Correlation of Patristic Distance with Nominal Specimen Collection Date in Influenza A/H1N1 Hemagglutinin-Encoding Segments

Jack K. Horner
P.O. Box 266
Los Alamos NM 87544 USA

Abstract

The influenza hemagglutinins are viral coat glycoproteins that facilitate viral binding to the host cell wall; as a result, the virulence of any strain of flu depends significantly on how well the hemagglutinin of that strain promotes that binding. Characterizing the evolution of the hemagglutinins is thus fundamental to predicting the virulence of the virus. Here, I describe a linear regression of patristic distance in Influenza A/H1N1 hemagglutinin-encoding segments on the nominal specimen-collection date contained in the label field of the hemagglutinin genomic sequence descriptors; the regression predicts an average mutation rate of ~2 bp/year (implying, on average, ~0.1 mutations in the hemagglutinin active site per year).

Keywords: Influenza, H1N1, hemagglutinin

1.0 Introduction

The influenza hemagglutinins are viral coat glycoproteins that bind to sialic acid residues on the glycoproteins exposed at the surface of the epithelial cells of the host respiratory system. As a result, the virulence of any strain of flu depends significantly on how well the hemagglutinin of that strain promotes that binding. Characterizing the evolution of the hemagglutinins is thus fundamental to predicting the virulence of the virus.

The influenza A viruses responsible for the pandemic of 1918 were derived from avian viruses, which typically recognize the cell-wall glycan SAa2,3Gal. The hemagglutinins of early isolates from humans infected in these pandemics seem to have recognized SAa2,6Gal in preference to SAa2,3Gal, suggesting that conversion of the avian hemagglutinin to one that can recognize SAa2,6Gal-terminated polysaccharides on host cells is an important step in the generation of pandemic strains. The principal amino acid substitutions involved in this shift of receptor recognition are residues 226 and 228 in the H2 and H3 hemagglutinins (equivalent to residues 222 and 224 in the H5 hemagglutinin). The introduction of these mutations into the H5 hemagglutinin permitted its binding to an a2,6 glycan, although neither change has been found in the hemagglutinins of H5N1 viruses isolated from humans ([13]). A first-principles theory of hemagglutinin evolution is highly desirable but currently beyond the state of the art. First-principles computational methods such as molecular dynamics could provide insight into relevant drug-site free-energetics, but such methods are often computationally expensive and in the case of the hemagglutinins, would require an initial, realistic specification of the *in situ* environment. Relatively few H1N1 hemagglutinin structures are available at present, and none address the effect of the molecules' environment on their active sites. In contrast, phylogenetic comparisons

of the genomic encoding of the hemagglutinins might, by translational proxy, provide insight; some phylogenetic methods, furthermore, are computationally inexpensive. Over 10000 hemagglutinin-encoding (HA) segments of the viral genomes are available for A/H1N1 ([4]).

2.0 Method

The general method of this study has four steps: downloading H1N1 HA segment descriptors, aligning the descriptors, computing the patristic distances among the segments, and analyzing the correlation of segment patristic distance with segment collection-date. Unless otherwise noted, all processing described in this section was performed on a Dell Inspiron 545 with an Intel Core2 Quad CPU Q8200 (clocked @ 2.33 GHz) and 8.00 GB RAM, running under the *Windows Vista Home Premium (SP2)* operating environment, connected by a 1.5 Mbit/s DSL link to the Internet.

Influenza H1N1 HA segments were downloaded from the Influenza Research Database ([4]) on 13 January 2011. The query/download parameters are shown in Figure 1.

```
Query parameters:

   Select Segments: 4 (HA)
   Subtype: H1N1
   Date Range: 1915 to 2011
   Geographic Grouping:  All
   Host: All
   Data to Return: Segment/Nucleotide

   Advanced Options:
         Display Fields:  Sequence Accession, Date

   Display:  sort on (increasing) date


Download parameters:
   Select: All segments
   Select Download Type:  Segment FASTA
   Label Sequence By:  Custom -- Accession Number, Date
```

Figure 1. Influenza Research Database ([4]) query/download parameters for the Influenza A/H1N1 HA segment descriptors used in this study.

The file resulting from the previous step was edited in *BioEdit* v7.0.5.3 ([6]) to remove any sequences shorter than 1600 bp or longer than 1800 bp, a range chosen by inspection of the sequence descriptors to include some of the descriptors with the earliest collection dates in the set, while excluding descriptors that were 50% shorter or longer that the average descriptor length in the set. The *BioEdit* navigation for this filtering was

```
Sequence --> Filter Out
  Sequences Containing
   Certain Characters -->
   Delete them  -->  are <x
     [>x] long (x = 1600
    [1800]) -->
     File  --> Save as (type
```

```
    = Fasta, filename =
ten.fasta)
```

If fewer than 10 sequences for a given year were in the resulting file, all sequence descriptors for that year were saved. Else, only the first 10 sequence descriptors in each year were saved. (This helps to reduce time bias in the sample, some of which, due to the scarcity of specimens collected before 1930, is unavoidable). The result was a collection of FASTA-formatted sequence descriptors 1600-1800 bp long. *BioEdit* was then used to save the descriptor Labels of this length- filtered set to a separate file.

The "Label" fields in the FASTA-formatted sequence descriptors obtained from the previous step were edited in *Word 2007* so that each had the form "GenBankAccessionID_yyyy", where yyyy is the year contained in the Label. (In this paper, that year is called the "collection date". It should be noted that such a date is merely part of a free-text field; thus, in principle, that "date" could be, and mean, anything. It is relatively common practice, however, for such a date to represent the date on which the organism from which the sequence was derived was collected.)

The FASTA-formatted sequences from the previous step were aligned using *MAFFT* v6.847b-win32 ([2]), invoked from a *Vista* Command Prompt window. The parameters for the alignment were

```
Order: input
Output format: clustal
Strategy: FFT-NS-i
    (Standard)
Iterative refinement
(Maximum of 2 iterations)
All other parameters:
    defaulted
```

The resulting CLUSTAL-formatted ([11]) file was edited in *Notepad* to remove blank lines and lines containing asterisks.

A *PAUP* ([8]) neighbor-joining (NJ, [12]) script was built in *Notepad*, incorporating the descriptor labels and aligned sequences obtained in previous steps. The template for the *PAUP* script is shown in Figure 2.

```
#NEXUS
begin taxa;
      dimensions ntax=389;
      taxlabels
    [descriptor labels go here (not shown)]
;
end;

begin characters;
      dimensions nchar=1794;
      format missing=? gap=- matchchar=. interleave datatype=dna;
      matrix
    [aligned data goes here (not shown)]
;
end;
begin paup;
   [1]  log start file=H1N1_HA_nj_patdist.log replace;
   [1]  nj;
   [3]  savedist file=tenpatdist.txt format=oneColumn;
end;
```

Figure 2. Template of PAUP script used to obtain the patristic distances used in this study.

Patristic distances from a 1918 "reference" segment (AF117241 in [4]), and corresponding label-times expressed as years-since-1918, were extracted using the *get_pats* software ([7]) running under *Cygwin* (in turn running under *Vista*) from the patristic distance file produced by *PAUP*. The output of *get_pats* is a comma-separated file. This file was converted to a space-separated file using *Notepad*. A linear regression of patristic distance on time was performed by the *Mathematica* ([5] script shown in Figure 3 ([9]).

```
patdistimedata = ReadList[ToFilename[{"C:",
  "BIOCOMP2011",   "Influenza_H1N1_HA"},
    "tenpatdistime.txt"], {Number, Number}];

model=LinearModelFit[patdistimedata,x,x]

model["BestFit"]

Show[ListPlot[patdistimedata, AxesOrigin -> {0,0},
   AxesLabel -> {"Years After 1918", "Patristic Distance from
      AF117247"}], Plot[model["BestFit"], {x, 0, 100}]]

model["ParameterTable"]

model["RSquared"]

model["AdjustedRSquared"]
```

Figure 3. *Mathematica* script used for linear regression in this study.

3.0 Results

10147 sequences were produced by the Influenza Research Database query/download described in Section 2.0.

The length-filtering and time-debiasing steps in *BioEdit* described in Section 2.0 yielded 389 FASTA-formatted sequences.

The *MAFFT* alignment step described in Section 2.0 yielded CLUSTAL-formatted sequence descriptors with 1794 characters per sequence. 388 patristic-distance/time pairs were produced by the patristic-distance/time extraction (via *get-pats*) from the patristic distance file produced by *PAUP*.

The linear regression computed by *Mathematica* was

```
patristic_distance_from_AF117241
    = 0.0621597 +
0.00128348*Years_Since_1918
```

A scatterplot and the best linear fit to that data is shown in Figure 4.

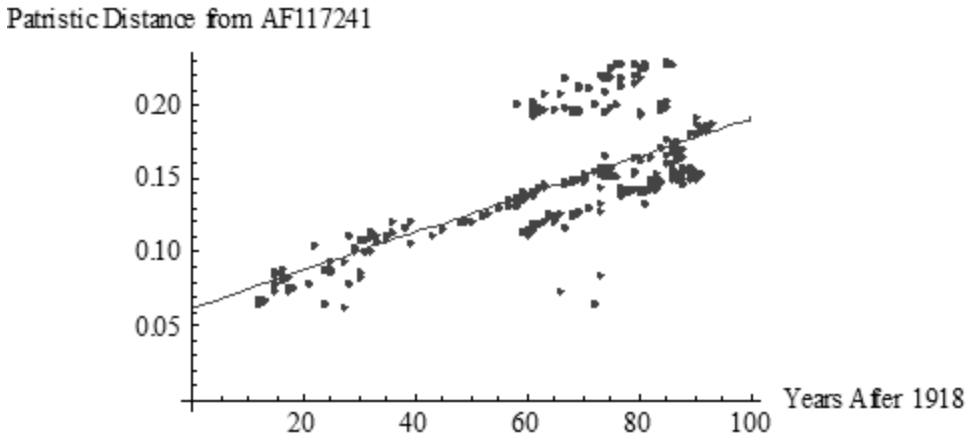

Figure 4. Scatterplot and best linear fit of patristic-distance/time data used in this study.

Some parameter statistics for this regression are:

Parm	Estimate	Standard Error	t Statistic	P-Value
b	0.0621597	0.00437122	14.2202	$3.38122*10^{-37}$
m	0.00128348	0.0000625213	20.5286	$7.74847*10^{-64}$

where b is the intercept on the patristic-distance axis, and m is the slope of the regression. The regression coefficient, r^7, is 0.521937 ; the adjusted r^2, 0.520698.

4.0 Discussion

The method described in Section 2.0 and the results of Section 3.0 motivate several observations:

1. The slope of the regression line suggests that the typical Influenza A/H1N1 HA segment experiences, on average, ~0.001 change per year. Since a nominal HA segment has length ~1700 bp, we would, based on the regression formula in Section 3.0, expect (~1700 bp x ~0.001 =) ~2 bp change per year. Such a change would be sufficient to alter at least one amide in the active site of the hemagglutinin encoded by the segment about every 5 years, if we assume the active site is determined by ~50 bp and that mutations are uniformly distributed across the molecule. This rate is consistent with the nominal mutation rate suggested by other considerations ([10]).

In general, we could not expect "collection date" to provide any information about mutation rate. However, if specimens are collected at a rate that is comparable to the mutation rate (as is the case with flu genomic segments), collection dates will tend to exhibit a strong correlation with mutation rates.

2. In contrast to a similar study performed on H1N1 NA segments ([14]), the regression reported in Section 3.0 is relatively small. Inspection of Figure 4 suggests why this is so. Beginning in ~1978, HA segments diverged into three relatively distinct cohorts, two of which were well removed from a linear extrapolation from earlier segments. This sharp change coincides with the beginning of a flu epidemic in swine in the US.

3. The sequence-descriptor sampling protocol described in Section 2.0 is intended to help mitigate time-biasing in the sample by restricting the number of sequence descriptors sampled per year to no more than 10. The results aren't perfect: for some years, [4] contains fewer than 10 (for some years, no) sequence descriptors. Other protocols are of course possible, but the one used in this study is a practical compromise between under-, or over-, sampling any given year, given the data available in [4].

5.0 Acknowledgements

This work benefited from discussions with Town Peterson of the University of Kansas Biodiversity Institute, George Hrabovsky of the Madison Area Science and Technology Institute for Scientific Computing, Tony Pawlicki, and Richard Barrett. For any problems that remain, I am solely responsible.

6.0 References

[1] Barry JM. *The Great Influenza*. Viking. 2004.

[2] Katoh K and Toh H. Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics* 9 (1 July 2008), 286-298.

[3] Butler D. Avian flu special: The flu pandemic: were we ready? *Nature* 435 (26 May 2005), 400-402. doi: 10.1038/435400a.

[4] Squires B, Macken C, A. García-Sastre A, Godbole S, Noronha J, Hunt V, Chang R, Larsen CN, Klem E, Biersack K, and Scheuerrmann RH. BioHealthBase: informatics support in the elucidation of influenza virus host pathogen interactions and virulence. *Nucleic Acids Research* 36 (Database issue), D497-503 (2008). http://www.fludb.org/brc/home.do?decorator=influenza.

[5] Wolfram Research. *Mathematica Home Edition* v7.0 (2010).

[6] Hall TA. *BioEdit*: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* 41 (1999), 95-98.

[7] Horner JK. *get_pats*, a perl program for extracting patristic distances from a PAUP "one-column" patristic distance file. 2011.

[8] Swofford D. *Phylogenetic Analysis Using Parsimony (PAUP)* v4.0b10. URL http://paup.csit.fsu.edu/. Sinauer Associates. 2004.

[9] Horner JK. *statpats.nb*, a *Mathematica* notebook for performing linear regression of patristic-distance on time. Available from the author on request.

[10] Horner JK. An estimate of the mutation rates of the active sites of Influenza A/H5N1 neuraminidases. *Proceedings of the 2010 International Conference on Bioinformatics and Computational Biology*. CSREA Press. 2010. pp. 344-349.

[11] Higgs DG, Thompson JD, and Gibson TJ. Using CLUSTAL for multiple sequence alignments. *Methods in Enzymology* 266 (1996), 383-402.

[12] Felsenstein J. *Inferring Phylogenies*. Sinauer Associates. 2004.

[13] Yamada S, Suzuki Y, Suzuki T, Le MQ , Nidom CA, Sakai-Tagawa Y, Muramoto Y, Ito M, Kiso M, Horimoto T, Shinya K, Sawada T, Kiso M, Usui T, Murata T, Lin Y, Hay A, Haire LF. Haemagglutinin mutations responsible for the binding of H5N1 influenza A viruses to human-type

receptors. *Nature* 444 (16 November 2006), 378-382. doi:10.1038/nature05264.

[14] Horner JK. Correlation of patristic distance with specimen collection date in Influenza A/H1N1 neuraminidase-encoding segments. *Proceedings of the 2011 International Conference on Bioinformatics and Computational Biology*. CSREA Press. 2011. Forthcoming.