Molecular epidemiology and evolution of emerging infectious diseases

Edward C Holmes

The Wellcome Trust Centre for the Epidemiology of Infectious Disease, Department of Zoology, University of Oxford, Oxford, UK

Molecular epidemiology is an emerging science. The development of new and rapid protocols to isolate and identify pathogens, coupled with the sophisticated phylogenetic analysis of their gene sequences, is providing a new and fascinating insight into the biology, origin and spread of infectious diseases. In this essay, I describe some of the ways in which the techniques of modern molecular biology and evolution have equipped us to face the challenge of these new infections.

It may come as a surprise to learn that the worst epidemic of infectious disease on record only occurred 80 years ago. Yet the 1918–1919 outbreak of ‘Spanish’ influenza killed at least 20 million people worldwide and infected a staggering 28% of the US population. Knowing the factors which allowed the emergence of such a virulent virus would obviously be of considerable benefit to medical science. Until recently there was no way we could recover such information, but, with the advent of modern molecular biology, the picture has changed dramatically: gene sequences from this ‘fossil’ virus have been now recovered from stored tissues and may yet unlock the secrets of this epidemic.

The case of ‘Spanish’ influenza illustrates how profoundly modern molecular biology, and especially the availability of gene sequence data, can change our approach to the study of infectious diseases. The invention of the polymerase chain reaction (PCR) and DNA sequencing, along with the development of more accurate methods for recovering epidemiological information from sequence data, has enabled us to build up a more detailed picture of pathogens than has ever been achieved before. We are clearly entering a new age. Crucially, it is not always necessary to have fossil gene sequences on hand to learn about the past history of diseases: those from contemporary organisms contain an archaeological record of their evolutionary history in the form of the mutations that have accumulated between them over time. This information can be revealed by reconstructing the mutational pathways...
which link gene sequences. These genealogies of gene sequences, more usually referred to as phylogenetic trees, are analogous to the pedigrees used in clinical genetics, and simply specify how closely related sequences are to each other. In the context of emerging diseases, phylogenetic trees of pathogen genomes have allowed us to determine from where infections originated and then to track their spread through populations. By looking at the structure of these trees more closely we can also reveal what forces have shaped pathogen evolution, including those which may have initiated emergence. Finally, the comparison of gene sequences from pathogens which differ in biological properties, such as virulence, tropism and transmissability, allows us to pinpoint the very nucleotides which control these life-style characteristics. Information of this kind will greatly improve the development of effective vaccines and drugs. The more we know about the origin and structure of genetic variation in pathogen populations, the better we will be able to control them.

In this paper I will review the ways in which modern molecular and evolutionary techniques, especially the phylogenetic analysis of gene sequence data, have greatly enhanced our study of emergent pathogens. As my own research interests lie with viral populations, I shall focus on viruses, although the principles outlined could equally well be applied to other infectious agents.

Pathogen isolation and identification

One of the most notable benefits of the expanding armoury of molecular biology is the ability to isolate and identify quickly new pathogens. This is of crucial importance if we are to respond effectively to the threat posed by emergent infections. In the first instance, newly developed molecular techniques have enabled the isolation of pathogens which had previously proven unculturable. An elegant case concerns the aetiology of Kaposi’s sarcoma (KS), a vascular neoplasm often associated with AIDS in homosexual men. Using a PCR technique known as ‘representational difference analysis’, in which pathogen genomes present in diseased tissues, yet absent from healthy tissues, are preferentially isolated, DNA sequences similar to those of herpesviruses were isolated from a KS lesion. This implies that Kaposi’s sarcoma is, at least in part, caused by a herpesvirus, now christened KSHV (or human herpesvirus 8 – HHV8). Similar techniques also led to the recent discovery of hepatitis G virus (HGV), also known as GB virus C, an RNA virus which has a high prevalence in some populations, although an uncertain role in liver disease. In the case of HGV, viral isolation
Disease molecular epidemiology proved more difficult because its RNA genome made it harder to distinguish from cellular mRNAs, so that the candidate agent had to be passed initially through primates and then subjected to representational difference analysis in cell-free extracts of primate plasma.

Once isolated, it is necessary to identify the pathogen. This can be done most precisely by the comparison of gene sequences. For emergent pathogens it may also be important to do this quickly. This is spectacularly demonstrated by the rapid response to the highly publicised outbreak of Ebola virus in Kitwit, Zaire in 1995 (Khan et al., this issue): glycoprotein sequences from the Kitwit strains were obtained within 48 hours of the virus arriving at the CDC in Atlanta. With gene sequence data available, initial identification can then be achieved by comparing overall sequence similarity between the reference sequences and those already present in the sequence databases (which, for example, led to the recognition that KSHV was a gamma herpesvirus) or, more precisely, through phylogenetic analysis. In the case of Ebola, the phylogenies constructed showed that the Kitwit strain differed only subtly from the Ebola strains isolated from previous outbreaks. Other notable triumphs in the field of viral isolation and identification were the discovery that a hantavirus ('Sin Nombre virus') circulating in deer mice was the cause of a high mortality respiratory illness in southwestern US (Young et al., this issue), and the rapid identification of a new and highly pathogenic strain of influenza A virus, H5N1, isolated in Hong Kong and which had unusually emerged directly from a chicken reservoir population.

The phylogenetics of emergent diseases

The uses of phylogenetic trees described thus far have primarily been of a taxonomic nature, showing how closely related the pathogen in question is to other microorganisms. Taxonomic studies of emergent pathogens are now commonplace (see, for example, the extensive work on the arenaviruses, some of which cause haemorrhagic fevers in humans) and will not be discussed further here.

At another level, molecular phylogenies are also being used to determine the genetic structure of pathogen populations, and to elucidate the evolutionary forces which have produced this structure. For example, on a worldwide basis, sequences from HIV-1 form clusters, or 'subtypes', on phylogenetic trees (Weber, this issue). Some of these subtypes, and most notably subtype A from central and west Africa, contain substantial genetic diversity, suggesting that they have been resident in their host populations for some time. Other viral subtypes,
particularly subtype B from much of the world outside of Africa, and frequently in Europe, Japan and North America, as well as subtype E from Southeast Asia (most notably Thailand), possess less genetic diversity (form a tighter cluster of branches on the phylogenetic tree), suggesting that they have emerged only recently and then spread quickly (Fig. 1). A similar situation is observed in hepatitis C virus (HCV), another important blood-borne pathogen of humans and a major cause of serious liver disease. Phylogenetic studies of genetic variation in HCV reveal that infected populations in Africa and Asia are exposed to viral genotypes (the generic term for phylogenetically distinct clusters of sequences) richer in genetic variation than those making their way through developed countries. This implies that the virus has only recently emerged in the developed world, where most people are infected through needle sharing and contaminated blood products.

Phylogenetic trees can be even more informative when the branches of the phylogeny are correlated with other biological variables, such as geographical location of the pathogen strains, their host organism, time of sampling, virulence, or mode of transmission (Fig. 1). Analyses of this sort performed on a multitude of viruses have revealed that geographical structuring is common in many populations, with particular viral genotypes often restricted to specific parts of the world. This is indicative of restricted viral movement ('gene flow') between populations. As an example, hepatitis B virus (HBV), one of the most common infectious diseases of humans, has six genotypes worldwide. Genotype A viruses are most commonly found in northern and central Europe and sub-Saharan Africa, while genotypes B and C dominate in Asia. Genotype D is found mainly in the Mediterranean region, while the remaining genotypes, E and F, are indigenous to Africa and the Americas, respectively. Other viral infections are more structured in time than space, so that strains continually replace each other as evolution proceeds. Influenza A virus represents just such a case. Phylogenetic trees of this virus are characterised by a single dominant lineage, the main 'trunk' of the tree, which connects successive epidemics, and short side branches that only persist for a single epidemic. This tree structure suggests an evolutionary process in which mutations in the haemagglutinin and neuraminidase proteins that allow the virus to evade immune responses are continually fixed (because they are selectively advantageous) on the central trunk of the tree, so taking the virus from one epidemic to the next ('antigenic drift'), whereas the side branches represent those strains eventually eliminated by the immune system.

Other viral populations are structured by characteristics such as tropism and even virulence. In these cases, phylogenetic trees may be of direct clinical importance. This can also be illustrated by HBV, where
Disease molecular epidemiology

Fig. 1 The anatomy of a phylogenetic tree. This tree describes the evolutionary history of the gag (capsid) gene of HIV-1 (1401 bp) from 56 viral isolates representing subtypes A (from central and west Africa) and B (currently found in much of the non-African world, most notably Europe, North America and Japan). The tree was constructed with the popular neighbor-joining method. The tree is labelled to highlight some of the things which can be inferred from it, such as the high rate of evolutionary change along some (long) branches, the location of mutations which define clusters of sequences, such as the subtypes (i.e. those mutations which would give each subtype their characteristic features) and the accelerated rate of branching which reflects the rapid spread of HIV-1 in the west during the early 1980s. An exemplar cluster of sequences is also pinpointed, along with some of the biological factors which could have given rise to it. Relevant for viruses other than HIV-1. All horizontal branch lengths are drawn to reflect the actual amount of evolutionary change along them.

Long branch = high rate of evolutionary change

HIV-1 Subtype A

What do clusters of sequences correspond to?
- geographical location
- time of sampling
- tropism
- virulence (clinical outcome)
- host*
- vector*

Mutations on this lineage define this subtype

Branches arising in a very short time period suggest a rapid rate of population growth

HIV-1 Subtype B

10% sequence divergence

Evidence of a strain basis to fulminant hepatitis has recently been presented: many cases of this high mortality acute disease are associated with clusters of viruses, most often from genotype D (Carman, this issue). Likewise, in HCV, it is well documented that some...
viral genotypes respond less well to interferon treatment than others\textsuperscript{16}. Knowing the viral genotype of a patient may, therefore, influence clinical management. Finally, phylogenetic analysis of another important human pathogen, human papillomavirus (HPV), led to the discovery of five viral ‘supergroups’, which generally correspond to different disease syndromes\textsuperscript{17}. For example, genital and mucosal HPVs are found in supergroup A, whilst those which cause the rare skin disorder, epidermodysplasia verruciformes, come from supergroup B.

In all these cases it is now possible to hunt for the mutations responsible for the different phenotypes, which will be located on the branches leading to each phenotypic group (Fig. 1). Determining whether the groupings observed in pathogen phylogenies correlate with important biological characteristics, and pinpointing the base changes responsible for these characteristics, will be a major component of the molecular epidemiology of the future.

Molecular data may even allow us to date, in real time, when the events leading to pathogen emergence took place. This is possible if mutations between genomes accumulate at an approximately constant rate over time, so giving rise to a ‘molecular clock’ of evolutionary change. If the tick speed of this clock can be calculated, for example by calculating the amount of sequence evolution between samples taken at different and known time points, then the divergence times of other branches on the tree can be estimated. For example, when we combine estimates of the mutation rate of HTV-1 and HCV (calibrated in the manner described above) with the phylogenetic trees of these viruses we find that both started to spread quickly within the last 50 years\textsuperscript{18}. From this we can infer that behavioural changes in some human populations, such as the more widespread use of blood products and increased needle sharing by injecting drug users, were the key events in allowing these viruses to spread through the developed world. One of the most exciting applications of molecular data in the future will be to place the evolution of many other human pathogens within a chronological framework.

Despite its potential, it is also important to stress the limitations of phylogenetic analysis. First, the evolutionary tree is in reality an assumption about how evolution takes place, not a truism. To be more specific, if recombination between pathogens is widespread, as appears to the case in many bacteria\textsuperscript{19} and some viruses\textsuperscript{20}, then the relationships between sequences should not be represented as a tree with steadily bifurcating lineages, but rather as a ‘network’ which depicts all the possible pathways by which they can be connected\textsuperscript{21}. Second, phylogenetic analysis should be based on a representative (unbiased) sample of sequences. Not only does this improve the analysis by diluting the evolutionary ‘noise’ in the data (such as multiple mutations at single nucleotide sites) which confound recovery of the true phylogeny\textsuperscript{22}, but the
correlation between branch of the tree and biological characteristic will only be possible if sequences are correctly sampled with respect to such variables as geography, age, sex and clinical status. Proper sampling is in many ways the key to good molecular epidemiology.

Finally, phylogenies are not the only aspect of gene sequence analysis, although they do provide the essential frame of reference. Other useful analytical techniques involve a comparison the relative numbers of nonsynonymous (amino acid changing) and synonymous (non-amino acid changing) substitutions in different gene regions, and the ‘sliding window’ analysis of genetic variability along sequences, both of which provide information about spatial variation in selection and mutation pressure.

**Case study: the emergence of dengue**

The power of many of the molecular tools described above can be illustrated with respect to the emergence of dengue virus, the agent of disease syndromes ranging from febrile infection (‘dengue fever’) to haemorrhagic fever in large parts of Africa, Asia and Latin America, and for which no vaccine currently exists.

Dengue is a single-stranded RNA virus, about 11 kb in length, that preferentially replicates in the monocyte cells of higher primates, most notably humans, and which is principally transmitted by the urban adapted mosquito *Aedes aegypti*. The genetic diversity in the virus is such that it exists as four antigenically distinct serotypes (denoted DEN-1 to DEN-4). Although difficult to measure accurately, at least 100 million cases of dengue occur year, with more than 500 000 classified as the most serious disease forms – dengue haemorrhagic fever (DHF) and dengue shock syndrome (DSS), the aetiology of which remain uncertain. The increasing prevalence of dengue, especially in Latin America (coincident with the re-invasion of that continent by *A. aegypti*), give it all the characteristics of an emerging virus. But where did it come from and what factors have facilitated its spread? Molecular phylogenies are helping to provide answers.

Dengue is a flavivirus, a genus of RNA viruses which contain two other important human pathogens, Japanese encephalitis virus and yellow fever virus. Comparison of sequences from the E (envelope) gene from a variety of flaviviruses reveal the dengue is most closely related to the other mosquito-borne flaviviruses, and distinct from those transmitted by ticks. Applying a molecular clock to this data set reveals that the flaviviruses as a whole have only arisen within the last 10 000 years, and that the divergence of dengue from the other flaviviruses took place about 2000 years ago (Fig. 2). So, whilst dengue is considered emergent, the virus itself has had a much longer evolutionary history.
Resurgent/emergent infectious diseases

Massive increase in viral genetic diversity and population size (rapid rate of branching)
~ 200 years ago

Origin of dengue virus
~ 2000 years ago

10% sequence divergence

Fig. 2 A phylogenetic tree for dengue virus. This tree was reconstructed using E (envelope) gene sequences (1485 bp) from 108 worldwide viral isolates. The four serotypes of the virus are marked, along with their point of origin, estimated to be at around 2000 years ago. Also shown is the point in time (approximately 200 years ago) when the virus started to spread quickly, as reflected in the rapid increase in the number of dengue lineages in all four serotypes (only shown for DEN-2 for the sake of simplicity). A cluster of sequences which corresponds to a genotype of DEN-2 is also denoted (although this tree does not contain all the dengue genotypes known to date).
But why has dengue become so prevalent today? The key observation here is that each of the four serotypes contains substantial genetic variation in the form of distinct genotypes: four in DEN-2 and DEN-3 and two in DEN-1 and DEN-4. As our sampling of dengue virus improves, particularly in Africa, it is likely that we will uncover even more genetic variation. More significantly, these genotypes have different geographical distributions which provide important clues about the emergence and spread of dengue. This is best documented in DEN-2, the serotype most often associated with severe disease, where some of the genotypes have huge geographical distributions, ranging from Asia, across the Pacific to Latin America. Such a geographical range implies that dengue virus frequently moves between populations, with epidemics perhaps due to the importation of new strains into populations with no existing immunity. Also of interest is the observation that one of the genotypes of DEN-2 only contains sylvatic (‘jungle’) viruses isolated from monkeys in west Africa. The separation of the sylvatic from the human strains on phylogenetic trees suggests that the former also have little role to play in dengue epidemics.

A second and even more worrying aspect to genetic diversity in dengue virus is that it appears to be on the increase. This can also be inferred from the phylogenetic tree: for much of its evolutionary history dengue appears to have possessed very few lineages, but these have dramatically increased in number in the recent past (Fig. 2). This in itself should be cause for concern as it means that there is a greater possibility for viruses to evolve new phenotypic characteristics, such as increased transmissibility, increased virulence, faster replication in mosquito vectors or even the adoption of new vectors. The increasing number of branches on the phylogeny also means that the population size of the virus is growing. The reason for such a rapid population growth seems to be tied to the fact that dengue virus is now dependent only on humans for its transmission, so that the more susceptible human hosts available, the greater the likelihood that the virus will be able to sustain its spread. The molecular clock of dengue supports this: the rapid increase in viral genetic diversity began about 200 years ago, when the size of urban human populations began to rise and humans also became more mobile. Large-scale dengue epidemics are also first recorded at this time. So dengue seems to have broken free from its simian past and now replicates solely in vectors and humans, and when more humans became available and started to move more freely around the world, the virus began to spread more quickly and so increase its population size. Given the current growth of human populations, particularly in urban areas, this represents an important lesson from the past.

A final, more vexed question, is whether some strains of dengue are more associated with severe clinical outcomes than others. In other
Resurgent/emergent infectious diseases

words, do DHF and DSS have a strain basis? Whilst it is clear that the reason some people succumb to these serious forms of dengue infection may well be differences in host genetic susceptibility\(^26\) and antibody-dependent enhancement, whereby sequential infection with multiple serotypes increases the number of infected monocytes and hence the extent of viral replication\(^27,28\), it is also possible that strains of dengue differ in virulence. For example, an outbreak of DHF which swept through the Caribbean region (most notably Cuba) in the early 1980s, is associated with a strain of DEN-2 which had previously been circulating in Southeast Asia where serious disease is especially common. This may represent a highly virulent strain of dengue virus\(^29\) and, if so, the mutations responsible for its heightened virulence can now be tracked down.

Gene sequence data have documented an unprecedented, if disturbing, picture of the emergence of dengue. Most striking is that these sequences not only contain information about when dengue first appeared and how it has spread since this time, but also about the rate at which its population size has grown. As gene sequencing technology develops further, and methods of sequence analysis become even more sophisticated, our ability to track the spread of new pathogens like dengue will continue to improve, as will the prospects for their eventual eradication.

Acknowledgements

I thank The Royal Society, The Wellcome Trust and BBSRC for financial support.

References

8 Cohen J. The flu pandemic that might have been. *Science* 1997, 277: 1600–1


