ELSEVIER

Vaccine 21 (2003) 1758–1761

Vaccine

www.elsevier.com/locate/vaccine

Applications of bioinformatics and computational biology to influenza surveillance and vaccine strain selection

Derek J. Smith[a,b,c,*]

[a] *Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK*
[b] *Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA*
[c] *Department of Virology, Erasmus University, P.O. Box 1738, Dr. Molewaterplein 50, 3015 GE Rotterdam, The Netherlands*

Abstract

In recent years, collaborations often between mathematical and computational biologists and scientists in the World Health Organization (WHO) global influenza surveillance network, have resulted in a number of mathematical and computational advances including: increasing the resolution at which antigenic surveillance data can be analyzed, providing methods for genetic analysis and prediction, and an increased understanding of the determinants of repeated influenza vaccination. These advances increase the information extracted from influenza surveillance and increase the quantitative data available for the vaccine strain selection process. This mathematical and computational work is possible because of the wealth of information collected over many years by the WHO global influenza surveillance network, and further advances will be greatly facilitated by implementation of the proposed strengthening of virological and epidemiological surveillance in the WHO global agenda on influenza surveillance and control.

Keywords: Antigenic maps; Phylogenetic analysis; Cluster analysis; Genetic maps; Epidemiology

Key messages

Methods in the fields of bioinformatics and computational biology are finding new applications in the study of influenza, providing:

- **quantitative data on influenza surveillance; and**
- **insights into the process of vaccine strain selection.**

1. Introduction

New high-throughput methods in biology are producing data at an unprecedented rate. In many areas of biology, traditionally experimental scientists are collaborating with computational scientists from many disciplines to create new methods in bioinformatics and computational biology to analyze these data. New patterns are being discovered which would not be detectable without systematic and automated approaches because of the volume and often noisy nature of the data. Although many of these methods are in their infancy, there is no doubt that major advances in our basic understanding of biology, and practical applications in medicine and public health will ensue.

The influenza surveillance and vaccine strain selection processes are in a good position to take advantage of these new methods. The long-standing WHO global influenza surveillance network has accumulated a great deal of institutional knowledge and has collected an extensive dataset of the antigenic and genetic evolution and epidemiology of influenza [1], a subset of this dataset exists in centralized databases [2–5], and the WHO global agenda on influenza surveillance and control proposes further critical strengthening virological and epidemiological surveillance [6,7].

Application and development of these new methods to influenza surveillance and vaccine strain selection is already ongoing. In recent years, collaborations often between scientists in the WHO global influenza surveillance program and mathematical, evolutionary, and computational biologists have resulted in a number of mathematical and computational advances in understanding the antigenic and genetic evolution of influenza and an increased understanding of the determinants of the efficacy of repeated influenza vaccination. Here we give a brief overview of the data used in the vaccine strain selection process, and then review how mathematical and computational methods are being applied to these data, and what might be possible in the future.

[*] Tel.: +1-917-378-2599; fax: +1-212-202-6455.
E-mail address: dsmith@santafe.edu (D.J. Smith).

0264-410X/03/$ – see front matter © 2003 Elsevier Science Ltd. All rights reserved.
doi:10.1016/S0264-410X(03)00068-9

2. The necessity to update the influenza vaccine, and the data used to do so

Human influenza is a complex pathogen, mostly because of its capacity to vary its surface proteins to escape immune surveillance. There are two main patterns of change [8]. The first, *antigenic shift*, is the result of a new influenza A subtype entering the human population, either directly or indirectly from birds. There is generally little pre-existing immunity to the new subtype and antigenic shifts can cause worldwide influenza pandemics. The second, *antigenic drift*, is the result of changes in existing human influenza viruses, to escape immune surveillance. If the influenza vaccine were not updated to track the antigenic changes in the virus, the vaccine would cease to be effective; hence, twice-yearly vaccine strain selection meetings recommend any necessary updates to the influenza strains used in the vaccine.

The vaccine strain selection decision is primarily based on three criteria: (1) antigenic, is there a significant antigenic difference between emerging strains and the existing vaccine strain; (2) genetic, is there supporting evidence of change in the genetic data; and (3) epidemic, are the emerging strains likely to cause widespread epidemics in the coming season. The data to assess these criteria are generated by the WHO global influenza surveillance network of sentinel physicians, National Influenza Centers, and collaborating centers for reference and research. There is mathematical and computational work ongoing to increase the information available to the vaccine strain selection process for all three of these criteria.

3. Increasing the resolution and visualization of antigenic data

Antigenic changes in the influenza virus are measured using the hemagglutination inhibition (HI) assay. The HI assay has been a tool for vaccine strain selection and research for many years [9,10]. The assay works well for distinguishing major drift variants, but finer-grain differences are difficult to judge reliably. New mathematical methods have been applied to the analysis of HI data which increase the resolution at which antigenic differences can be reliably measured, and also produce a visualization of the antigenic relationships among many strains [11] (Smith et al., manuscript in preparation) (see [12–14] for related work). These *antigenic maps* reveal details of both the short- and long-term patterns of antigenic evolution, increase the granularity at which antigenic surveillance data can be examined, and thus increase the information available to the vaccine strain selection process. Antigenic maps can ameliorate some of the difficulties in comparing HI data from different laboratories and thus open the way for curation of antigenic data in a centralized database. Also, the finer-grain quantification of antigenic data opens up basic research in, among other areas, antigenic evolution, and the relationship between genetic mutations and their antigenic effects.

4. Predicting genetic evolution from the genetic data

Genetic data is more precise than antigenic data and there is a long history of detailed quantitative work on the genetic evolution of influenza. Most of this work is focuses on the hemagglutinin gene because of its primary role in antigenic drift [15,16]. Advances in the methods of evolutionary biology [17] and careful analyses of the intricacies of influenza data [18], have resulted in the identification of 18 positively selected codons in the hemagglutinin gene [19]. In a retrospective analysis, these codons predicted, from among circulating strains, the genetic variants that were the progenitors of future lineages in 9 of 11 seasons [20]. Monitoring changes in these 18 codons might help to predict the future evolution of the virus.

Mathematical techniques have been used to identify clusters in genetic data. In a retrospective analysis, a strain chosen from the most dominant genetic cluster of one season matched the WHO vaccine choice for the following season in 9 of 16 seasons [21] (see [40] for commentary). These clusters also enabled new analyses and quantification of the antigenic sites on the hemagglutin gene. Genetic clusters can also be visualized in *genetic maps* constructed using similar techniques to those used to construct antigenic maps. High-throughput sequencing of many strains will further enable genetic analyses, by providing not only more data, but also by reducing the sampling bias in the current data.

5. Modeling influenza epidemiology

The third consideration in vaccine strain selection is whether newly emerging strains are likely to cause widespread epidemics in the coming season. Current epidemiological models are not yet at the point of being able to help answer this question. However, influenza epidemics in closed settings, such as single outbreaks in nursing homes, can be accurately modeled [22] using modifications of classic *susceptible-infectious-recovered* techniques [23], and these models might provide guidance for the most effective use of antivirals in a pandemic situation [24]. However, to model interpandemic epidemiological patterns, models may will have to take into account antigenic drift of the virus, immunity after infection or vaccination with multiple related but different strains [25], vaccine coverage [41–43], seasonal variation [44], and spatio-temporal effects [45] (see [26] for a review). Adding antigenic dirft and cross-immunity into classical models greatly complicates the mathematics and although advances have been made, much work remains [27–29]. Advances in incorporating spatio-temporal information into epidemiological models has been successful in models of other pathogens including measles, whooping cough, and foot and mouth disease [30–34], and some progress has been made for influenza though much work remains [35,46–48].

A major reason for successes in epidemiological modeling of other pathogens is the availability of detailed spatio-temporal data. For influenza one needs not only spatio-temporal data but also virological data; thus, the proposed improvements coverage and harmonization of epidemiological surveillance, the linking of epidemiological and virological data, and the entering of these data in the influenza databases as proposed by the global agenda on influenza surveillance and control, will be of significant benefit to future epidemiological modeling work.

6. Optimizing vaccine strain selection for increased efficacy in repeat vaccinees

Vaccine strain selection is currently optimized to give a good match between the vaccine strains and the strains expected to circulate in the coming influenza season. This is the optimal strategy for first-time vaccines, but there is some evidence to suggest that modifications to this strategy might improve efficacy in repeat vaccinees. The efficacy of repeated vaccination has been difficult to determine definitively as different studies have come to different conclusions [36,37]. A meta-analysis of repeated vaccination studies showed that, on average, repeat vaccinees are as well protected as first-time vaccinees, but that vaccine efficacy in repeat vaccinees is more variable than efficacy in first-time vaccinees [38]. The *antigenic distance hypothesis* has been proposed to explain this variability, and a corollary of the hypothesis is that there is a trade-off in vaccine strain selection: while selecting a vaccine strain close to the expected epidemic strains increases vaccine efficacy, a vaccine strain too close to previous vaccine strains will, in some circumstances, be less effective in repeat vaccinees than in first-time vaccinees [25]. This reduced efficacy is likely due to elimination of vaccine antigen by pre-existing cross-reactive antibodies raised by prior influenza vaccination or infection. Influenza vaccination guidelines recommend annual revaccination for at-risk individuals; thus, a case can be made for optimizing vaccine strain selection for repeat vaccinees. Mathematical modeling of vaccine strain selection strategies, that take into account the antigenic distance hypothesis, suggests strategies that have the potential to increase vaccine efficacy in repeat vaccinees [39] (Smith et al., manuscript in preparation). These strategies are based on modifications of the current strategy and all give higher efficacy when there is a good estimate of the next drift variants; thus, the potential to optimize the vaccine choice for repeat vaccinees is dependent on the current methodology and any improvements that can be made to it, such as those described above and those proposed by the global agenda, will also increase the potential of these alternate strategies.

7. Summary

The wealth of data collected by the WHO global influenza surveillance network, and the subset of it stored in the influenza sequence database and influenza epidemiological databases, have enabled recent mathematical and computational advances in our basic understanding of the genetic and antigenic evolution of influenza. Coupled with an increased understanding of the determinants of the efficacy of repeated vaccination, these new methods increase the quantitative information available to the influenza surveillance and vaccine strain selection processes. Further advances in mathematical and computational biology, and its application to influenza, will be greatly facilitated by implementation of the proposed strengthening of virological and epidemiological surveillance by the WHO global agenda on influenza surveillance and control.

Acknowledgements

Many thanks to Robin Bush, Nancy Cox, Roṅ Fouchier, Alan Hay, Alan Lapedes, A.D. Osterhaus and Alan Perelson. This work was supported by Grant QLRT-2001-01034 from the European Union.

References

[1] Cox NJ, Brammer TL, Regnery HL. Influenza: global surveillance for epidemic and pandemic variants. Eur J Epidemiol 1994;10(4):467–70.

[2] Flahault A, Dias-Ferrao V, Chaberty P, Esteves K, Valleron AJ, Lavanchy D. FluNet as a tool for global monitoring of influenza on the web. J Am Med Assoc 1998;280(15):1330–2.

[3] Snacken R, Manuguerra JC, Taylor P. European influenza surveillance scheme on the Internet. Methods Inf Med 1998;37(3):266–70.

[4] Macken CA, Lu H, Goodman L, Boykin L. The value of a database in surveillance and vaccine selection. In: Osterhaus AD, Cox NJ, Hampson A, editors. Options for the control of influenza IV. Crete, Greece: Excerpta Medica; 2000. p. 103–6.

[5] Manuguerra JC. Surveillance of influenza: a pan European perspective. In: Osterhaus AD, Cox NJ, Hampson A, editors. Options for the control of influenza IV. Crete, Greece: Excerpta Medica; 2000. p. 131–7.

[6] Global agenda on influenza—adopted version I. Weekly Epidemiol Rec 2002;77(22):179–82.

[7] Adoption of global agenda on influenza II. Weekly Epidemiol Rec 2002;77(23):191–5.

[8] Cox NJ, Subbarao K. Global epidemiology of influenza: past and present. Annu Rev Med 2000;51:407–21.

[9] Both GW, Sleigh MJ, Cox NJ, Kendal AP. Antigenic drift in influenza virus H3 hemagglutinin from 1968 to 1980: multiple evolutionary pathways and sequential amino acid changes at key antigenic sites. J Virol 1983;48(1):52–60.

[10] Raymond FL, Caton AJ, Cox NJ, Kendal AP, Brownlee GG. The antigenicity and evolution of influenza H1 haemagglutinin, from 1950 to 1957 and 1977 to 1983: two pathways from one gene. Virology 1986;148(2):275–87.

[11] Lapedes A, Farber R. The geometry of shape space: application to influenza. J Theor Biol 2001;212(1):57–69.

[12] Weijers TF, Osterhaus AD, Beyer WE, et al. Analysis of antigenic relationships among influenza virus strains using a taxonomic cluster procedure. Comparison of three kinds of antibody preparations. J Virol Methods 1985;10(3):241–50.

[13] Beyer WE, Masurel N. Antigenic heterogeneity among influenza A(H3N2) field isolates during an outbreak in 1982/83, estimated by methods of numerical taxonomy. J Hyg (London) 1985;94(1):97–109.

[14] Perelson AS, Oster GF. Theoretical studies of clonal selection: minimal antibody repertoire size and reliability of self–non-self discrimination. J Theor Biol 1979;81(4):645–70.

[15] Fitch WM, Leiter JM, Li XQ, Palese P. Positive Darwinian evolution in human influenza A viruses. Proc Natl Acad Sci USA 1991;88(10):4270–4.

[16] Ina Y, Gojobori T. Statistical analysis of nucleotide sequences of the hemagglutinin gene of human influenza A viruses. Proc Natl Acad Sci USA 1994;91(18):8388–92.

[17] Fitch WM, Bush RM, Bender CA, Cox NJ. Long term trends in the evolution of H(3) HA1 human influenza type A. Proc Natl Acad Sci USA 1997;94(15):7712–8.

[18] Bush RM, Smith CB, Cox NJ, Fitch WM. Effects of passage history and sampling bias on phylogenetic reconstruction of human influenza A evolution. Proc Natl Acad Sci USA 2000;97(13):6974–80.

[19] Bush RM, Fitch WM, Bender CA, Cox NJ. Positive selection on the H3 hemagglutinin gene of human influenza virus A. Mol Biol E 1999;16(11):1457–65.

[20] Bush RM, Bender CA, Subbarao K, Cox NJ, Fitch WM. Predicting the evolution of human influenza A. Science 1999;286(5446):1921–5.

[21] Plotkin JB, Dushoff J, Levin SA. Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus. Proc Natl Acad Sci USA 2002;99(9):6263–8.

[22] Stilianakis NI, Perelson AS, Hayden FG. Emergence of drug resistance during an influenza epidemic: insights from a mathematical model. J Infect Dis 1998;177(4):863–73.

[23] Anderson RM, May RM. Infectious diseases of humans. Oxford: Oxford University Press; 1991.

[24] Stilianakis NI, Perelson AS, Hayden FG. Drug resistance and influenza pandemics. Lancet 2002;359(9320):1862–3.

[25] Smith DJ, Forrest S, Ackley DH, Perelson AS. Variable efficacy of repeated annual influenza vaccination. Proc Natl Acad Sci USA 1999;96(24):14001–6.

[26] Earn DJ, Dushoff J, Levin SA. Ecology and evolution of the flu. Trends Ecol Evol 2002;17(7):334–40.

[27] Andreasen V, Lin J, Levin SA. The dynamics of cocirculating influenza strains conferring partial cross-immunity. J Math Biol 1997;35(7):825–42.

[28] Lin J, Andreasen V, Levin SA. Dynamics of influenza A drift: the linear three-strain model. Math Biosci 1999;162(1/2):33–51.

[29] Gog JR, Grenfell BT. Dynamics and selection of many-strain pathogens. Proc Natl Acad Sci USA 2002;99(26):17209–14.

[30] Grenfell BT, Bjornstad ON, Kappey J. Travelling waves and spatial hierarchies in measles epidemics. Nature 2001;414(6865):716–23.

[31] Rohani P, Earn DJ, Grenfell BT. Opposite patterns of synchrony in sympatric disease metapopulations. Science 1999;286(5441):968–71.

[32] Ferguson NM, Donnelly CA, Anderson RM. The foot-and-mouth epidemic in Great Britain: pattern of spread and impact of interventions. Science 2001;292(5519):1155–60.

[33] Keeling MJ, Woolhouse ME, Shaw DJ, et al. Dynamics of the 2001 UK foot and mouth epidemic: stochastic dispersal in a heterogeneous landscape. Science 2001;294(5543):813–7.

[34] Keeling MJ, Woolhouse ME, May RM, Davies G, Grenfell BT. Modelling vaccination strategies against foot-and-mouth disease. Nature 2002;421(6919):136–42.

[35] Bonabeau E, Toubiana L, Flahault A. The geographical spread of influenza. Proc R Soc London B: Biol Sci 1998;265(1413):2421–5.

[36] Hoskins TW, Davies JR, Smith AJ, Miller CL, Allchin A. Assessment of inactivated influenza-A vaccine after three outbreaks of influenza A at Christ's Hospital. Lancet 1979;1(8106):33–5.

[37] Keitel WA, Cate TR, Couch RB, Huggins LL, Hess KR. Efficacy of repeated annual immunization with inactivated influenza virus vaccines over a 5 year period. Vaccine 1997;15(10):1114–22.

[38] Beyer WE, de Bruijn IA, Palache AM, Westendorp RG, Osterhaus AD. Protection against influenza after annually repeated vaccination: a meta-analysis of serologic and field studies. Arch Intern Med 1999;159(2):182–8.

[39] Smith DJ, Lapedes AS, Forrest S, et al. Modeling the effects of updating the influenza vaccine on the efficacy of repeated vaccination. In: Osterhaus AD, Cox NJ, Hampson A, editors. Options for the control of influenza IV. Crete, Greece: Excerpta Medica; 2000. p. 655–60.

[40] Fergnson NM, Anderson RM. Predicting evolutionary change in the influenza A virus. Nat Med 2002;8(6):562–3.

[41] Longini IM, Ackerman E, Elverback LR. An optimization model for influenza A epidemics. Math Biosci 1978;38:141–57.

[42] The Japanese experience with vaccinating schoolchildren against influenza. New Engl J Med 2001;344(12):889–96.

[43] Reichert TA. The Japanese program of vaccination of schoolchildren against influenza: implications for control of the disease. Semin Pediatr Infect Dis 2002;13(2):104–11.

[44] Reichert TA, Sharma A, Pardo S. A global pattern for influenza activity. In: Osterhaus AD, Cox NJ, Hampson A, editors. Options for the control of influenza IV. Crete, Greece: Excerpta Medica; 2000. p. 87–94.

[45] Grenfell BT, Kleczkowski A, Gilligan CA, Bolker BM. Spatial heterogeneity, nonlinear dynamics and chaos in infectious diseases. Stat Meth Med Res 1995;4:160–83.

[46] Rvachev LA, Longini IM. A mathematical model for the global spread of influenza. Math Biosci 1985;75:3–22.

[47] Flahault A, Letrait S, Blin P, Hazout S, Menares J, Valleron A-J. Modelling the 1985 influenza epidemic in France. Stat Med 1988;7:1147–55.

[48] Flahault A, Deguen S, Valleron A-J. A mathematical model for the European spread of influenza. Eur J Epidemiol 1994;10:471–4.