Evolutionary Trajectory Analysis of Pandemic H1N1 2009

R. Burke Squires1, Elizabeth McClellan3, Jyothi Noronha1, Victoria Hunt1, Richard H. Scheuermann1,2
1Department of Pathology, 2Division of Biomedical Informatics, University of Texas Southwestern Medical Center at Dallas, Dallas, Texas 2 Department of Statistics, Southern Methodist University, Dallas, Texas.

Introduction

In late April and early May 2009 the first pandemic strain of the 21st Century arrived in the form of the Pandemic H1N1 2009 influenza virus. Initially named “Swine Flu,” the origin of the Pandemic H1N1 2009 sequence presented many initial questions about each of the segments of the virus.

We began by examining the relationship between nucleotide differences of the A/California/04/2009 and a number of top BLAST hits versus the isolation year differences.

Although we expected that a correlation would exist between the nucleotide difference and isolation year difference would exist, we were surprised by the triangular pattern, shown in Figure 1.

Alignments

To further examine the sequence relationships in the different strain clusters, we assembled working sets of sequences for each cluster in the Influenza Research Database (IRD) (www.ird-db.org).

We aligned representative sequences from strains in cluster 1:

• Sorted sequences by ascending isolation years,
• We observed a persistence of amino acid changes in the sequences over time.

We aligned representative sequences from strains in cluster 2 (Figure 3):

• We observed no persistence. Amino acid changes appeared to occur randomly.

Quantitate Alignments

To quantify the difference we were observing in the alignments, we developed a metric to count the number of amino acid changes at each position.

A/California/04/2009 represents a lack of persistence of amino acids changes (white background).

Phylogenetic Trees

Phylogenetic trees for cluster 1 and 2 sequences were generated

• Trees are rooted at the earliest strain with colored subtrees.
• Observed a gradual progression of strains with short branch lengths in the cluster 1 tree.

In contrast, cluster 2 trees lacked the gradual progression of strains and appeared to possess longer branch lengths.

Tree Quantitation

We compared the average branch lengths of each tree and performed a T-test with the results shown in Table 2.

Results confirm that cluster 1 shows a gradual accumulation of sequence alterations as evidenced by relatively short branch lengths. In addition, the gradual transition in branches seemed to reflect the normal progression of time based on the year of isolation.

Table 3. Shape of the evolutionary trajectory clusters and the corresponding transition rates for influenza A virus segments.

<table>
<thead>
<tr>
<th>Segment</th>
<th>K</th>
<th>Slope</th>
<th>Slope</th>
<th>Exp. Determined Mutations Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>PB2</td>
<td>6.9</td>
<td>24.9</td>
<td>4.3</td>
<td>1 (1)</td>
</tr>
<tr>
<td>PB1</td>
<td>7.7</td>
<td>26.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA</td>
<td>6.0</td>
<td>23.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HA</td>
<td>5.7</td>
<td>28.8</td>
<td>5.7</td>
<td>2 (2)</td>
</tr>
<tr>
<td>NP</td>
<td>3.2</td>
<td>18.2</td>
<td>3.6</td>
<td>1</td>
</tr>
<tr>
<td>NA</td>
<td>3.9</td>
<td>23.1</td>
<td>3.2</td>
<td>6 (2)</td>
</tr>
<tr>
<td>M</td>
<td>4.4</td>
<td>1.5</td>
<td>1.5</td>
<td>6 (7)</td>
</tr>
<tr>
<td>NS</td>
<td>2.1</td>
<td>12.5</td>
<td>1.5</td>
<td>2.6</td>
</tr>
</tbody>
</table>


Conclusion

In conclusion, we have shown that sequence similar to the Pandemic H1N1 strain as identified by BLAST analysis fall into different categories based on other measures of sequence comparison. We propose that:

• Cluster 1 represents the true “Evolutionary Trajectory” (ET) of each segment of the A/California/04/2009 pandemic H1N1 2009 strain.
• Cluster 2 contains sequences that are “Similar, but Unrelated” (SUR).
• Cluster 3 contains sequences that are only moderately similar and also unrelated.

It has become apparent through this analysis that sequence similarity, in the absence of a time component, is not a stringent enough criterion to determine relationships among sequences.