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

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

Abstract

Neuraminidases are viral coat glycoproteins that facilitate the transmission of influenza from cell to cell. Characterizing the evolution of the neuraminidases is essential to effective development and deployment of neuraminidase-inhibitor therapeutics. Here, I describe a linear regression of patristic distance in Influenza A/H1N1 neuraminidase-encoding segments on the nominal specimen-collection date contained in the label field of the neuraminidase genomic sequence descriptors; the regression predicts an average mutation rate of ~1 bp/year (implying, on average, ~0.1 mutations in the neuraminidase active site per year).

Keywords: Influenza, H1N1, neuraminidase

1.0 Introduction

The most widely used anti-influenza therapeutic, oseltamivir (Tamiflu™, [4]), a neuraminidase inhibitor, was decreasingly effective against the dominant influenza strain (an Influenza A/H1N1 mutant) in the US in the 2009 "Spring/Fall" pandemic ([10]). Characterizing the evolution of the neuraminidases is essential to effective development and deployment of neuraminidase-inhibitor therapeutics.

Influenza type A is divided into nine sero-subtypes, and these subtypes correspond at least roughly to differences in the active-site structures of the corresponding neuraminidases. The subtypes fall into two groups ([3]): group-1 contains the subtypes N1, N4, N5 and N8, whereas group-2 contains the subtypes N2, N3, N6, N7 and N9. Oseltamivir was designed to target the group-2 neuraminidases.

The known molecular structures of the neuraminidases are broadly consistent with this sero-taxonomic characterization. The available crystal structures of the group-1 N1, N4 and N8 neuraminidases ([1]) reveal that the active sites of these enzymes have a very different three-dimensional structure from that of group-2 enzymes. The differences lie in a loop of amino acids known as the "150-loop", which in the group-1 neuraminidases has a conformation that opens a cavity not present in the group-2 neuraminidases. The 150-loop contains an amino acid designated Asp 151; the side chain of this amino acid has a carboxylic acid that, in group-1 enzymes, points away from the active site as a result of the 'open' conformation of the 150-loop. The side chain of another active-site amino acid, Glu 119, also has a different conformation in group-1 enzymes compared with the group-2 neuraminidases ([13]).

The Asp 151 and Glu 119 amino-acid side chains form critical interactions with neuraminidase inhibitors. For neuraminidase subtypes with the “open conformation” 150-loop, the side chains of these amino acids might not have the precise alignment required to bind inhibitors tightly ([13]).

The difference in the active-site conformations of the two groups of neuraminidases may also be caused by differences in amino acids that lie outside the active site. This means that an enzyme inhibitor for one target will not necessarily have the same activity against another with the same active-site amino acids and the same overall three-dimensional structure ([17]).

A first-principles theory of neuraminidase 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 neuraminidases, would require an initial, realistic specification of the *in situ* environment. Relatively few H1N1 neuraminidase 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 neuraminidases might, by translational proxy, provide insight; some phylogenetic methods, furthermore, are computationally inexpensive. ~6800 neuraminidase-encoding (NA) segments of the viral genomes are available for A/H1N1 ([7]).

2.0 Method

The general method of this study has four steps: downloading H1N1 NA 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.

Because typical influenza neuraminidases are ~400-mers, ~1200 base-pairs (bp, 3 bp per mer) are required to encode them in the viral genome. Influenza H1N1 NA segments of length at least 1000 bp were downloaded from the Influenza Research Database ([7]) on 13 January 2011. The query/download parameters are shown in Figure 1.

```
Query parameters:

   Select Segments: 6 (NA)
   Subtype: H1N1
   Date Range: 1915 to 2011
   Geographic Grouping:  All
   Host: All
   Data to Return: Segment/Nucleotide

   Advanced Options:
         Minimum segment length:  (Segment 6)  1000
         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 ([7]) query/download parameters for the Influenza A/H1N1 NA segment descriptors used in this study.

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 "GenBank_accession_ID-yyyy", where yyyy is the year represented 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, 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.) Any sequence descriptor that did not contain year information was subsequently deleted using *Word 2007.*

The file resulting from the previous step was edited in *BioEdit* v7.0.5.3 ([9]) to remove any sequences longer than 1450 bp. The *BioEdit* navigation for this filtering was

```
   Sequence    -->   Filter   Out
Sequences   Containing   Certain
Characters --> Delete them   -->
are >x long (x = 1450) --> 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 1000-1450 bp long. *BioEdit* was then used to save the descriptor Labels to a separate file.

The FASTA-formatted sequences from the previous step were aligned using *MAFFT* v6.847b-win32 ([5]), 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 ([16]) file was edited in *Word 2007* to remove blank lines and lines containing asterisks.

A *PAUP* ([12]) neighbor-joining (NJ, [18]) script was built in *Notepad*, incorporating the descriptor labels and aligned sequences obtained in previous steps. Hyphens in the descriptor labels were replaced by underscores. The template for the *PAUP* script is shown in Figure 2.

```
#NEXUS
begin taxa;
      dimensions ntax=385;
      taxlabels
    [descriptor labels go here (not shown)]
;
end;
```

```
begin characters;
      dimensions nchar=1477;
      format missing=? gap=- matchchar=. interleave datatype=dna;
      matrix
    [aligned data goes here (not shown)]
;
end;

begin paup;
   [1]  log start file=H1N1_NA_nj_patdist.log replace;
   [2]  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 (AF250356 in [7]), and corresponding label-times expressed as years-since-1918, were extracted using the *get_pats* software ([11]) 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* ([8] script shown in Figure 3 ([14]).

```
patdistimedata = ReadList[ToFileName[{"C:",
    "BIOCOMP2011",   "Influenza", "Branch_and_age"},
      "tenpatdistime.txt"], {Number, Number}];

  model=LinearModelFit[patdistimedata,x,x]

  model["BestFit"]

  Show[ListPlot[patdistimedata, AxesOrigin -> {0,0},
     AxesLabel -> {"Years After 1918", "Patristic Distance from
        AF250356"}], 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

6816 sequences were produced by the Influenza Research Database query/download described in Section 2.0. 23 sequence descriptors in this file had no identifiable date information and were deleted, netting a file containing 6793 sequence descriptors.

The time-debiasing step in *BioEdit* yielded 385 FASTA-formatted sequences.

The *MAFFT* alignment step described in Section 2.0 yielded 385 CLUSTAL-formatted sequence descriptors with 1477 characters per sequence. 384 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_AF250356
  = 0.00113267*years_since_1918
  +  0.0497421
```

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

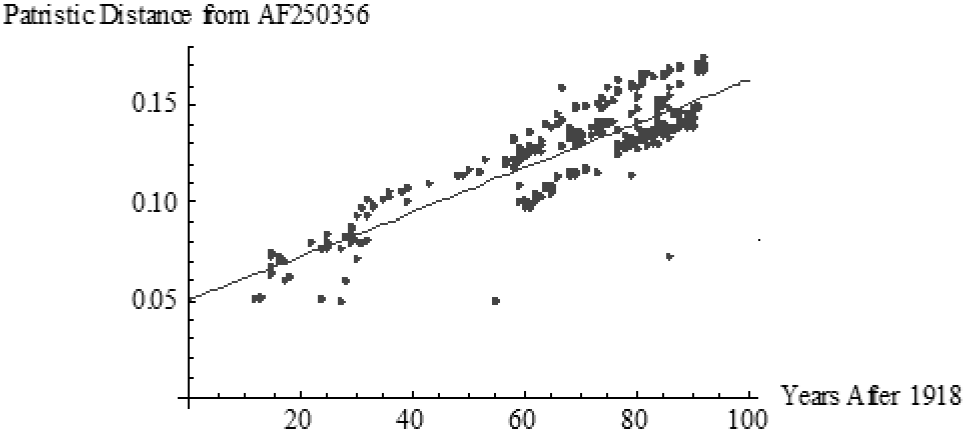

Figure 4. Scatterplot and best linear fit of patristic-distance/time data used in this study.

Some parameter statistics for this regression are:

Parm	**Estimate**	**Standard Error**	**t Statistic**	**P-Value**
b	0.0497421	0.00211585	23.5092	$3.16958*10^{-76}$
m	0.00113267	0.0000301765	37.535	$3.17033*10^{-130}$

where b is the intercept on the patristic-distance axis, and m is the slope of the regression. The regression coefficient, r^2, is 0.786696; the adjusted r^2, 0.786138.

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 NA segment experiences, on average, ~0.001 change per year. Since an NA segment has length ~1000 bp, we would, based on the regression formula in Section 3.0, expect (~1000 bp x ~0.001 =) ~1 bp change per year. Such a change would be sufficient to alter at least one amide in the active site of the neuraminidase encoded by the segment about every 7 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 ([15]).

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. The regression reported in Section 3.0 has robust significance statistics, strongly suggesting that current flu genomic segment sampling and sequencing practices are sufficient to characterize the average mutation rate of the H1N1 NA segments.

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, [7] 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 [7].

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] Russell RJ et al. The structure of H5N1 avian neuraminidase suggests new opportunities for drug design. *Nature* 443 (6 September 2006), 45-49.

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

[3] World Health Organization. A revision of the system of nomenclature for influenza viruses: a WHO memorandum. *Bulletin of the World Health Organization* 58 (1980), 585-591.

[4] Ward P et al. Oseltamivir (Tamiflu) and its potential for use in the event of an influenza pandemic. *Journal of Antimicrobial Chemotherapy* 55, supplement 1 (2005), i5-i21.

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

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

[7] 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.

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

[9] 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.

[10] US Centers for Disease Control. *Summary: Interim Recommendations for the Use of Influenza Antiviral Medications in the Setting of Oseltamivir Resistance among Circulating Influenza A (H1N1) Viruses, 2008-09 Influenza Season.* 19 December 2008. URL http://www.cdc.gov/flu/professionals/antivirals/summary.htm.

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

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

[13] Luo M. Structural biology: antiviral drugs fit for a purpose. *Nature* 443 (7 September 2006), 37-38. doi:10.1038/443037a, published online 6 September 2006.

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

[15] 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.

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

[17] Stoner TD, Krauss S, DuBois RM, Negovetich NJ, Stallknecht DE, Senne DA, Gramer MR, Swafford W, DeLiberto T, Govorkova EA, and Webster RG. Antiviral susceptibility of avian and swine influenza virus of the N1 neuraminidase subtype. *Journal of Virology* 84 (October 2010), 9800-9809.

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