EVOLUTIONARY TRAJECTORY ANALYSIS: RECENT ENHANCEMENTS

R. Burke Squires
Pandemic H1N1 2009 Origin?

- April / May 2009
  - Cases of an Influenza-like Illness (ILI) occurred in California, Texas and Mexico
- New strain of influenza was found to be the cause.
  - Pandemic H1N1 2009 influenza virus - 1st pandemic strain of the 21st Century
- Like many, we also wondered what the origin of the virus and its segments were.
- Our analysis
  - Different conclusions
  - More accurate description of virus lineage
Original Analysis

- Reference strain
  - A/California/04/2009

- BLAST
  - Each segment against all flu sequences in IRD
  - Return top 1000 hits (June 2009)

- Graph
  - Nucleotide differences vs. isolation year differences
Nucleotide Diff. vs Isolation Year Diff. – Seg 5 (NP)
The Pattern Repeats

Segment 7 (M)

Segment 8 (NS)

Segment 4 (HA)

Segment 6 (NA)
Characterizing Groups: Seg 5 (NP)

Group “A”

Group “B”
Evolutionary Trajectory

Similar, Distantly Related

Evolutionary Trajectory (E.T.)
## Evolutionary Trajectory Slopes vs. Mutation Rate

<table>
<thead>
<tr>
<th>Segment</th>
<th>E. T. Slope</th>
<th>S.U.R. Slope</th>
<th>Mutation Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>PB2</td>
<td>6.8</td>
<td>24.9</td>
<td>4.3</td>
</tr>
<tr>
<td>PB1</td>
<td>7.6</td>
<td>26.9</td>
<td></td>
</tr>
<tr>
<td>PA</td>
<td>5.9</td>
<td>23.2</td>
<td></td>
</tr>
<tr>
<td>HA</td>
<td>5.5</td>
<td>28.8</td>
<td>5.7</td>
</tr>
<tr>
<td>NP</td>
<td>2.9</td>
<td>18.2</td>
<td>3.6</td>
</tr>
<tr>
<td>NA</td>
<td>3.8</td>
<td>23.1</td>
<td>3.2</td>
</tr>
<tr>
<td>M</td>
<td>1.3</td>
<td>5.6</td>
<td>1.5</td>
</tr>
<tr>
<td>NS</td>
<td>2.0</td>
<td>12.5</td>
<td>1.6</td>
</tr>
</tbody>
</table>
Recent Enhancements

- Dynamic BLAST of reference sequence
- Alignment of BLAST hits
- Distance matrix scoring of aligned sequences
- Estimation of genetic distance
- Selection of ancestral sequences
Dynamic BLASTing

- 1000 BLAST Results – overkill
- BLAST reference sequence with increasing 50 hit results
  - Use default BLAST options
  - BLAST against all influenza sequences isolated before 2009
- Extract the oldest year of isolation from hits
- Stop when the oldest year of isolation does not change between iterations
- Using CDS sequences; not full length sequences
Alignment of BLAST Hits

- Align all of the BLAST hits using MAFFT
- Order alignment by year of isolation
Distance Matrix

- Calculate distance matrix aligned sequences
  - Utilizing Criscuolo & Michel codon model
  - DNAdistree

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>Y</td>
</tr>
<tr>
<td>2</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Y</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

Genetic Distance Estimation

- Find lowest slope of genetic distance / isolation year difference when compared to reference sequence (2009)
- Use that genetic distance as a cut off, eliminating all sequences with greater genetic distance
- Use remaining sequences to estimate slope with exponential curve fit
Estimation of Genetic Distance / Year

Genetic Distance

Isolation Year Differences
Estimation of Genetic Distance / Year
Estimation of Genetic Distance / Year

Isolation Year Differences

Genetic Distance

Isolation Year Differences
Estimation of Genetic Distance / Year

Genetic Distance

Isolation Year Differences
Estimation of Genetic Distance / Year

Genetic Distance

Isolation Year Differences
Genetic Distance Per Year - Actual

\[ y = 0.0125x^{0.5184} \]

\[ R^2 = 0.87804 \]
Eliminate Non-ancestral Sequences

- Use genetic distance per year curve fit to eliminate sequences in distance matrix

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Y</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>
## Determine Ancestral Sequences

- Sum number of sequences in a row of matrix
- Examine distribution of sums to determine ancestral sequences
- Fill-in gaps with hypothetical ancestor for each nucleotide change

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Y</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Final Analysis

- Unfortunately, we will have to wait for the publication
- Prospectively linked 2009 H1N1 Pandemic sequence to 1918 “Spanish Flu” pandemic
Summary

- Developed the Evolutionary Trajectory Analysis
- Data-driven approach to ancestry
- Ancestry = similarity component + time component
Acknowledgements

- JCVI / Influenza Research Database (IRD)
  - Richard H Scheuermann
  - Brett Picket
- UT Southwestern / IRD
  - Jyothi Noronha
  - Victoria Hunt
- Erasmus MC / SMU
  - Elizabeth McClellan

My thanks to JCVI / IRD for travel support
Q & A

R. Burke Squires
NIAID, Bioinformatics & Computational Biosciences Branch (BCBB)
richard.squires@nih.gov