On the origin of influenza A hemagglutinin

Derek Gatherer

Received: 22 September 2009 / Accepted: 30 September 2009
© Association of Microbiologists of India 2009

Abstract Recent advances in phylogenetic methods have produced some reassessments of the ages of the most recent common ancestor of hemagglutinin proteins in known strains of influenza A. This paper applies Bayesian phylogenetic analysis implemented in BEAST to date the nodes on the influenza A hemagglutinin tree. The most recent common ancestor (MRCA) of influenza A hemagglutinin proteins is located with 95% confidence between 517 and 1497 of the Common Era (AD), with the center of the probability distribution at 1056 AD. The implications of this revised dating for both historical and current epidemiology are discussed. Influenza A can be seen as an emerging disease of mediaeval and early modern times.

Keywords Influenza A · Hemagglutinin · H1N1

Introduction

Influenza A was the cause of three pandemics in the 20th century, in 1918, 1957 and 1968, associated with hemagglutinin serotypes H1, H2 and H3 respectively. The arrival of novel hemagglutinin serotypes in influenza A viruses infecting human populations is referred to as antigenic shift. Sixteen different serotypes of influenza A hemagglutinin are found in avian wildfowl, the major reservoir of the disease, and it is believed that zoonotic transfer from avian species, is the major mechanism for antigenic shift in mammals. In addition to the three serotypes introduced to humans in the 20th century pandemics, H5, H7 and H9 have also been found in human influenza A cases, although without extensive human-to-human transmission.

Serological, epidemiological and clinical evidence suggests that the pandemic of 1889 was also caused by influenza A, although hemagglutinin sequences remain to be isolated [1]. Prior to this, historical accounts are suggestive of possible influenza pandemics in 1173, 1510, 1557, 1580, 1729, 1781 and 1830 [2]. The present pandemic, declared by the World Health Organization on 11 June 2009, is of serotype H1N1 [3, 4]. However, it is apparent that this strain is of porcine origin and the H1 hemagglutinin is considerably different to that found in seasonally circulating strains of H1, being about 20–24% different at the amino acid level (cf. the 44–45% difference between H1 and H2) [5]. Although the present pandemic is therefore not technically an antigenic shift, it does represent the intrusion of a considerably novel hemagglutinin into human populations.

Saitou and Nei [6] initially estimated the most recent common ancestor (MRCA) of serotypes H1 to H13 at 200–300 years ago, using an amino acid replacement rate based on H1 (human), H2 (duck), H3 (human) and H11 (duck). Suzuki and Nei [7] later revised this figure upwards by one order of magnitude based on a recalibration of the

D. Gatherer
MRC Virology Unit, Institute of Virology,
University of Glasgow, Church Street,
Glasgow G11 5JR, UK

E-mail: d.gatherer@mrcvu.gla.ac.uk

molecular clock of hemagglutinin using 28 hemagglutinin sequences from duck hosts. In this new scenario, influenza A hemagglutinins evolve relatively slowly in their main aquatic wildfowl host reservoir, with bursts of faster evolution after entry of a novel hemagglutinin into humans via an antigenic shift. This issue has important implications for assessment of historical candidate epidemics. If extant hemagglutinin serotypes have a MRCA only a few hundred years old, then it is likely that candidate influenza pandemics of the Middle Ages did not involve hemagglutinin genes that would be recognisable today or may even have involved ancestral orthomyxoviruses rather than a true influenza A, and influenza A may be regarded as an “emerging disease” [8, 9]. By contrast, if the MRCA of extant serotypes is thousands of years old, all influenza pandemics of the past could have involved hemagglutinin serotypes similar to those found today. Influenza A would therefore be an established disease [10]. Recently, O’Brien et al. [11] have derived a radically shortened estimate of the MRCA of H3 and influenza B hemagglutinins of only 100 years, based on multiple assessments of substitution rate in triplet sets of H3, B and C hemagglutinin sequences. This implies that all the recent pandemics have effectively been due to novel hemagglutinins, and indeed that the viruses we

currently recognise as influenzas A and B did not really exist as such prior to the turn of the 20th century. All previous pandemics, if they can be regarded as influenza pandemics, must therefore have been a result of a proto-influenza orthomyxovirus.

This paper further investigates this reawakened controversy by application of the relaxed clock Bayesian approach of BEAST [12, 13] to reassess the dating of nodes on the hemagglutinin tree, then discusses the implications for the historical study of past pandemics and the prediction of future pandemics.

Materials and methods

Hemagglutinin sequences

Influenza A hemagglutinin protein sequences were downloaded from the NCBI Influenza Resource (<http://www.ncbi.nlm.nih.gov/genomes/FLU/>). For the first stage in the analysis, calculation of the MRCA of each individual serotype, sequences were obtained over a range of dates, balancing speed of analysis with sufficient time depth to

Fig. 1 Phylogenetic tree of influenza A hemagglutinin proteins derived using BEAST. Numbers on nodes are years before 2009, indicating the mean of the Bayesian probability distribution. The branches within each serotype are collapsed (triangles).

Table 1 Dates of most recent common ancestors (MRCA) of influenza A hemagglutinin serotypes

Serotypes	n	Mean	Upper	Lower	Poisson
H1	88	1951	1920	1973	0.046
H2	57	1952	1947	1957	0.062
H3	56	1933	1904	1956	0.039
H4	56	1936	1920	1952	0.034
H5	59	1929	1913	1944	0.068
H6	60	1914	1888	1938	0.099
H7	61	1884	1864	1904	0.110
H8	31	1958	1947	1967	0.021
H9	54	1953	1945	1960	0.070
H10	65	1935	1926	1944	0.043
H11	86	1939	1929	1949	0.043
H12	39	1958	1941	1972	0.030
H13	32	1934	1911	1956	0.102

n: number of sequences used; mean: mean point of Bayesian probability distribution for the date of the MRCA; upper: 95% confidence oldest bound of date of MRCA; lower: 95% confidence most recent bound of date of MRCA; Poisson: diversity of serotype as measured by number of amino acid substitutions per site with Poisson correction [15, 17]

allow accurate estimation of the substitution rate. The range of numbers of sequences within each serotype was from 31 (H8) to 88 (H1), and the range of time depth was from 29 years (H13) to 66 years (H7). For H14, H15 and H16, only 2, 6 and 14 sequences were available respectively, so these were omitted from the first stage. Initial analysis was carried out within each serotype to estimate the MRCA of that serotype. Then a selection of 4 to 6 sequences was taken from each serotype group for estimation of the MRCAs at deeper nodes on the tree. The MRCAs obtained from individual serotype sets and their 95% confidence limits were used to calibrate this more extensive analysis.

Alignments and phylogenetic trees

Within individual serotypes, sequence alignment of the proteins is relatively straightforward and was carried out in MAFFT v.6.708 [14] (<http://align.bmr.kyushu-u.ac.jp/mafft/software>) under default parameters. Sequence diversity was calculated using MEGA v.4.1 [15] (<http://www.megasoftware.net>). For the second stage in the analysis, when different serotypes are aligned, structure-assisted alignment was carried out in MOE v.2008.10 (<http://www.chemcomp.com>), using the following structures: 1RV0 (H1), 1RU7 (H1), 1JSM (H5), 2FK0 (H5), 1JSD (H9), 1EO8 (H3), 1TI8 (H7), 1HA0 (H3), 1RD8 (H1) from PDB [16] (<http://www.pdb.org>). Bayesian phylogenetic analysis was carried out in BEAST v.1.4.8 [12, 13] (<http://beast.bio.ed.ac.uk>). The Blosum62 model of amino acid substitution with a gamma-(4 categories)-plus-invariant distribution was used. The set of 15 duck hemagglutinin proteins from

Saitou and Nei [6] was used in Bayes Factor analysis in BEAST to determine the best model (data not shown). A relaxed uncorrelated lognormal model was selected for the molecular clock with the population model set as a constant size coalescent.

Bayesian estimation was run to convergence, in all cases within 30 million generations, with tree sampling every 1000 generations. For each analysis, half of the trees were discarded as burn-in and the final tree produced by summarization of the remaining trees (or the final 5000 trees whichever was the greater number) and mid-point rooting.

Results

Most recent common ancestors (MRCAs) of individual hemagglutinin serotypes

Table 1 shows the MRCAs for hemagglutinins of each serotype. H14, H15 and H16 could not be included in this initial phase owing to a lack of sequences. The range of the mean of the MRCA age distribution is from 1884 (H7) to 1958 (H8 and H12). In all hemagglutinin serotypes in avian hosts, with the exceptions of H7 and possibly H14–H16 (not assessed), the mean of the MRCA age distribution is within the 20th century. The 95% confidence limits for the MRCA in all serotypes are quite tight. Leaving aside H14–H16 again, the widest ranges are for H3 at 95% confidence intervals of 1904–1956, and H1 at 95% confidence intervals of 1920–1973. The narrowest is H2 at 95% confidence intervals of 1947–1957. The most variable serotype is H7

Table 2 Dates of nodes of the influenza A hemagglutinin serotype tree

Clade	Node	Date	Upper	Lower
All A	953	1056	517	1497
Top half	620	1389	996	1781
Lower half	607	1402	1090	1657
H8, 9, 11, 12, 13, 16	506	1503	1196	1730
H1, 2, 5, 6	393	1616	1341	1786
H7, 15, 10	337	1672	1411	1827
H11 and H13	322	1687	1467	1861
H3, H4 and H14	305	1704	1489	1890
H8, H12, H9	294	1715	1521	1883
H1, H2 and H5	261	1748	1587	1869
H8 and H12	184	1825	1705	1921
H7 and H15	178	1831	1760	1885
H4 and H14	163	1846	1748	1921
H2 and H5	151	1858	1779	1919
H13 and H16	141	1868	1786	1932

node: mean Bayesian age estimate for the node; date: mean age translated into Common Era (AD) date; upper: 95% confidence oldest bound of date of node; lower: 95% confidence most recent bound of date of node.

with an average of 0.11 substitutions per site in pairwise comparisons, after Poisson correction. The least variable is H8 with 0.021 substitutions per site. The correlation between estimated age of MRCA and the diversity of the serotypes is 0.71 (t-test of correlation significant at $p < 0.01$), indicating that the oldest serotypes tend to be the most diverse.

Bifurcation dates in the phylogenetic tree of hemagglutinin serotypes

The phylogenetic tree for influenza A hemagglutinin sequences is shown in Figure 1. Each node is dated at its mean age. In Table 2, these ages are translated into dates and the 95% confidence date intervals given.

Discussion

Comparison with previous studies

This is the first application of Bayesian phylogenetic analysis across the full range of hemagglutinin serotypes. Previously, Smith et al. [18] used BEAST to analyse the origins of serotypes H1, H2 and H3. They concluded that the MRCA of a sample of 61 avian H1 hemagglutinins is located approximately at 1953 with a 95% CI range of 1941–1962 which is in agreement with the estimate derived here of a MRCA at 1950 with 95% CI of 1920–1974. However, Smith et al. also included swine and human sequences in

the same Bayesian estimation. Differing rates of evolution within the three hosts, particularly higher rates within human influenza hemagglutinins [7], do not appear to have disturbed calculation of rates across the whole tree. Using the older method of estimation of molecular clock followed by regression analysis, Suzuki and Nei [7] estimated the MRCA of a sample of 6 avian and 6 avian-like swine H1 hemagglutinins to be 1965 with 99% CI of 1843–1977 and for a set of human, avian and swine H1 sequences the MRCA is at 1849 with 99% CI of 1660–1886. Bayesian analysis using BEAST therefore allows for far tighter limits on MRCA dating. Smith et al. also found that the uncorrelated exponential clock was the best fit for their data set by Bayes Factor analysis whereas the present analysis found that uncorrelated lognormal was more likely.

Epidemiological implications

Influenza A hemagglutinin serotypes are all believed to be ultimately of avian origin, although some may have passed through intermediate hosts, principally pigs, on their route to humans. Therefore, study of the evolution of the virus is complicated by the fact that important evolutionary events such as lineage bifurcations may have taken place in more than one host species. However, in the absence of any strong indication to the contrary, the default hypothesis is to assume that most of influenza's evolution has taken place within its main reservoir, the aquatic wildfowl. The results presented here suggest that the MRCA of all extant

influenza A hemagglutinin serotypes was approximately in the mid 11th century AD, with 95% confidence limits of the early 6th century to the late 15th century. This does not mean that the process of diversification within influenza A hemagglutinin only began at that point. It is possible that an ancestral form of influenza A, comprising extinct serotypes, existed in the more distant past. However, it does imply that prior to the early Middle Ages, infections of influenza A in the avian reservoir and also those that may have been zoonotically transferred to humans, did not involve serotypes known today.

The symptoms of influenza infection are similar to those of many feverish respiratory diseases of both viral and bacterial etiology. Identification of outbreaks of influenza from historical records is therefore inevitably a little conjectural. For instance, the “English Sweat” that caused high mortality and considerable social panic in north-western Europe from 1485 to 1551 had many of the symptoms of influenza, but has also been speculated to have been scarlet fever [19, p. 221], anthrax [20], an enterovirus [21] or an hantavirus [22], all of which seem equally plausible on the basis of symptomology alone. The three pandemic events which contributed to the decline of the Roman Empire: the Antonine Plague of 165–180, the Plague of Cyprian of 251–266 and the Justinian Plague of 541–542 have been conventionally identified as measles, smallpox and bubonic plague respectively [19, pp. 116–124], but influenza has been recently suggested as a cause of both the Justinian Plague and the Black Death of 1346–1349 [23, 24]. The justification for this is that the hemolytic and cytokine storm symptoms occurring in many individuals during the 1918 H1N1 influenza pandemic often led to the appearance of petechiae (skin eruptions) and bullae (lymphatic swellings), also seen in bubonic plague. Other epidemiological reasons have also been advanced in favor of influenza as opposed to bubonic plague, such as its rapid spread and spread into areas with relatively small rodent populations, but these factors could equally well favor other respiratory-borne diseases including the pneumonic variant of bubonic plague.

This study allows us to reconsider these hypotheses, informed by the molecular evolution of influenza A as calculated using Bayesian methods. The two earlier of the three major pandemics of the Roman world occurred prior to the 95% confidence MRCA of known influenza A hemagglutinins. Therefore if the Antonine Plague or the Plague of Cyprian were influenza A, they would have involved extinct serotypes of hemagglutinin. The Justinian Plague of 541 is situated just inside the 95% confidence limits for the MRCA (earliest date 517AD) and cannot therefore be excluded at this confidence level. However, if influenza A were involved, it would have incorporated a highly ancestral form of hemagglutinin. The same applies to the Black Death

of 1346–1349. Figure 1 shows that the initial bifurcation from the MRCA had not yet further bifurcated (mean dates 1389 and 1402 respectively – Table 2), so the lineages extant in the 14th century were most likely to have been ancestral H5/H2/H1/H12/H8/H9/H11/H16/H13 and ancestral H10/H15/H7/H14/H4/H3. Since both of these are ancestral to at least one serotype known to infect humans today (H2/H1 and H3 respectively) there is no reason to exclude either lineage as a potential pandemic pathogen. This, of course, assumes that the mean is the correct date. If the oldest date of the 95% confidence range is used, it is possible that five bifurcation events had occurred before 1341 AD (Table 2), the last of which would be the bifurcation leading to H1/H2/H5 and H6. An ancestral H1/H2/H5 serotype might conceivably have been a potent pathogen. However, this hypothesis requires the use of divergence dates at the limits of their 95% confidence ranges, and therefore must be considered to be far less likely.

Subsequent candidate pandemics [2] in 1510, 1557, 1580, 1729, 1780, 1830 and 1898 may be similarly assessed. Considering the mean node dates in Fig. 1, by about 1868 (the date of the bifurcation of the lineages leading to H16 and H13), all currently extant serotypes of influenza hemagglutinin A were in existence in the avian reservoir. Of the 3 serotypes known to infect humans, H1 bifurcated with the ancestor of H2/H5 in 1748, H2 bifurcated with H5 in 1858 and H3 bifurcated with the ancestor of H4/H14 in 1704. One might therefore postulate that the four subsequent influenza pandemics of 1729 onwards were potentially H3, the three of 1780 onwards potentially H1 and the 1898 epidemic potentially H2. The three pandemics of the 16th century occurred at a time when the tree in Fig. 1 had only achieved four branches.

Such a reassessment of the ages of important evolutionary events in the history of influenza virus can thus help to clarify some of controversies concerning its historical occurrence. It will never be possible to prove the etiology of any ancient pandemic simply by examination of a phylogenetic tree, but it is possible to exclude certain hypotheses on that basis. The results presented here exclude influenza A as it is known today from being the pathogen in pandemics much before the early Middle Ages. Influenza A may conceivably be the pathogen of the Black Death, but if so it must have been a highly ancestral form of the modern serotypes, or possibly an extinct serotype.

Conclusion

This analysis of influenza evolution suggests that influenza A is a relatively recent pathogen in avian populations, gradually emerging in the mediaeval and early modern period. At any

point in this process, it would presumably have been possible for influenza to have zoonotically transferred to humans. It may be noticed that the nodes on the hemagglutinin tree are temporally clustered (Table 2) with a group of four nodes having mean dates of 1672 to 1715 AD and a further group of five having nodes of 1825 to 1868 AD. This may be a coincidental clustering due to stochastic cladogenesis [25]. However, since evolution within hemagglutinin lineages is generally regarded as being under strong selective pressures [26], production of a selective or ecological explanation for the past activity in influenza A hemagglutinin cladogenesis and its absence since the late 19th century, is tempting. Two obvious candidates present themselves: climate change – the northern hemisphere “Little Ice Age” overlaps this period [27]; or rapid population expansion in China from the mid-17th century onwards, and its associated expansion of avian agriculture. Global temperature fluctuations may have disturbed avian migration patterns creating isolated populations where serotype divergence could occur. Likewise, farming practices may have similarly fragmented the gene pools of several wildfowl species, creating local conditions where influenza could rapidly evolve. Current slowing of population growth in the Far East and global warming from the end of the Little Ice Age to the present, might have marked the end of these processes in the 20th century. However, extant serotypes of influenza still pose considerable danger to human populations.

References

- Altschuler EL, Kariuki YM and Jobanputra A (2009) Extant blood samples to deduce the strains of the 1890 and possibly earlier pandemic influenzas. *Medi Hypotheses* 73:846–848
- Potter CW (2001) A history of influenza. *J Appl Microbiol* 91:572–579
- Fraser C, Donnelly CA, Cauchemez S, Hanage WP, Van Kerkhove MD, Hollingsworth TD, Griffin J, Baggaley RF, Jenkins HE, Lyons EJ, Jombart T, Hinsley WR, Grassly NC, Balloux F, Ghani AC, Ferguson NM, Rambaut A, Pybus OG, Lopez-Gatell H, Alpuche-Aranda CM, Chapela IB, Zavala EP, Guevara DM, Checchi F, Garcia E, Hugonnet S and Roth C (2009) Pandemic potential of a strain of influenza A (H1N1): early findings. *Science* 324:1557–1561
- Dawood FS, Jain S, Finelli L, Shaw MW, Lindstrom S, Garten RJ, Gubareva LV, Xu X, Bridges CB and Uyeki TM (2009) Emergence of a novel swine-origin influenza A (H1N1) virus in humans. *N Engl J Med* 360:2605–2615
- Gatherer D (2009) The 2009 H1N1 influenza outbreak in its historical context. *J Clin Virol* 45:174–178
- Saitou N and Nei M (1986) Polymorphism and evolution of influenza A virus genes. *Mol Biol Evol* 3:57–74
- Suzuki Y and Nei M (2002) Origin and evolution of influenza virus hemagglutinin genes. *Mol Biol Evol* 19:501–509
- Webster RG (1998) Influenza: an emerging disease. *Emerg Infect Dis* 4:436–441
- Webster RG, Wright SM, Castrucci MR, Bean WJ and Kawaoka Y (1993) Influenza--a model of an emerging virus disease. *Intervirology* 35:16–25
- Meissner HC (2001) Influenza: emerging control of an old disease. *Pediatr Emerg Care* 17:465–470
- O'Brien JD, She ZS and Suchard MA (2008) Dating the time of viral subtype divergence. *BMC Evol Biol* 8:172
- Drummond AJ and Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* 7:214
- Drummond AJ, Ho SY, Phillips MJ and Rambaut A (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biol* 4:e88
- Katoh K, Kuma K, Toh H and Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* 33:511–518
- Tamura K, Dudley J, Nei M and Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution* 24:1596–1599
- Berman H, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, Shindyalov I and Bourne P (2000) The Protein Data Bank. *Nucl Acids Res* 28:235–242
- Zuckerkindl E and Pauling L (1965) Evolutionary divergence and convergence in proteins, p. 97–166. In: Bryson V and Vogel H (ed.), *Evolving Genes and Proteins*. Academic Press, New York.
- Smith GJ, Bahl J, Vijaykrishna D, Zhang J, Poon LL, Chen H, Webster RG, Peiris JS and Guan Y (2009) Dating the emergence of pandemic influenza viruses. *Proc Natl Acad Sci U S A* 106:11709–11712
- McNeill WH (1977) *Plagues and Peoples*. Basil Blackwell, Oxford.
- McSweeney E (2004) Anthrax and the etiology of the English sweating sickness. *Med Hypotheses* 62:155–157
- Hunter PR (1991) The English sweating sickness, with particular reference to the 1551 outbreak in Chester. *Rev Infect Dis* 13:303–306
- Bridson E (2001) The English 'sweate' (*Sudor Anglicus*) and hantavirus pulmonary syndrome. *Br J Biomed Sci* 58:1–6
- Altschuler EL and Kariuki YM (2009) Was the Justinian Plague caused by the 1918 flu virus? *Med Hypotheses* 72:234
- Altschuler EL and Kariuki YM (2008) Did the 1918 flu virus cause the Black Death? *Med Hypotheses* 71:986–987
- Wollenberg K, Arnold J and Avise JC (1996) Recognizing the forest for the trees: testing temporal patterns of cladogenesis using a null model of stochastic diversification. *Mol Biol Evol* 13:833–849
- Hay AJ, Gregory V, Douglas AR and Lin YP (2001) The evolution of human influenza viruses. *Philos Trans R Soc Lond B Biol Sci* 356:1861–1870
- Robock A (1979) The "Little Ice Age": Northern Hemisphere average observations and model calculations. *Science* 206:1402–1404