–4608.

2 Shinde V, Bridges CB, Uyeki TM *et al.* Triple-reassortant swine influenza A (H1) in humans in the United States, 2005–2009. N Engl J Med 2009; 360:616–2625.

3 Thompson JD. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighing, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994; 22:4673–4680.

4 Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004; 32:1792–1797.

5 Notredame C, Holm L, Higgins DG. COFFEE: an objective function for multiple sequence alignments. Bioinformatics 1998; 14:407–422.

6 Saitou N, Nei M. The neighbour-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 1987; 4:406–425.

7 Swofford DL. Phylogenic Analysis Using Parsimony. Sinauer Associates: Sunderland, MA; 1998.

8 Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 2001; 17:754–755.

9 Altschul SF, Gish W, Miller W *et al.* Basic local alignment search tool. J Mol Biol 1990; 215:403–410.

10 Wan XF, Chen G, Luo F *et al.* A quantitative genotype algorithm reflecting H5N1 Avian influenza niches. Bioinformatics 2007; 23:2368–2375.

11 Wan XF, Wu X, Lin G *et al.* Computational identification of reassortments in avian influenza viruses. Avian Dis 2007; 51:434–439.

12 Wan XF, Ozden M, Lin G. Ubiquitous reassortments in influenza A viruses. J Bioinformatics Computational Biol 2008; 6:981–999.

13 Felsenstein J. PHYLIP – Phylogeny inference package (version 3.2). Cladistics 1989; 5:164–166.