1

Introduction and guide

ADRIAN GIBBS, CHARLES CALISHER
AND FERNANDO GARCÍA-ARENAL

This chapter briefly outlines the lines of evidence of virus evolution that are discussed in this book; hopefully it augments the information given in the Table of Contents by indicating where those lines anastomose.

Evolution is the process whereby the population of an organism changes genetically over a period of time. It occurs when genetic variation within a population, combined with selection from among the mixture of genotypes, results in change. Viruses, especially those with RNA genomes, sometimes evolve very rapidly and so their evolution can be studied as it occurs. By contrast, all cellular organisms evolve more slowly and so their evolution is deduced by comparing extant forms and, for some, the fossils of their ancestors. No fossils of viruses are known.

Pre-molecular evidence of virus evolution

It has been realized for a long time that pathogens evolve. The fact that new epidemics of disease appear (Chapter 3), and that some viral diseases, such as measles and smallpox, induce life-long immunity in those individuals that survive, whereas others, such as the common cold or influenza, do not, all indicate that some viruses change. Of course, such differences influence choice of control strategies because stable viruses are amenable to control by vaccination, whereas rapidly changing viruses are not. Another type of evidence of virus evolution is the fact that avirulent strains of some viruses can be selected for use as vaccines, usually by growing them in unusual hosts, or under unusual conditions.

One of the first deliberate attempts to study natural virus evolution was the classic study of myxoma leporipoxvirus when it was liberated in Australia to attempt to control European rabbits (Chapter 2).
Molecular evidence of virus evolution

Since the 1970s the development of methods for sequencing genes and determining the structures of proteins has transformed the study of evolution, particularly that of viruses. Although viruses had previously been classified into groups using various characters, such as their host range, virion characteristics and the serological specificity of their proteins, there was no evidence that the resulting groupings reflected evolutionary relationships, nor whether any of the groups were phylogenetically related. However, sequences of viral genes revealed for the first time which genes are shared and homologous, and how closely related they are. This is because there is redundant information at all levels of biological systems; 61 nucleotide triplets encode 20 amino acids, and innumerable different amino acid combinations can give the same fold in a protein. As a consequence, genes encode information essential for life mixed with clues of their evolutionary history. Thus, by comparing the nucleotide sequences of genes, or the sequences or structure of proteins, from different lineages, one can attempt to separate the different components of change (Chapter 36). In this way, the mode and tempo of evolution of viral populations, species, genera and even higher taxa can be deduced. It has been found, most surprisingly, that many viral genes are widely distributed among viruses that seemed quite unrelated (Chapters 4, 5 and 6), some are related to those of their hosts and may have very ancient origins, and there is clear evidence of genetic recombination in all viral genomes (Chapter 9).

Evolution of viral species

The need to define and name viral species caused much controversy in the past. Viruses reproduce asexually and, in early studies, only a few gave evidence of genetic recombination. Thus it was unclear how viruses could maintain apparently stable species because it was thought that sexual reproduction, involving gene shuffling among individuals constituting a ‘gene pool’, was the cohesive force that maintained the genetic integrity of species. None the less some viruses are known to be genetically stable, notably some of the viral strains used as vaccines. This paradox was resolved when the genomic sequence of QB levivirus, a bacteriophage, was first determined. During this work, QB was passaged 55 times, and only a single variant nucleotide was found in the genomic sequence over a two-year period. However, when individual clones were
prepared from inocula saved during the work, it was found that their
genomic nucleotide sequences differed by an average of 1.6 sites from
that of the unselected stock. The variant nucleotides were apparently
distributed at random in the genome, and thus the stock gave the
appearance of being a single genomic sequence, the ‘master copy’.
When variant clones were mixed with the stock population and passaged,
they rapidly disappeared, and the ‘master copy’ became dominant. Thus
stabilizing (‘purifying’) selection operating on the variants of a genome
that arise from a common ancestor usually favours a single master copy
and close variants of it, resulting in a stable ‘quasispecies’ or ‘mutant
spectrum’ (Chapter 13).

It seems that most variant genomes produced by point mutation are
less fit than their parents, as in the Q8 study, but not all are, especially
when selection pressures change and these allow the evolution of
populations (Chapters 14 and 15). Some variants may be favoured by
extant conditions; for example, it was a single point mutation in
the haemagglutinin gene of a wild fowl plague orthomyxovirus in 1982
that resulted in a changed glycosylation site and caused an epidemic
of virulent disease in American poultry farms resulting in the death
or slaughter of 18 million birds (Chapter 34). Similar variants allow
antigenic drift, and this enhances the survival of many viruses of
vertebrates.

Patterns of relatedness in viral populations

It is now becoming evident that all viral populations, except those that
have just been cloned, vary to a greater or lesser extent (Chapters 13,
35 and 36). Individual isolates in a population can be compared, their
relationships inferred and represented, usually as a network or ‘tree’.
Such a tree can give useful information on the origin and the mode
and extent of selection affecting that population. In many viral populations
the between-isolate relationships fit a quasi-species normal distribution,
and there is no obvious correlation between those relationships and the
time or place that the isolates were obtained. Such populations are
probably stable quasi-species that have established a balance between
mutation and selection. This pattern has been found in populations
of, for example, influenza A orthomyxovirus in wild bird populations
(Chapter 34), foot-and-mouth disease aphthovirus (Chapter 21) and
tobacco mild green mosaic tobamovirus (Chapter 23) and its satellite
virus (Chapter 26). Very few of these populations have been proven
to be quasi-species by competition experiments, and it is possible that some of the variation in them is maintained by host selection.

Another population pattern is that of a ‘shrub-like’ tree with the most recent isolates furthest from the centre of the shrub. In such populations, many independent lineages have evolved from a common ancestor, and there has, as yet, been little or no selection against any lineage. One of the first studies reporting this pattern was of an epidemic of enterovirus 70 (Miyamura et al., 1986); the current epidemic of HIV in the human population seems to be of this type (Chapter 30). By contrast, influenza A orthomyxovirus epidemics in human populations give trees with a single narrow dominant apex and few short branches, which indicates that, during the epidemic, there has been selection against all but one lineage (Chapter 34), probably the result of herd immunity selecting for the line that is most antigenically novel. A similar pattern is also found in successive isolates obtained from individual animals persistently infected with lentiviruses (Chapters 11, 12).

**Sequence variation**

Mutations (Chapter 8) apparently occur in random positions in viral genomes, though the different types of mutation, or their survival, may not be random. This is shown, for example, in biased nucleotide usage in some viral lineages, but not others; evidence that the bias is genetically determined by the virus. In ORFs, third codon positions are usually redundant and vary most frequently, transitions (changes from purine to purine or from pyrimidine to pyrimidine) occur more frequently than transversions (changes from purine to pyrimidine or vice versa), and insertions/deletions least frequently. There are large differences in the rate of change of different parts of each gene caused by differences in the ease with which their function is conserved despite changes. Mutations appear at a rate of about $10^{-3}$/ nucleotide position/replication cycle in RNA genomes, and this allows rapid nucleotide sequence changes, up to 1% /year, in some viral populations, such as HIV (Chapter 30) or influenza A virus (Chapter 34) in the human population. It also permits speculation on the timing of recent changes (Chapter 33). By contrast, DNA genomes change at least a million times more slowly, and there is evidence that the observed mutation rate per genome per replication is virtually constant (.0033), not the mutation rate per nucleotide, which is fastest in the smallest genomes (Drake, 1991, 1993).

Few studies, however, have determined which changes are adaptive,
and which are mere evolutionary noise. Clearest evidence comes from viruses with virion proteins of known structure, like the influenzas and FMDV, where the changes can be shown to have occurred in antigenically important surface regions, or to have been avoided in receptor sites. However, little is known of the molecular basis of host range and of the changes that enable viruses to acquire and adapt to new hosts; a feature of great evolutionary significance.

**Evolution of viral genera or groups**

As mentioned above, most of the taxonomic groupings of viral species, that were defined long before the genomic sequences of the viruses were known, were confirmed by genomic sequence analysis. The relatedness of species of each viral genus usually vary greatly, perhaps because the criteria for defining genera are artificial and because new species can probably arise at any time.

Species of the same viral genus often have different natural hosts, though their experimental host range may be very much wider (Selling, Allison & Kaesberg, 1990; Ball, Amann & Garrett, 1992). This implies that adapting to a new natural host and speciation are linked; however, the genomic sequences of the viruses give few, if any, clues to the molecular basis of host adaptions. In some viral genera, for example, the papillomaviruses (Chapter 31), the poxviruses (Chapter 18) and the tobamoviruses (Chapter 23), the molecular taxonomy of the viral species is mostly congruent with the taxonomy of the natural hosts. This suggests that the origins of such viral groups and their hosts pre-date their present divergences, also that the viruses and hosts have mostly co-evolved and co-speciated. However, incongruous host affiliations in such groupings indicate that taxonomically unrelated hosts can be acquired occasionally. For example, all the tobamoviruses found naturally in solanaceous plants form a single close-knit group, but that group also includes a virus of orchids and another of cacti.

Past studies of viral host ranges (Bald & Tinsley, 1967) have been hampered by poor knowledge of the real evolutionary relationships of viral hosts, especially at the higher taxonomic levels. However, molecular studies are now producing a flood of information, especially for ‘prokaryotes’ (Chapter 16) and higher plants (Chapter 17), and, more slowly, for animals (Graur, 1993; Marshall & Schultze, 1992). Soon it should be possible to search sensibly for correlations between
the taxonomies of hosts, viruses and their vectors and genes (Chapters 19, 20, 22, 25 and 32).

The species in each viral genus have genomes with unequivocally related sequences, but these differ in ways, and on a scale, that is greater than that found between different isolates of each viral species. First, there are sequence alterations that often produce amino acid differences, usually conservative, in the encoded proteins. Secondly, there are also length differences. These result from insertions, additions or deletions in individual genes, and also from the acquisition or loss of genes; for each gene, comparative taxonomic analysis may show which of these options is correct. Many are acquired by recombination from other genomes, both viral and cellular (Chapter 7). For example, recombination has clearly been common among the retroviruses (Chapter 27) and among the luteo-like viruses (Chapter 24), but it has also occurred, though less frequently, in other taxa. For example, it is clear that western equine encephalitis alphavirus (WEEV) originated by recombination (Chapter 33) indeed, recombination may be very common but the recombinants usually less fit than their parents.

Pseudo-recombination, the reassortment of the segments of multipartite genomes, is especially common among some viruses. It is important in the success of orthomyxoviruses in the human population (Chapter 34). The major epidemics of influenza A in the human population in 1957 (Asian ‘flu) and 1968 (Hong Kong ‘flu) were caused by strains of the virus that had acquired novel virion surface antigens. This probably resulted when genome segments re-assorted during mixed infection of a ‘bridging’ host, perhaps a pig, by bird and human isolates.

Although most recombinant viruses acquire their genes from viral parents, sometimes they acquire them from their hosts (Chapters 7 and 27). Best known are the oncoviruses acquired by some retroviruses to yield defective variants that require help from the parental virus but which rapidly transform susceptible cells into a cancerous state.

Comparative taxonomic analysis also shows that some genes have arisen de novo (Chapter 6). For example, the methyl transferase gene of tymoviruses and the virion protein gene of luteoviruses are shared with a large number of closely or distantly related viruses; however, the genes that overlap them are only found in tymo- and luteoviruses respectively, indicating that they arose de novo in the progenitors of these groups. Similarly, the overlapping spliced regulatory genes of lentiviruses can be shown to have arisen de novo, because whereas the associated env gene
is related to the env genes of all other retroviruses, only the lentiviruses have tat and rev genes.

Origins of viral genera or groups

Sequence comparisons show that the virion protein genes and many of the basic metabolic enzymes of viruses, such as those involved in nucleic acid and protein metabolism, have evolved from a limited number of parental genes and are shared by many genera (Chapter 4). For example, many of the viruses with RNA genomes have replicases, nucleotide binding or helicases and methyl transferases that probably evolved from single ancestral genes, and virion protein genes that evolved from, perhaps, three genes; Qβ levivirus has a virion protein with a structure that is a β-sheet with an α-helical clip, whereas all the others that have isometric virions about 30 nm in diameter have virion proteins that are eight-stranded anti-parallel β-barrels, and, by contrast, many, if not all, with rod-shaped or filamentous virions have four-stranded α-helical bundle virion proteins. Sequence similarities also link many of the genes of viruses with DNA genomes, and some of these genes are also unequivocally related to genes of their hosts (Chapters 5 and 18).

The viral phylogeny indicated by one gene may, however, be quite different from that indicated by another. This indicates that major viral groups arise by ‘modular evolution’ (Botstein, 1980), the assembly by recombination of gene modules from several ancestral sources. Viral genomes also contain genes unique to each group, such as the rev and tat genes of lentiviruses. Some of these genes arise de novo, like the overlapping genes of tymo- and luteoviruses, but the sources of many such genes are unknown.

The fact that some genes are shared between viral genera/families has been used to propose higher viral taxonomic groupings. The taxonomic value of such groups is clear when its members share most of their genes (Chapter 29), but is of less obvious utility when few genes are shared as the taxonomist then has to decide on the ‘importance’ of different genes. However, such groupings are interesting as indicators of past biological linkages that have permitted genetic recombination, and they are not comparable to higher groupings in cellular organisms, whose genes are mostly linked by descent and not acquired by genetic recombination from disparate sources; the genomes of the different cellular organelles probably came from
different sources, and there has been limited movement of genes between them.

However, an overview indicates that multicellular organisms are hosts of several major clusters of viruses that have one or more related genes. These are the viruses with single-stranded positive-sense (messenger) RNA genomes, those that are single-stranded negative-sense RNA, those that are double-stranded RNA, those that are DNA and those whose genomes alternate between RNA and DNA. However, even these major ‘viral gene pools’ are not genetically isolated from one another as is shown most dramatically by the clear homology (c. 20%) of the envelope glycoproteins of acariviruses and the gp64 membrane proteins of baculoviruses (Chapter 28); acariviruses have a negative-sense RNA genome and three of their genes are clearly related to those of orthomyxoviruses, whereas the baculoviruses have a double-stranded DNA genome.

**Origins of viruses and rates of evolution**

Viruses are clearly polyphyletic in origin, but there is no unequivocal evidence about when they originated and how quickly they evolve, although comparative taxonomic analyses consistently indicate that their origins are ancient, and their rates of evolution very variable.

Some viral genes, especially those of viruses with DNA genomes, are related to those of their hosts. For example, there are two families of viral thymidine kinases (TKs) (Chapters 18 and 20) and one of these, that found in poxviruses, African swine fever virus and T4 ‘phage, also includes the TKs of their hosts. The sequences of the poxvirus TKs and the central region of the other TKs are clearly homologous, although the non-poxvirus TKs have unique extra portions at their N- and C-termini. It could be that the TK genes have moved from viruses to animals or vice versa, or that both have inherited them from a common ancestor, or that there has been some mix of ‘horizontal’ and ‘vertical’ transmission. Taxonomic analysis of the TK sequences to distinguish between these possibilities is complicated by their differences in length, by the likelihood that the viral sequences are evolving much more quickly than the animal sequences, and by nucleotide biases.

Comparative taxonomy also indicates the great antiquity of the retroviruses and related ‘retroid elements’ (Chapter 27), which include the hepadna- and caulimoviruses, transposons and other mobile genetic elements that have genomes which alternate between DNA and RNA.
phases. Hybridization analysis, for example, has shown that there are virogenes derived from baboon endogenous retrovirus (BaEV) in all Old World monkey species, and that they have co-evolved with their hosts (Benveniste, 1985). Thus BaEV is at least 35 million years old, yet it is just one of many distinct retroviruses that form a small twig of the tree of retroid elements, that can be calculated from comparisons of their reverse transcriptase genes. Retroid elements are found in all types of cellular organisms, giving credence to the suggestion that they are very ancient, and that their reverse transcriptase may have been involved in the earliest phases of life on earth when genomes based on DNA evolved from those based on RNA.

The fast mutation of some viruses with RNA genomes has led some virologists to conclude that ‘most (RNA genome) viruses we know today probably arose since the last Ice Age’! However, there is no evidence for this and although some viruses with RNA genomes do evolve at extraordinary rates, most, like orthomyxoviruses in birds or retroviruses in their wild hosts, give little evidence of current change, their large mutation rate matched by equivalently stringent selection.

The most convincing evidence that viruses with RNA genomes have origins at least as ancient as those with DNA genomes is that, like the latter, most of their genera or groups, and even higher taxa, are ‘phylogenetically contained’; they infect a restricted group of phylogenetically related hosts, and, when their molecular taxonomy is known, it often largely mirrors that of their hosts or vectors. Some are probably very ancient, for example, the rhabdoviruses (Chapter 29), whose two types, the lyssa-like viruses and vesiculovirus-like viruses, are found in both animals and plants. When more of the viruses of lower organisms are known, these links may be better understood.

The ancient origins of viruses with RNA genomes are also indicated by sequence or structural similarities of some of their genes and those of cellular organisms. For example, viral helicases share sequence motifs with some cellular enzymes, and the β-barrel fold found in some virion proteins is also found in the catabolic activator and other cellular proteins. However, these similarities are very tenuous and may have arisen by convergence.

In summary, viruses are fascinating and are starting to attract more attention than they have in the past, from those interested in the study of evolution (Szathmary, 1993). Viruses are an alternative genetic lifestyle to that of cellular organisms, not so much a family, more a way of life, and are probably as ancient as life itself.
10 A. Gibbs, C. Calisher and F. García-Arenal

References