A Simple Measure of the Dynamics of Segmented Genomes: An Application to Influenza

Stéphane Aris-Brosou[1,2]

[1] Department of Biology and Center for Advanced Research in Environmental Genomics, University of Ottawa, Ottawa, ON K1N 6N5, Canada
[2] Department of Mathematics and Statistics, University of Ottawa
sarisbro@uottawa.ca

Abstract. The severity of influenza epidemics, which can potentially become a pandemic, has been very difficult to predict. However, past efforts were focusing on gene-by-gene approaches, while it is acknowledged that the whole genome dynamics contribute to the severity of an epidemic. Here, putting this rationale into action, I describe a simple measure of the amount of reassortment that affects influenza at a genomic scale during a particular year. The analysis of 530 complete genomes of the H1N1 subtype, sampled over eleven years, shows that the proposed measure explains 58% of the variance in the prevalence of H1 influenza in the US population. The proposed measure, denoted nRF, could therefore improve influenza surveillance programs at a minimal cost.

1 Introduction

In March 2009, a new influenza A virus emerged in Mexico and in the United States, and spread quickly to the rest of the world in the following weeks. On May 11, the World Health Organization declared the situation as the first pandemic of the 21st century (e.g., [21]). A number of studies have now confirmed that this particular virus emerged following a series of particular exchanges of genetic material between viruses circulating in different hosts (e.g., [21]). One lesson learned during this outbreak is that nobody expected this very particular virus to emerge and to cause a human pandemic. This lack of perspective can essentially be attributed to the very traditional way the dynamics of these genomes are monitored: by studying the history (phylogeny) of each segment (i) independently and (ii) over a long period of time simultaneously analyzing several years (e.g., [12, 15, 21]). The objective of this work is to provide us with a simple tool that can be used to assist surveillance programs by monitoring the dynamics of influenza viruses at the genomic level, (i) integrating the information about the history of all segments simultaneously and (ii) on a yearly basis in order to be able to track the dynamics of these genomes in time.

Influenza viruses are the etiologic agents of the 'flu', a seasonal illness that in a 'regular' season kills 250,000-500,000 people, globally [18]. The virus itself is

E. Tannier (Ed.): RECOMB-CG 2010, LNBI 6398, pp. 149–160, 2010.

made of a protein capsid that encloses its genome, along with a few structural proteins. The influenza genome is composed of eight segments of negative-sense single-stranded RNA molecules, each of which encodes 1-2 proteins for a total of 12 proteins [24]. By approximate order of decreasing size, these genes code for polymerase subunits (PB2, PB1 and PA), the hemagglutinin (HA) and neuraminidase (NA) antigens, a nucleoprotein (NP), a ribonucleoprotein exporter (NS2, also called NEP), an interferon antagonist (NS1), an ion channel protein (M2) and a matrix protein (M1). The last two of the twelve proteins, PB2-F1 [5] and PB1-N40 [24], are less well characterized. Five major different types of influenza viruses are recognized, A B and C being the three most common. The bases of the division into types is made according to genetic information (phylogenies), mutation rates (A > B > C) or host range (A: a large number of vertebrates, B: humans and seals, C: humans and swines).

Influenza A viruses are the principal source of epidemics in the human population, and unlike the other two common types, are further subdivided into subtypes. This subtype classification is based on the type of HA and NA proteins, and therefore genes, that each virus is made of. Almost all combinations can be formed between the 16 known subtypes of HA (H1-H16) and the nine that are known for NA (N1-N9) [11, p.157]. All of these subtypes are present in wild waterfowl, but the most prevalent subtypes in the human population are H1N1 and H3N2 [18].

Because of their structure as single-stranded RNA molecules and as segmented genomes, influenza A viruses evolve quickly under two general mechanisms: antigenic drift and antigenic shift [14]. Antigenic drift is caused by the accumulation of mutations, which occur at a high rate ($\sim 10^{-3}$ substitutions per site per year [18]) due to the lack of a proof-reading mechanism during replication of the viral genome, while antigenic shift is due to the exchange of segments when at least two different viruses, potentially of different subtypes, co-infect the same cell. Such an exchange is called reassortment. With the potential to generate new combinations of antigenes and potentially new subtypes, reassortment is the main source of antigenic novelty [15,14], and has been directly implicated in the emergence of the 2009 pandemic [21]. However, the methods used to quantify reassortments are virtually inexistent. The current practice consists, first, in estimating a phylogenetic tree for each individual segment for a set of genomes sampled through many years, and then, in reconciling these trees, optimally using a cophylogenetic method [4] in order to obtain a snapshot of the history of reassortment. Although this approach allows us to reconstruct what happened *a posteriori* [21], it does not allow us to quantify the dynamics of influenza genomes in real time. Hence, the predictive power of the current approach is vanishingly small.

Here I describe a method that permits the quantification of the dynamics of viruses with segmented genomes, and apply it to a follow-up of H1N1 influenza A viruses through the eleven years between 2000 and 2010. The method, called nRF, is based on a measure of the dissimilarity of gene trees estimated for the different segments of the genome. To use nRF, I further introduce the notion

of a *genome tree*, which is a tree whose leaves represent the different genes constituting the influenza genome, and whose topology represents the amount of reassortment occurring between the different segments. If reassortment actually drives the evolution of influenza genomes [14], we should expect to observe a correlation between a quantitative measure of reassortment, namely, nRF, and prevalence (proportion of affected individuals in a given population) of the viruses under study. I show here that the nRF measure is able to explain 58% of the variance in prevalence of H1 influenza in the US population.

2 Methods

2.1 Rationale

Under the assumptions that (i) reassortment drives the evolution of influenza genomes [14] and (ii) there is no recombination [3, 2], the motivation is to find a measure of reassortment, at the scale of the genomes, which are sampled on a yearly basis. For now, I further assume that all the phylogenetic inferences below return the "true" tree, an assumption to which I come back later in section 2.3.

In the absence of reassortment, all ten segments should have exactly the same phylogeny. Therefore, if we compute the pairwise distances between estimated gene topologies, for instance with the Robinson and Foulds (RF) distance [19], we should obtain a matrix of pairwise distances full of zeros. Briefly, the RF distance, also called the symmetric distance, is the number of branch partitions that differ between two topologies.

On the other hand, if a segment has undergone a reassortment event, then its estimated phylogeny should differ from the others'. The matrix of pairwise distances computed as above should now contain a row and a column of nonzero entries. If we use this matrix of pairwise distances between the gene trees to build a new tree, that I call here a *genome tree*, its total tree length will be a measure of the amount of reassortment that occurred between the different segments (*genes*, in actuality). Reciprocally, the proposed measure can also be understood as a measure of linkage, in the sense that segments transmitted together will cluster together in the estimated genome tree.

Because different years can have different numbers of sampled genomes, RF distances should be scaled by their maximum value, $2(n-3)$ for a tree with n leaves, within each year, so that they become comparable across several years. Hereafter, I denote this measure nRF for normalized RF distance.

The hypothesis of interest is then that a correlation should be observed between nRF and a measure of prevalence of influenza A H1N1 in the human population. To test this hypothesis, I used data on the US population affected by H1 viruses from 2000-2010, as available at the US Centers for Disease Control website (`www.cdc.gov/flu/weekly`). Note that these data are theoretically for all H1 viruses, which include several subtypes; this is related to the assay used to type samples; in practice however, the circulation of subtypes other than H1N1 is negligible in the human population. Because (i) the prevalence data is available by season (in the Northern hemisphere: weeks 40-20), while the genome

data collected for this study run from week 1-52 (see below) and (ii) I specifically want to test for the predictive power of the method, I shifted the prevalence data to compare nRF during year y with prevalence during y, rather than the season running between week 40 of year y to week 20 of year $y + 1$.

2.2 Algorithm

The procedure is divided into six steps:

1. sample influenza genomes for one particular year;
2. extract protein-coding sequences (CDSs) within each segment, translate to amino acid sequences [9], align [6] and back-translate to DNA alignments [22];
3. for each CDS, reconstruct phylogenetic trees under maximum likelihood with a mixture of NNI and SPR searches [10], assuming the GTR $+\Gamma$ model of substitution [26, p.33, 44] (optimally: select the most appropriate model of evolution for each data set [16]);
4. compute the matrix of RF distances [8] for each pair of trees and scale by $2(n-3)$ (for n sequences in alignment) to obtain nRF;
5. reconstruct the Neighbor-Joining tree [8] from the matrix of pairwise RF distances to visualize "correlation" among CDSs [optional];
6. compute the tree length of the NJ tree built on normalized RF distance.

The procedure is repeated for as many years as desired or, in practice, years for which there are "enough" available data ($\geq$ 4 genomes, since with $<$ 4 genomes there is only one single unrooted topology and the phylogenetic problem becomes nonexistent). Perl scripts were written to make most of these steps automatic. Unless otherwise specified, the default settings were used for all the programs cited in the above algorithm. Computations were distributed with the `ForkManager` library (available at `search.cpan.org/dist/Parallel-ForkManager/`).

Branch lengths can affect the comparison of two trees in the following way. Consider the two rooted trees ((1, 2) ,3) and (1, (2, 3)). They have different topologies, but if we consider their respective branch lengths, then the same two trees ((1:0.1, 2:0.1):0.001, 3:0.1) and (1:0.1, (2:0.1, 3:0.1):0.001) are actually very similar. Although the trees considered here are all unrooted, a similar effect of branch lengths could artificially increase the effect of reassortment from the perspective of the RF distance. Therefore, in addition to the RF distance, I also computed the Branch Score Distance or BSD [13,8]. As with nRF, I denote the scaled measure $nBSD$.

2.3 Measure of Support

In order to estimate confidence intervals for both nRF and $nBSD$, I performed a bootstrap analysis [7] during step 3 in the algorithm given above. Here I used 100 replicates to keep computations to a minimum while demonstrating the concept. Optimally, one to ten thousand replicates would be preferable.

Steps 4–end of the algorithm are then repeated on each bootstrap replicate. For simplicity's sake, I considered that a bootstrapped genome would be formed by the bootstrap replicates taken replicate by replicate: bootstrapped genome 0, denoted b_0^G, was formed by the set of bootstrapped genes $\{b_0^{\mathrm{PB2}}, b_0^{\mathrm{PB1}}, \ldots\ b_0^{\mathrm{NS1}}\}$, b_1^G was constructed as $\{b_1^{\mathrm{PB2}}, b_1^{\mathrm{PB1}}, \ldots\ b_1^{\mathrm{NS1}}\}$, *etc.*. This matches the null assumption of independence between segments (*genes*, here), although it might be preferable to take physical linkage into account for the genes on segments 7 on the one hand (genes M2 and M1) and 8 on the other hand (genes NS2 and NS1).

The bootstrapped trees were summarized by computing the majority-rule consensus tree [7,8], constructed from the bipartitions that appeared in at least 50% of the bootstrapped replicates. Just like with the computation of bootstrapped nRF distances, a bootstrap genome was formed by the bootstrap replicates taken replicate by replicate.

2.4 Sampled Genomes

All publicly-available complete influenza A genomes of subtype H1N1 were extracted from the National Center for Biotechnology Information (available at `www.ncbi.nlm.nih.gov`, [1]) for the eleven years spanning 2000-2010. As of May 26, 2010, a total of 2,435 genomes were available (2000: 80 genomes; 2001: 118; 2002: 8; 2003: 26; 2004: 8; 2005: 25; 2006: 19; 2007: 326 [100 of which were randomly chosen for computational expediency]; 2008: 70; 2009: 1706 [100 of which were randomly chosen]; 2010: 49), sampling across swine and human hosts distributed around the world. Alignments were visually inspected and edited where necessary [23], with a particular attention to genes on segments 2, 7 and 8, which are occasionally misannotated; dubious entries were discarded, which demanded to check that each year had sequences from exactly the same individuals. As a result, the final data sets had the following sizes: 2000: 69; 2001: 96; 2002: 7; 2003: 18; 2004: 7; 2005: 19; 2006: 16; 2007: 100; 2008: 54; 2009: 100; 2010: 44, for a total number of genomes equal to 530. Over the sampled years, the distribution of genomes coming from the US was very uneven (actual numbers: 2000: 2; 2001: 24; 2002: 0; 2003: 13; 2004: 0; 2005: 1; 2006: 6; 2007: 89; 2008: 11; 2009: 44; 2010: 36), with years 2000, 2002, 2004 and 2005 having too few US genomes to perform any phylogenetic study (hence the worldwide genome sampling adopted here). Because the PB2-F1 and PB1-N40 genes are small and not always present and / or correctly annotated, I focused on the ten 'canonical' genes: PB2, PB1, PA, HA, NP, NA, M2, M1, NS2 and NS1. Note that the M genes are both on the same segment (# 7); likewise, the NS genes are both on segment # 8. All alignments used in this study are available at `www.bioinformatics.uottawa.ca/stephane`.

3 Results

3.1 Genome Dynamics

Figure 1 shows the trees based on nRF, the proposed measure for monitoring genome dynamics. A number of results are clear from that figure. First,

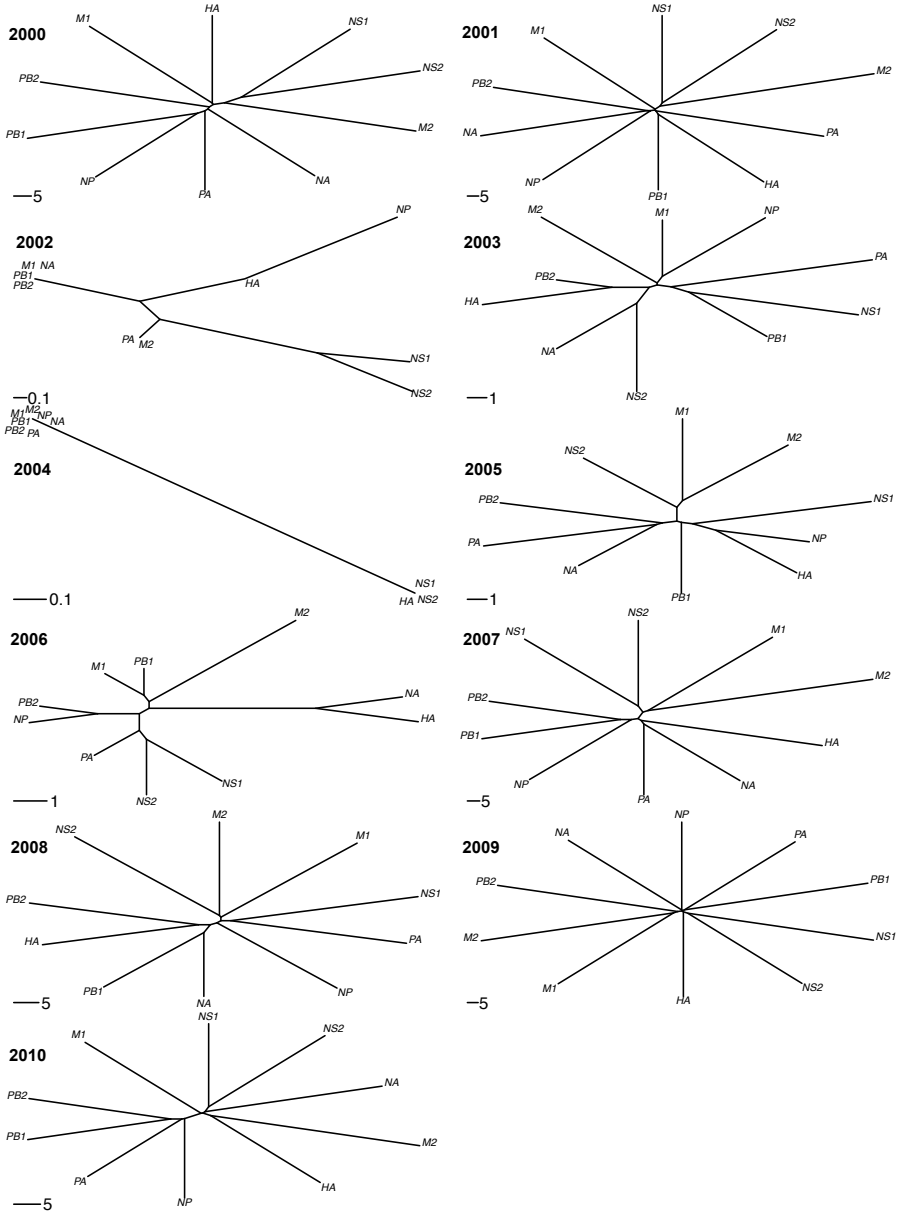

Fig. 1. Genome trees estimated over the eleven years sampled. RF distances were estimated between the gene trees for the ten 'canonical' protein-coding genes of influenza A genomes, and genome trees were reconstructed from the from the pairwise RF distances by NJ. Bootstrap support values not shown (see Figure 2).

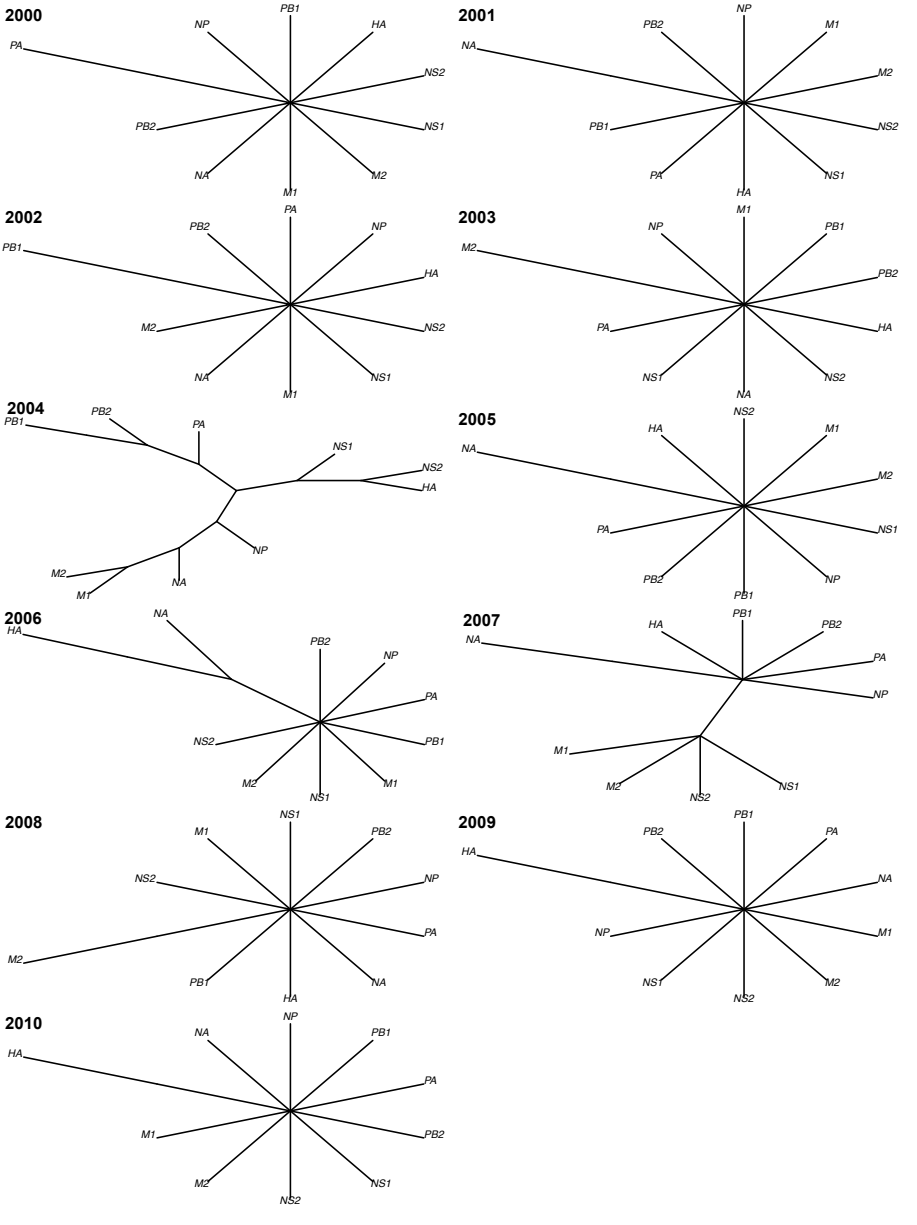

Fig. 2. Majority-rule consensus of the 100 bootstrapped genome trees estimated over the eleven years sampled. Internal branch lengths are nonzero only if the node bipartition they sustain is > 0.75; otherwise, they are set to an arbitrary nonzero number.

reassortment is a rampant process affecting these viruses every year. In the eleven years analyzed here, all the genome trees have a nonzero tree length.

Second, there is extensive variation in the amount of reassortment observed from year to year, with tree lengths in unit of normalized RF distance ranging from 0.05 in 2004 to 0.42 in 2009 (which, according to the confidence intervals calculated below, is significant). Despite the fact that 2004 is the year where only seven genomes were analyzed and 2009 comprises 100 genomes (see Methods), a robust linear regression [27] shows that there is no evidence for an association between tree lengths and the size of the data sets analyzed at the 1% level ($P = 0.037$). This result shows that the nRF measure is robust to sample size.

Third, linkage or reassortment patterns are expected, such as those for the M and NS genes, which are respectively encoded on segments 7 and 8. This is visible on trees where the M2/M1 and NS2/NS1 genes cluster together. But in most years (eight years out of eleven, or 73% of the time for the M genes), the pattern is due to very short internal branch lengths, and is probably artifactual. Indeed, when the consensus trees are computed over the 100 bootstrap replicates (Figure 2), almost all evidence of linkage disappears. Some exceptions exist, such as in 2006 and 2007, but the consensus trees show that (i) all segments are extensively exchanged among individual viruses, so that no clear linkage seems to exist, and (ii) the year of the H1N1 pandemic, 2009, was no exception: this particular pandemic was not preceded and does not exhibit any particular linkage between segments or the genes encoded on these segments.

3.2 *nRF* and *nBSD* as Predictors of Prevalence

Figure 3 shows the estimated nRF values plotted against the yearly log prevalence of H1 infections in the US population. The linear model fitted to these data is highly significant ($F_{1,9} = 14.59$; $P = 0.0041$; adjusted $R^2 = 0.576$). This means that more than half (58%) of the prevalence of influenza H1 in the US is determined by the genome dynamics at the global scale as measured by nRF. A robust linear regression gives similar results ($P = 0.0006$, $R^2 = 0.181$; P(M-bias) = 0.7402, P(LS-bias) = 0.8850), and this regression remained highly significant even after removing the 2004 data ($P = 7 \times 10^{-6}$, $R^2 = 0.271$; P(M-bias) < 0.0001, P(LS-bias) = 0.7294). Therefore, the regression is not solely driven by this point (see Figure 3), and the predictive power of nRF is not an artifact of the sampled data.

We also fitted the same kind of model with nRF of H1 viruses against the prevalence of H3 viruses. In this case, nRF carries no predictive power for this influenza subtype ($F_{1,9} = 3.22$; $P = 0.1063$; adjusted $R^2 = 0.182$), as expected since H1 and H3 are different subtypes.

However, contrary to the expectation that including knowledge of branch lengths might improve prediction, there was still some significant predictive power with $nBSD$ ($F_{1,9} = 10.12$; $P = 0.0111$), but R^2 decreased to 0.477 (compared to 0.576 with nRF). This lack of significance at the 1% level is probably due to the rate heterogeneity across segments, which can be substantial [18]; indeed, in presence of such rate heterogeneity, estimated branch lengths are going

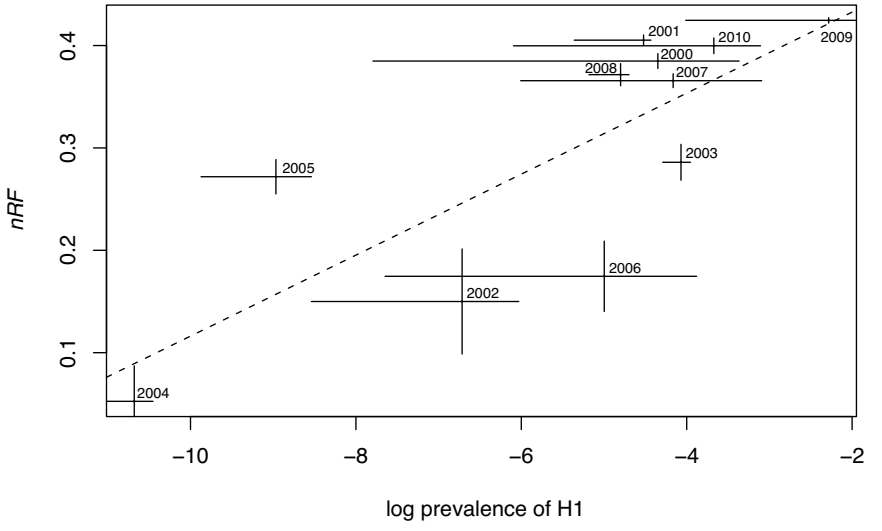

Fig. 3. nRF **as a function of the log prevalence of H1 subtypes.** The broken line represents the fitted regression ($P = 0.0041$). Vertical bars: 2 standard deviations; horizontal bars: upper and lower 5% quantiles of the distribution of prevalence over a calendar year.

to be dramatically different from one segment to the other, which is expected to increase BSD if the tree topologies compared share at least one branch bipartition.

4 Discussion

This work represents a proof of concept demonstrating the possibility of monitoring the severity of epidemics based on a simple measure. The measure described, nRF, intuitively relies on our understanding of the basic biology of RNA viruses, in that reassortment of segmented viral genomes generates new properties with respect to antigenicity, pathogenicity, virulence *etc.*, all of which can compromise the integrity of the immune system of their hosts, and potentially lead to epidemics or pandemics. Our results demonstrate that nRF is a powerful proxy for prevalence, at least in the case of influenza A viruses of subtype H1 over the eleven years and 530 genomes studied here.

One of the reasons why the method has good predictive power is because intra-segmental recombination is negligible in influenza A viruses [3, 2]. If this were not the case, then recombination would affect each individual gene tree and would confound the reassortment signal. As a result, the method described here only applies to non-recombining segmented viral genomes such as influenza A viruses. Independent tests of the nRF measure should be performed on other such viruses, which probably include most single-stranded negative sense segmented RNA viruses [17] (for a full list of segmented RNA viruses, see [11, p.4]),

for which we have prevalence data; potential candidates include the Rift Valley Fever virus [20], as well as Ebolavirus and Marburgvirus or Lyssavirus (rabies). An extension to recombining genomes is however possible through the use of ancestral recombination graphs [25].

In the context of the influenza data analyzed here, a number of points might demand to be refined in order to fully demonstrate the power of the approach. First, sampling was done on a yearly basis, not on a seasonal basis, and genomes were sampled on a global scale while compared to prevalence data from the US only. Yet, I was able to demonstrate the existence of a significant ($P = 0.0041$) and reasonably good predictive power ($R^2 = 0.576$). This suggests two mutually exclusive interpretations: (i) the approach returns a random result, which is unlikely given both the size of the data analyzed (eleven years, 530 complete genomes) and the robustness of the results to the use of robust linear regression; (ii) prevalence in the US is representative of the global dynamics of influenza A subtype H1, and that the nRF method is quite powerful in predicting the severity of an epidemic as measured by prevalence. Second, intra-host dynamics were not studied, as a few genomes from swine hosts were present in the data analyzed here. It is likely that the inclusion of these genomes from swine hosts did not affect the results because most of the genomes (93.21%) were coming from human hosts. Third, the rates of evolution, in units of substitutions per site per year, were not analyzed in correlation with the nRF measure of genome dynamics. Considering rates of evolution in conjunction with nRF would represent a further development of the method. Note however that all the three points above are linked to the sampling scheme used in this work, and therefore do not affect the approach described to monitor genome dynamics through time by means of nRF. A last potential caveat is the expected correlation between prevalence and sequencing effort: it can indeed be expected that years of high prevalence lead to an increased surveillance effort and hence to the deposition of a larger number of genomes in publicly-available databases, such as in 2009 with the H1N1 pandemic. Although there was no significant association between the length of genomes trees and the number of fully sequenced genome in public repositories (section 3.1), more independent data sets should be examined to further demonstrate the power of the nRF measure.

Could we have predicted the 2009 pandemic? Maybe, in the sense that nRF measures for 2009 (evaluated from data sampled between the end of the 2008-2009 season and the beginning of 2009-2010 season) were relatively high. Yet, the question of the existence of a threshold for nRF above which the emergence of a pandemic becomes likely is still open, as nRF measures for 2001 were also high (Figure 3), albeit significantly lower than in 2009, and no pandemic occurred back in 2001.

Finally, in the wake of the 2009 pandemic, it has been realized that the current surveillance system failed to track the emerging pathogens [21]. The approach and results presented here demonstrate that even in the absence of such information on genomes from the past, it is still possible to implement a cost-effective warning system based on the nRF of *currently* circulating genomes, provided

that the surveillance programs track full influenza genomes rather than limiting their effort to HA and NA sequencing.

Acknowledgments

I wish to thank three reviewers for very insightful comments that helped improve this paper. This work was funded by the Natural Sciences Research Council of Canada and by the Canada Foundation for Innovation to SAB.

References

1. Bao, Y., Bolotov, P., Dernovoy, D., Kiryutin, B., Zaslavsky, L., Tatusova, T., Ostell, J., Lipman, D.: The influenza virus resource at the National Center for Biotechnology Information. J. Virol. 82(2), 596–601 (2008)
2. Boni, M.F., de Jong, M.D., van Doorn, H.R., Holmes, E.C.: Guidelines for identifying homologous recombination events in influenza A virus. PLoS One 5(5), e10434 (2010)
3. Boni, M.F., Zhou, Y., Taubenberger, J.K., Holmes, E.C.: Homologous recombination is very rare or absent in human influenza A virus. J. Virol. 82(10), 4807–4811 (2008)
4. Charleston, M.A., Perkins, S.L.: Traversing the tangle: algorithms and applications for cophylogenetic studies. J. Biomed. Inform. 39(1), 62–71 (2006)
5. Chen, W., Calvo, P.A., Malide, D., Gibbs, J., Schubert, U., Bacik, I., Basta, S., O'Neill, R., Schickli, J., Palese, P., Henklein, P., Bennink, J.R., Yewdell, J.W.: A novel influenza A virus mitochondrial protein that induces cell death. Nat. Med. 7(12), 1306–1312 (2001)
6. Edgar, R.C.: MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5), 1792–1797 (2004)
7. Felsenstein, J.: Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39(4), 783–791 (1985)
8. Felsenstein, J.: PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle (2005)
9. Gilbert, D.: Sequence file format conversion with command-line readseq. Curr Protoc Bioinformatics Appendix 1, Appendix 1E (2003)
10. Guindon, S., Delsuc, F., Dufayard, J.F., Gascuel, O.: Estimating maximum likelihood phylogenies with PhyML. Methods Mol. Biol. 537, 113–137 (2009)
11. Holmes, E.C.: The evolution and emergence of RNA viruses. Oxford series in ecology and evolution. Oxford University Press, Oxford (2009)
12. Holmes, E.C., Ghedin, E., Miller, N., Taylor, J., Bao, Y., St. George, K., Grenfell, B.T., Salzberg, S.L., Fraser, C.M., Lipman, D.J., Taubenberger, J.K.: Whole-genome analysis of human influenza a virus reveals multiple persistent lineages and reassortment among recent h3n2 viruses. PLoS Biol 3(9), e300 (2005)
13. Kuhner, M.K., Felsenstein, J.: A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11(3), 459–468 (1994)
14. Nelson, M.I., Holmes, E.C.: The evolution of epidemic influenza. Nat. Rev. Genet. 8(3), 196–205 (2007)

15. Nelson, M.I., Simonsen, L., Viboud, C., Miller, M.A., Taylor, J., George, K.S., Griesemer, S.B., Ghedin, E., Ghedi, E., Sengamalay, N.A., Spiro, D.J., Volkov, I., Grenfell, B.T., Lipman, D.J., Taubenberger, J.K., Holmes, E.C.: Stochastic processes are key determinants of short-term evolution in influenza A virus. PLoS Pathog 2(12), e125 (2006)
16. Posada, D., Crandall, K.A.: MODELTEST: testing the model of DNA substitution. Bioinformatics 14(9), 817–818 (1998)
17. Pringle, C.R.: The Bunyaviridae, chap. Genetics and genome segment reassortment, pp. 189–226. Plenum Press, New York (1996)
18. Rambaut, A., Pybus, O.G., Nelson, M.I., Viboud, C., Taubenberger, J.K., Holmes, E.C.: The genomic and epidemiological dynamics of human influenza A virus. Nature 453(7195), 615–619 (2008)
19. Robinson, D.F., Foulds, L.R.: Comparison of phylogenetic trees. Mathematical Biosciences 53(1-2), 131–147 (1981)
20. Sall, A.A., Zanotto, P.M., Sene, O.K., Zeller, H.G., Digoutte, J.P., Thiongane, Y., Bouloy, M.: Genetic reassortment of Rift Valley fever virus in nature. J. Virol. 73(10), 8196–8200 (1999)
21. Smith, G.J.D., Vijaykrishna, D., Bahl, J., Lycett, S.J., Worobey, M., Pybus, O.G., Ma, S.K., Cheung, C.L., Raghwani, J., Bhatt, S., Peiris, J.S.M., Guan, Y., Rambaut, A.: Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature 459(7250), 1122–1125 (2009)
22. Suyama, M., Torrents, D., Bork, P.: PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34(Web Server issue), W609–W612 (2006)
23. Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., Barton, G.J.: Jalview version 2: a multiple sequence alignment editor and analysis workbench. Bioinformatics 25(9), 1189–1191 (2009)
24. Wise, H.M., Foeglein, A., Sun, J., Dalton, R.M., Patel, S., Howard, W., Anderson, E.C., Barclay, W.S., Digard, P.: A complicated message: Identification of a novel PB1-related protein translated from influenza A virus segment 2 mRNA. J. Virol. 83(16), 8021–8031 (2009)
25. Wiuf, C., Hein, J.: Recombination as a point process along sequences. Theor. Popul. Biol. 55(3), 248–259 (1999)
26. Yang, Z.: Computational molecular evolution. Oxford University Press, Oxford (2006)
27. Yohai, V.J.: High breakdown-point and high efficiency robust estimates for regression. The Annals of Statistics 15(2), 642–656 (1987)