Predicting evolutionary change in the influenza A virus

Each year a new flu vaccine is produced, and judging which strains to target is a tricky business. A new study evaluating viral evolution suggests a more systematic approach to predicting next year's virus.

NEIL M. FERGUSON & ROY M. ANDERSON

In any discussion of the public health impact of influenza, the catastrophic death toll of global pandemics such as that of 'Spanish flu' in 1918 comes first to mind[1]. But over the longer term the smaller flu epidemics that occur each winter between major pandemics are responsible for greater overall mortality. Underlying the damage caused by both forms of influenza is an unusual pattern of virus evolution. Pandemics are caused by rare events in which existing human strains recombine with bird or swine viruses, producing a novel (and potentially lethal) new virus to which the human population has had no prior exposure and therefore has no immune protection. The more common annual epidemics are the result of more gradual evolutionary processes termed 'antigenic drift'—the generation of new strains from existing ones through mutation. 'Drift' gives a somewhat misleading impression of the processes shaping influenza evolution, however. Although mutation is random, new variants are subject to the intense selective pressure[2] of the immunity built up in the human population against their (antigenically similar) viral ancestors. Hence, although new viral lineages constantly emerge, this process of immunological selection generates fierce competition between strains, so on a timescale of a few years, only one survives and the others go extinct. The surviving dominant lineage is the one from which future variants arise. These dynamics give rise to the characteristic shape of the influenza A phylogenetic tree[3]—a dominant 'trunk' from which short branches emerge, describing a pattern of continuous genetic change through time, but approximately constant viral diversity at any instant (Fig 1*a*).

Although it is of considerable theoretical interest, this pattern also has important implications for public health. The limited diversity of the virus at any point in time makes feasible the use of vaccines targeted against only a few viral strains, but viral evolution necessitates annual updating of that vaccine. Given that mass production of new vaccine takes several months, vaccine efficacy depends on accurately predicting the strain likely to dominate transmission in the next year. A major public health goal is therefore to refine the prediction techniques used to select vaccine strains. This challenging task motivates a recent study by Plotkin *et al.*[4], which presents a novel approach to describing changes in the viral genome and, perhaps, to improving our ability to predict the unpredictable; namely, the direction of evolution.

Optimally, prediction methods should occur within a quantitative framework that melds the dynamics of both evolu-

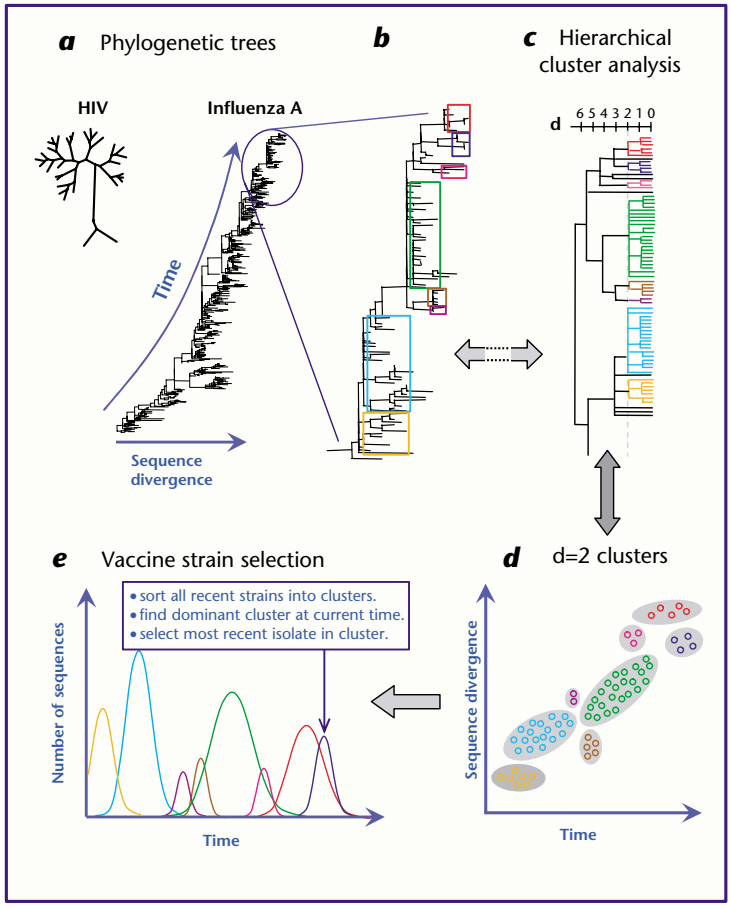

Fig. 1 Analyzing viral evolution. ***a***, Phylogenetic analysis reconstructs the evolutionary relationships among sequenced influenza viral isolates, the results of which are represented as a tree. External vertices represent individual isolates and internal nodes represent historical lineages from which the observed viral strains are inferred to have evolved. Unlike HIV, which is constantly growing in diversity, influenza A changes constantly but with limited diversity at any point in time—giving an unusual 'conifer' tree shape (thanks to R. Bush). ***b***, Looking in detail, clusters of closely related influenza sequences can be identified, where all the members of one cluster are further than some threshold distance on the tree from those in other clusters. ***c***, An alternative approach concentrates solely on how viral strains can be grouped in such clusters on the basis of sequence similarity. Hierarchical cluster analysis examines how cluster size changes for different values of the threshold distance ('*d*'), the minimum genetic distance (amino acid or base-pair differences between two sequences) separating all sequences of one cluster from those of another. Internal nodes in the resulting tree do not represent ancestral sequences—instead, vertical lines define clusters at a particular value of *d*. At *d*=0, all samples are their own cluster, at *d* = 1, closely related samples group into small clusters, whereas at *d* = 2, clusters include increasingly large numbers of related strains. Color is used to represent how clusters identified by the 2 types of analysis are likely to be similar (though not identical). ***d***, Plotkin *et al.* examine the sequence relatedness of isolates defining *d* = 2 clusters and how they are distributed in time (small circles represent strains). Any cluster dominates the viral population only for a few years, being displaced by a new cluster of strains that are significantly divergent at key epitopes. ***e***, Assuming the viral strain likely to cause most transmission next year will arise from the currently dominant cluster, Plotkin *et al.* propose the vaccine selection algorithm shown.

tionary and transmission processes. The growing volume of immunological and sequence data can help in this goal. The study by Plotkin *et al.*[4] focuses on these data by interpreting the variability present in viral isolates recorded in the Los Alamos National Laboratory Influenza Sequence Database (http://www.flu.lanl.gov/). This database is the result of a global surveillance effort targeted at detecting new strains. Such global surveillance typically uses immunological screens to preselect isolates for antigenic novelty before sequencing—resulting in a set of sequences that are far from a random sample of circulating viral types[5]. To analyze these complex data, Plotkin *et al.* used statistical cluster analysis methods[6] rather than the more usual phylogenetic models. There are some common elements to the two approaches (Fig 1*b–d*), but also some key conceptual distinctions. Phylogenetic methods try to infer the ancestral sequences (sometimes long extinct) from which all the sequences examined emerged, whereas cluster analysis solely concentrates on ordering samples by some criterion of similarity to give insight into existing patterns of diversity.

Plotkin *et al.* classify viral isolates into non-overlapping clusters based on the amino-acid sequence of HA1, the immunogenic part of the surface hemagglutinin (HA) protein. They then propose that these clusters, or 'swarms', better characterize the scale at which influenza evolution occurs than individual strains—given that strains within the same swarm will be subject to very similar selective pressures from host immunity. Insight into viral evolution is then gained by analyzing how clusters change through time and examining the differences among sequences within and between clusters. This approach reveals a pattern of emergence of clusters of closely related strains, with a few clusters growing to dominate the viral population but then being replaced within two to five years. Perhaps most interestingly, by examining changes at key epitopic sites in HA1 alone, the authors suggest each transition from one dominant cluster to another is associated with substantial change at a single epitope, with the affected epitope varying from transition to transition.

For vaccine selection (Fig 1*e*), it is the shorter-term evolutionary trends that are most relevant. The analysis of Plotkin *et al.* predicts that new strains are more likely to arise from the cluster of largest size (that is, the one containing the largest number of isolates collected over the previous year). Thus, focusing on that dominant cluster can improve predictions of the next season's influenza strain.

These are interesting results, but clearly more work is required before their techniques might replace current methods used in the annual process of vaccine strain designation by the World Health Organization. A number of key methodological issues also suggest caution at present. First, should the optimal measure of evolutionary distance between isolates be based on nucleotides, amino acids, stereochemical properties or immunogenic properties? Plotkin *et al.* use the number of HA1 amino-acid differences between two strains as a more phenotypically correlated description of isolate similarity than nucleotides (which give the greatest evolutionary information for phylogenetic analyses), but recognize the need for more research to develop predictive metrics capable of inferring antigenic properties from genotype data.

Second, although Plotkin *et al.* highlight the difficulties caused for phylogenetic analyses by the biased sampling scheme used to collect influenza sequence data[5], the methods they adopt are also vulnerable to such biases. If viral evolution had been more intensively observed (for example, using 5,000 sequence samples, rather than 500), sampling all single amino-acid changes, then single-linkage cluster analysis would have grouped all sequences into the same cluster—dramatically reducing insight into evolutionary processes. A consequent concern is that the classification of strains into clusters is at least in part a result of phenotypically biased sampling or random temporal variation in sequence collection, as well as any underlying discrete directional changes in viral evolution. Non-random sampling also undermines assumptions that cluster size reflects the frequency of member strains in the overall viral population. These issues need to be addressed through use of cluster analysis methods that are more robust to sampling biases[6,7], and by placing greater emphasis on representative sampling.

Understanding how cluster analysis relates to phylogenetic studies (Fig 1) is also clearly important to a deeper insight into the dynamics of influenza evolution. The power of phylogenetic approaches is typified by past work identifying the HA1 gene to be under intense positive selection[2]. These studies demonstrated an association between mutation at 18 key codons and lineage survival that might form the basis of a vaccine-strain prediction algorithm[3]. Phylogenetic techniques (which are also a type of cluster analysis) have the advantage of being based on statistically robust models of the random mutational process governing nucleotide substitutions. Testing evolutionary hypotheses, sensitivity analysis and sample-size power calculations can all be undertaken more easily for such models[5] than for the deterministic algorithms underlying simple cluster-analysis methods.

The novelty of the work of Plotkin *et al.* therefore lies less in the specifics of their analysis method, but more in the conceptual focus on influenza evolution as a dynamic process. The explicit representation of the development of 'swarms' of closely related viral strains over time gives new insight into processes of strain and quasi-species emergence and extinction. Despite the statistical concerns, by focusing less on inferring the microscopic structure of evolution, cluster methods do offer some potential advantages over conventional phylogenetic analyses—such as their ability to integrate both diversity and frequency data into approaches for predicting short-term viral evolution. Such ever more sophisticated descriptive analyses bring us closer to uncovering what is still missing from our understanding of influenza epidemiology and evolution: a robust theoretical framework capable of explaining the key population processes that shape the complex spatiotemporal patterns observed in epidemiological and sequence data.

1. Glezen, W.P. Emerging infections: pandemic influenza. *Epidemiol Rev.* **18**, 64–76 (1996).
2. Fitch, W.M., Leiter, J.M., Li, X.Q. & Palese, P. Positive Darwinian evolution in human influenza A viruses. *Proc. Natl. Acad. Sci. USA* **88**, 4270–4270 (1991).
3. Bush, R.M., Bender, C.A., Subbarao, K., Cox, N.J. & Fitch, W.M. Predicting the evolution of human influenza A. *Science* **286**, 1921–1925 (1999).
4. Plotkin, J.B., Dushoff, J. & Levin, S.A. Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus. *Proc. Natl. Acad. Sci. USA* **99**, 6263–6268 (2002).
5. Bush, R.M., Smith, C.B., Cox, N.J. & Fitch, W.M. Effects of passage history and sampling bias on phylogenetic reconstruction of human influenza A evolution. *Proc. Natl. Acad. Sci. USA* **97**, 6974–6980 (2000).
6. Gordon, A.D. *Classification.* (Chapman & Hall/CRC, London, 1999).
7. Everitt, B.S., Landau, S. & Leese, M. *Cluster Analysis.* (Edward Arnold, London, 2001).

Department of Infectious Disease Epidemiology
Faculty of Medicine
Imperial College of Science, Technology and Medicine
London University, London, UK
Email: neil.ferguson@ic.ac.uk