# Pandemic (H1N1) 2009 Cluster Analysis: A Preliminary Assessment

**Charles Wick,[1#] Samir Deshpande,[2^] Rabih Jabbour,[3] Michael Stanford,[1] Patrick McCubbin,[4] Evan Skowronski,[1] Alan Zulich[1]**

1 U.S. Army, Edgewood Chemical Biological Center, ATTN: RDCB-DRD-D, 5183 Blackhawk Road, Aberdeen Proving Ground, Maryland 21010

2. Science and Technology, 500 Edgewood Road, Edgewood, MD, 21040

3. SAIC, 1308 Continental Drive, Abingdon, MD 21009

4. OptiMetrics, Inc., 100 Walter Blvd, Suite 100, Abingdon, MD 21009

^ Towson University, 8000 York Road, Towson, MD 21252

# to whom correspondence should be addressed, E-mail: charles.h.wick@us.army.mil

## Abstract:

Pandemic (H1N1) 2009 virus has been causing major concerns around the world because of its epidemic potential, rapid dissemination, rate of mutations, and the number of fatalities. One way to gain an advantage over this virus is to use existing rapid bioinformatics tools to examine easily and inexpensively generated genetic sequencing data. We have used the protein sequences deposited with the National Center for Biotechnology Information (NCBI) for data mining to study the relationship among the Pandemic (H1N1) 2009 proteins. There are 11 proteins in the Pandemic (H1N1) 2009 virus, and analysis of sequences from 65 different locations around the globe has resulted in two major clusters. These clusters illustrate the Pandemic H1N1 2009 virus is already experiencing significant genetic drift and that rapid worldwide travel is affecting the distribution of genetically distinct isolates.

Keywords: pandemic H1N1 2009, Cluster Analysis, Bioinformatics

## ***Introduction:***

The H1N1 strain of influenza is a single stranded RNA virus composed of a segmented genome originated from various influenza viruses [1]. An infection of mixtures of various influenza viruses results in the release of progeny viruses containing novel arrangements of segments. In Asia, North America, and much of Europe, viruses of the H1N1 subtype are the most commonly isolated [3-4]. However, for the purposes of this study, we have chosen to focus on the Pandemic (H1N1) 2009 virus isolates [2] that have been of great worldwide public concern this year. The Pandemic (H1N1) 2009 viruses differ in the origins of their genomic components from these previously circulating H1N1 strains and belong to the classic swine lineage, which is genetically related to the human H1N1 viruses responsible for the 1918 Spanish influenza pandemic [5-6]. The reassortment and selection events that led to the recent outbreak of influenza viral infection in man could be traced to several factors including cellular receptors, oligosaccharides, the cohabitation of swine and poultry, and the presence of avian and human receptors in pigs [7-8].

The first reported outbreak of the mutated subtype influenza virus, Pandemic (H1N1) 2009 occurred in Mexico in April 2009 and was rapidly followed by more cases reported in the United States and many other countries. According to the World Health Organization [9], in a two month period, the Pandemic (H1N1) 2009 virus spread to more than 65 countries and infected more than 17,000 people, with 115 deaths.

Several studies have recently been reported that addressed the origin of the Pandemic (H1N1) 2009 using cluster analyses of genomic segments isolated from various infection cases. The overall agreement traces the Pandemic (H1N1) 2009 to a “triple-assortment” - swine influenza predominant in North America and H1N1 strains predominant in swine populations in Europe and Asia[1,10]. While such studies are of significant importance to the public and scientific communities, the correlation studies among the worldwide reported cases of Pandemic (H1N1) 2009 would expand the scope of coverage for this epidemic. This paper addresses the Pandemic (H1N1) 2009 protein sequences collected from National Center for Biotechnology Information (NCBI) [11] as of 06/11/2009. These sequences were specific to 2009 outbreak only, and originated from 65 different global locations [12]. Our results show that the there are significant clustering of Pandemic (H1N1) 2009 among the cases. Also, some cases with reported antiviral resistance (Oseltamivir) showed the same mutation of a single amino acid H274Y in the neuraminidase (NA)-protein of H1N1 that are in agreement with relevant reported data.

## ***Materials and Methods:***

The Pandemic (H1N1) 2009 protein sequences were downloaded from the NCBI website [11] in FASTA format as of 06/11/2009. These protein sequences were specific to the 2009 outbreak, and were collected from different infected people from 65 different locations [12].

The combined FASTA file was fragmented onto location-specific protein sequences. These location-specific FASTA files were then uploaded into a Structured Query Language (SQL) Server relational database management system (RDBMS). This database was used to create a 65 x 65 correlational matrix based on the locations. The corresponding matrix FASTA files were merged to create a two location-based protein sequence file; a total of 4225 FASTA files were created in this matrix. These files were indexed using DBIndexer[3] utility included in its Bio-Works® Suite of applications (Thermo Fisher, CA). The generated header files were used to read the unique number of peptides observed and the computed values were assigned in the corresponding cell of the matrix.

Jaccard's index was computed for each cell to determine the similarity of a given cell to the data set. The unique peptide values observed for each specific location were used to perform clustering calculations, using linkage rule of "Wards Method" and distance measure of "Euclidean distance". This analysis was automated in-house with coded software application labeled as "genTree".

## ***Results and Discussions:***

The resulting cluster analysis generated from the protein sequences of the Pandemic (H1N1) 2009 virus showed two distinct sub-clusters (figure 1). The distinct sub-clusters vary in their number of the locations with cluster-A containing 7 countries, and sub-cluster-B 17 countries (Table 1). Also, sub-cluster-A has fewer numbers of cases (29) than that of sub-cluster-B (36), although this may be the result of availability of sequence data rather than an indication of a true disparity in total numbers of cases. The proteomic variability of the cases within a sub-cluster shows some independence of the geographical location, which may demonstrate dissemination of distinct genetic isolates via commercial airline travel. For example in the topmost portion of sub-cluster-A, a significant grouping was observed among cases from Japan (6 cases), China (3 cases), USA (1 case), and Russia (1 case). This variability pattern is also observed with other cases in either sub-cluster. While it is expected to have same location cases grouped together, the presence of cases from geographically distant locations indicate other factors, such as airline travel, that should considered.

The two sub-clusters were also characterized by variation in the serotype of the Pandemic (H1N1) 2009 strains. Most notable are cases from China and USA which have the largest variability, as cases from these two locations are spread across both sub-clusters, perhaps an indication of individual mobility.  The serotypes of the Pandemic (H1N1) 2009 infecting humans in California are similar to those found in Texas and Mexico (sub-cluster-A), but different from those found in Vermont, DC and Pennsylvania cases (sub-cluster-B). Overall, this clustering analysis potentially provides a different perspective on the Pandemic (H1N1) 2009 strains in terms of geographic distribution of population and migration factors. The establishment of two distinct clusters from the cases studied is an indication of strain variability of Pandemic (H1N1) 2009 to be considered for diagnosis and prognosis purposes. This kind of automated bioinformatics analysis may be useful for the ongoing assessment of how viruses such as influenza spread across the globe, monitoring spread in drug-resistant or particularly virulent genotypes, and potential antigenic/genetic shifts that may impact detection and vaccine efficacy. The addition of more sequence information, easily and cheaply available, correlated with clinical and phenotypic data would be of particular interest to the public health community.

***References:***

1- Solovyov A, Palacios G, Briese T, Lipkin WI, Rabadan R., Cluster analysis of the origins of the new influenza A(H1N1) virus., Euro Surveill. 2009 May 28; 14 (21).

2- Christophe Fraser, *et al.* 2009. Pandemic Potential of a Strain of Influenza A (H1N1): Early Findings *Science* 324, 1557.

3- Campitelli, L., I. Donatelli, E. Foni, M. R. Castrucci, C. Fabiani, Y. Kawaoka, S. Krauss, and R. G. Webster. 1997. Continued evolution of H1N1 and H3N2 influenza viruses in pigs in Italy. Virology 232:310-318.

4- Scholtissek, C., V. S. Hinshaw, and C. W. Olsen. 1998. Influenza in pigs and their role as the intermediate host, p. 137-145. *In* K. G. Nicholson, R. G. Webster, and A. J. Hay (ed.), Textbook of influenza. Blackwell Science, Oxford, United Kingdom.

5- Reid, A. H., T. G. Fanning, J. V. Hultin, and J. K. Taubenberger. 1999. Origin and evolution of the 1918 "Spanish" influenza virus hemagglutinin gene. Proc. Natl. Acad. Sci. USA 96:1651-1656.

6- Schultz, U., W. M. Fitch, S. Ludwig, J. Mandler, and C. Scholtissek. 1991. Evolution of pig influenza viruses. Virology 183:61-73.

7- de Jong, J. C., A. P. van Nieuwstadt, T. G. Kimman, W. L. Loeffen, T. M. Bestebroer, K. Bijlsma, C. Verweij, A. D. Osterhaus, and E. C. Class. 1999. Antigenic drift in swine influenza H3 haemagglutinins with implications for vaccination policy. Vaccine 17:1321-1328.

8- Done, S. H., and I. H. Brown. 1999. Swine influenza viruses in Europe, p. 263-267. . Proceedings of the Allen D. Leman Swine Conference, vol. 26.

9- World Health Organization: http://www.who.int/csr/disease/swineflu/en.

10- Novel Swine-Origin Influenza A (H1N1) Investigation Team (2009) Emergence of a Novel Swine-Origin Influenza A (H1N1) Virus in Humans. N Engl J Med. Vol 361:115-119**.**

11- http://www.ncbi.nlm.nih.gov/genomes/FLU/SwineFlu.html

12- USA, Mexico, Chile, Brazil, Colombia, Poland, Turkey, Greece, Spain, Finland, Norway, Sweden, Italy, Ireland, France, Russia, China, Japan, Philippines, Thailand, Australia, New Zealand, protein sequences were extracted from http://www.ncbi.nlm.nih.gov/genomes/FLU

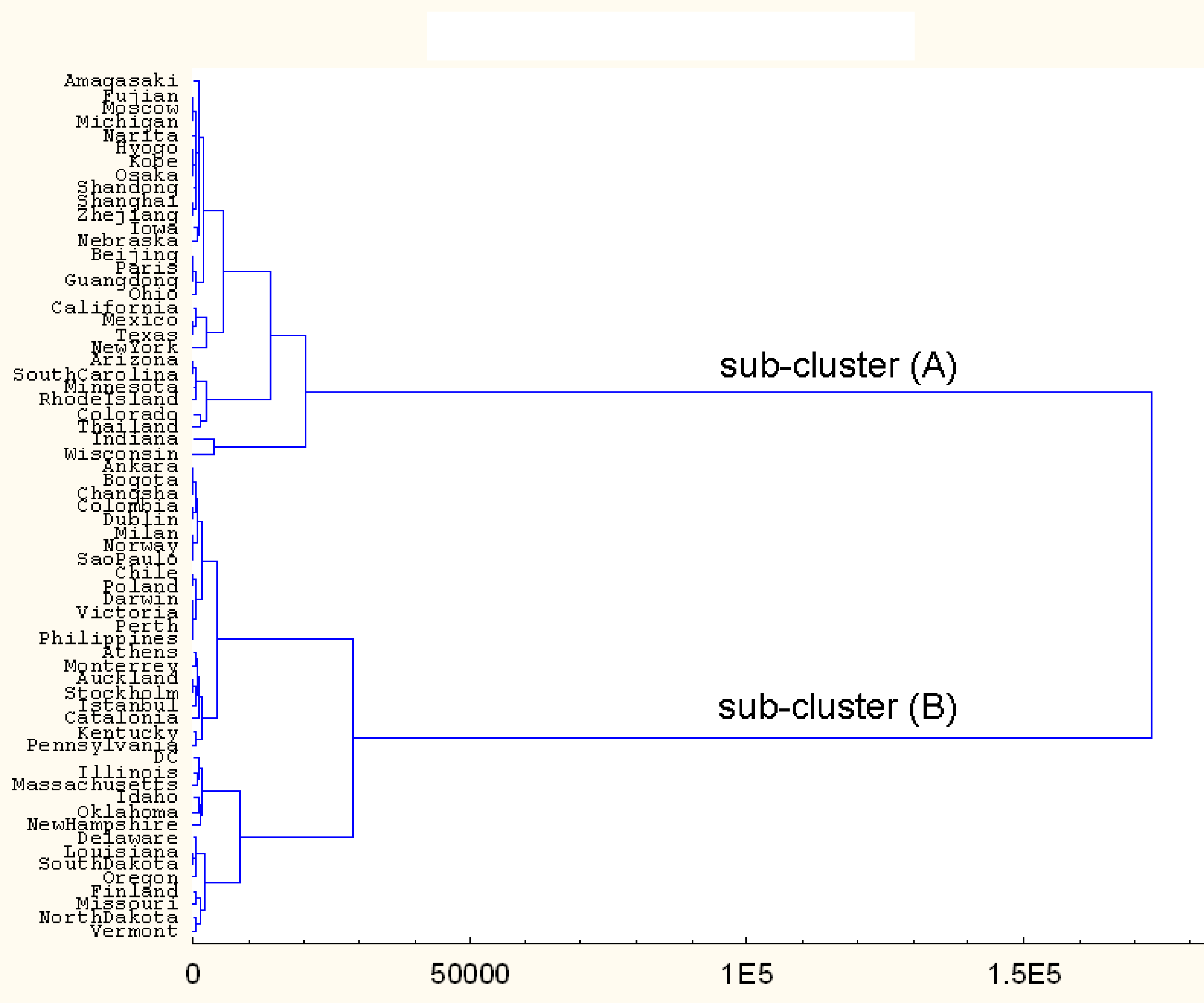

Figure 1: Cluster Analysis of Pandemic (H1N1) 2009 reported cases from different locations. The horizontal length represents the similarity among the protein sequences, and the vertical lines indicate spacing.

Table 1. Country Listing by Sub-cluster

| **Sub-cluster A** | **Sub-Cluster B** | | |
|---|---|---|---|
| China | Australia | Italy | Sweden |
| France | Chile | Mexico | Turkey |
| Japan | China | New Zealand | United States |
| Mexico | Columbia | Norway | |
| Russia | Finland | Philippines | |
| Thailand | Greece | Poland | |
| United States | Ireland | Spain | |