Virus systematics: relationships and names

Adrian Gibbs

Australian National University
Emeritus Faculty
Canberra, ACT 0200
7 Hutt Street, Yarralumla
ACT 2600
Tel (02) 62816971
Email adrian_j_gibbs@hotmail.com

The core activity of biological systematists is to devise systems – classifications – representing the relationships of groups of organisms, usually their evolutionary relationships. These classifications, together with the names of the organisms – taxonomies – can then be used by all to communicate about the organisms, their characteristics, identification, evolution, adaptations and so on.

Viruses: an alternative lifestyle

Viruses are distinguished from cellular organisms by being sub-cellular and having a two-phase life cycle. The transmissible phase consists of virions of distinctive sizes and shapes (that is, rods, spheres or more complex). Virions are metabolically inert until they infect the cells of susceptible hosts, and the viral genes then direct the host's metabolism to generate progeny virions.

Virions and genomes

Viruses present special problems to systematists wishing to devise a universal classification. The great diversity of the form and composition of the virions of different viruses suggests that they are polyphyletic in origin. This has been confirmed over the past two decades by gene sequencing, which has shown that there are no homologous gene families that unite all viruses, like the ribosomal genes found in all cellular organisms.

The simplest virions are regular geometric structures that are assembled from virus-encoded proteins to form a protective covering for the viral genome; a cartoon showing the shapes of the virions of most virus groups is at http://www.ictvdb.org/Images/Viroscoop2005_07minPoster.jpg. The genomes of different viruses differ greatly in size and composition. Some are of DNA, others RNA, they are single- or double-stranded, linear or circular, and either a single nucleic acid molecule or a complement of two or more different molecules.

Among the simplest are the virions of tobacco mosaic virus (TMV), which are rod-shaped, helically-constructed tubes made of many copies of a single type of coat protein with the genome, a single molecule of RNA about 6400 nucleotides long, wound into the basic helix. Many other viruses have isometric virions with the genome centrally folded within a protein shell made from protein subunits arranged as an icosahedron. Those of tobacco necrosis satellite virus (STNV) are about 18 nm in diameter and contain a genome of only 1239 nucleotides of single-stranded RNA, which only encodes the virion protein; STNV only replicates in plant cells already infected with tobacco necrosis virus.

The virions of some viruses, such as influenza virus, have an outer lipid envelope, which contains viral proteins acquired as the virion buds from an infected cell.

The largest virions, those of poxviruses and mimiviruses, are rounded with a very complex interior and are at least 300 nm in size, so just visible in a light microscope. Those of *Acanthamoeba polyphaga* mimivirus (APMV), for example, are about 400 nm in diameter, covered in fibres about 100 nm long. Each virion contains a genome, which is a double-stranded DNA molecule of 1,181,549 base pairs encoding at least 900 proteins. Thus APMV is more complex than some bacteria!

Virus species and their names

Virus species, like species of cellular organisms, are best defined as clusters of phylogenetically related individuals that occupy specific ecological niches[1-2]. Nowadays gene sequences provide the basic data for assessing the phylogenetic relationships of viruses, and although the sequences also contain the functional information which determines their ecology, we do not yet know how to extract such information directly from sequences, and require the hosts and others to do that for us. Thus the host ranges, symptoms and vector types can often only be ascertained by observation and experiment, and attempts to use sequence relatedness as a surrogate for niche occupancy[3] fail when the phylogenetic and functional components of the information in gene sequences are not congruent[4].

The naming of viruses, like all organisms, has had a lively history[5]. Initially they were mostly named after the disease they caused, such as smallpox virus or tobacco mosaic virus. In the 1930–1940s, FO Holmes and HH McKinney tried to introduce Latinised binomials (for example, *Marmor tabaci* for TMV), but most virologists continued to use vernacular names, because the nature of viruses was still largely unknown, and even the knotty problem of whether they were 'living or dead' was still being discussed! Those working with bacteriophages mostly use alphanumeric code names for their viruses, often including the Greek letter ϕ or the Latin letter P to indicate that the name is of a phage, for example ϕX174 and P22. Over the past half-century, as large numbers of distinct viruses have been isolated, it has become clear that many form distinct groups or genera. Various ways have been found for devising group/generic names (for example,[6]), and it seems likely that non-Latinised binomials[7], for example, frangipani mosaic tobamovirus, will become standard[8].

Trees, rates and date

As genes replicate they occasionally mutate; some genes are more mutable than others. Most mutations are deleterious or lethal, and progeny containing them do not survive, but a few are not deleterious and over time these accumulate in the surviving populations, which therefore evolve. The resulting genetic differences reveal the relationships and hence the evolutionary pathways that led to the survivors. There is a hierarchy of different sorts of mutational change as organisms evolve; nucleotide differences result from the most recent changes, while encoded proteins change much more slowly, and 'indels' (that is short ***in***sertions or ***del***etions, often of a single codon), or recombinants (that is chimaeral genes with different parts coming from two or more parental genes) occur even less frequently.

Virus sequences are compared using the same techniques as those used for other gene sequences[9]. Firstly the sequences are aligned using programs such as CLUSTAL[10] or MAFFT[10]. Before the relationships of the sequences can be represented as phylogenetic trees it is important to check for, and remove, recombinant sequences using RDP[11] because recombination, which is common in many virus populations, interferes with the calculation of trees. Many different methods are available for computing trees from sequences, so, to be sure that the results reflect the differences between the sequences not the method, it is best to compare the trees obtained using at least two fundamentally different methods, such as neighbour-joining (NJ)[12] and maximum likelihood (ML)[13]. Trees can be compared by eye, but better by simple graphical methods[14]. The correctness or support for different parts of individual trees can be checked by boot-strapping (that is random sampling of the data).

It is now very fashionable to estimate the age and rate of evolutionary change of populations using ML or Bayesian methods[15]. This requires dated events, such as samplings of the population on significantly different dates. In this way many extant populations of single-virus species have been found to be only hundreds to thousands of years old; however, these analyses give less credible estimates for earlier events, such as the origins of genera, as there is also evidence that many virus genes are as ancient as life itself[16]. This contradiction may merely indicate that extant populations do not preserve ancient evidence because they are of limited effective size and under strong positive selection; a 'Red Queen' scenario[17].

Data and databases; the International Committee on Taxonomy of Viruses (ICTV)

The earliest compilations of virus descriptions and groupings were books, such as Kenneth Smith's *Textbook of Plant Virus Diseases* (1957) and Christopher Andrewes' *Viruses of Vertebrates* (1978). In the 1970s a series of loose-leaf pamphlets of the best-known plant viruses, *Descriptions of Plant Viruses*, was printed in an attempt to find a more flexible way for recording viruses, and these have subsequently been transferred, first to CD-ROM, and then to a database on the internet (http://www.dpvweb.net/). The first computer database of viruses was started in the early 1980s as the VIDE (Virus Identification Data Exchange) project[18]. It used the pioneering DELTA (DEscription Language for TAxonomy) system[19], and aimed to be comprehensive and progressive. The first data collected was of viruses of legumes, and this was expanded in various stages to include all plant viruses (http://www.agls.uidaho.edu/ebi/vdie//refs.htm), and then all other viruses when, in 1991, it became the ICTVdB (http://phene.cpmc.columbia.edu/ICTVdB/index.htm). The work on this database, *per se*, was concluded in 2008. It contains data on 4949 species, 286 genera and 71 families of viruses.

The International Committee on Taxonomy of Viruses (ICTV) was established in 1966 by the Virology Division of the International Union of Microbiology Societies as the official body to develop an internationally agreed taxonomy for viruses, and to establish names for virus taxa and for "subviral agents" (that is, viroids and satellites; http://www.ictvonline.org/codeOfVirusClassification_2002.asp?bhcp=1). Over the years it has published a series of printed reports of which the most recent, the eighth, was published in 2005[2]. The reports provide lists of 'approved' names. They describe the virus genera, and list virus species, but do not describe them. The ICTV website (http://www.ictvonline.org/index.asp) provides (December 2010) a list of 2285 names of 'approved species' taxonomically arranged into 343 genera, each with a named type species, and some grouped into 84 families and six orders.

Gene and protein sequences of viruses are primarily stored in the Genbank/DDBJ/EMBL databases. Subsets of the data are also found in other databases, most notably Genbank's Viral Genomes database, which provides a curated set of complete viral genomes with a single sequence (that is, a type specimen) representing each virus species. Another useful subset is the curated list of plant virus sequence Accession Codes in the Database of Plant Viruses (DPV; http://www.dpvweb.net/), and also various metadatabases, such as the Expert Protein Analysis System (ExPASy; http://expasy.org/) and the Protein Families Database (Pfam; http://pfam.sanger.ac.uk/)

Using virus taxonomies

All current stores of virus taxonomic information have strengths and weaknesses and so must be used with caution. For example, although the ICTV reports describe the features of the 'approved' genera, and list the 'approved' species in them, those species are themselves neither described nor tied to type specimens (for example, stored isolates or genomic sequences), as in other biological taxonomies. Thus there is some uncertainty about what entity each name specifies, and this is compounded by a semantic argument within the ICTV about the uses of italicised and non-italicised names; the italic font is used for naming taxonomic entities, and the roman font for all other purposes.

A confusing feature of the ICTV's hierarchically arranged taxonomic list is that whereas genera consist of sets of viruses that have mostly homologous genes and come from a single ancestor, most families and orders have inherited their genes from more than one parental lineage, and reflect their polyphyletic origins. Thus, in the ICTV list, the lower taxa (that is, the crown groups) are 'natural', whereas the higher taxa are 'artificial'; some are linked by related replicases, others by their coat proteins, and some have few genes in common but have virions of similar shape (for example, the tailed phages or Caudovirales)!

The DELTA format ICTVdB is the most complete, well-organised store of virus data available, but is no longer maintained, *per se*, although its data is to be moved into a MySQL format ICTVdB (http://www.ictvonline.org/ictvdbDev.asp), but the launch date of the new version is not known. This change has been made because the DELTA system is not ideal for describing disparate polyphyletic organisms, like viruses; DELTA works best for sets of phylogenetically related organisms with characters that are homologous but variable.

The Descriptions of Plant Viruses database includes data on only around 400 of the best known plant virus species and genera, but it is very carefully edited and hence the items are readable, and the sequences database very useful.

The most useful tool for identifying novel viruses is the Genbank nucleotide database using its various Basic Local Alignment Search Tools (BLASTn/p/x). The Genbank database currently contains over one million virus sequences, and so a search with a novel virus gene or protein sequence quickly reveals its closest relatives. The calculation of a simple tree of the matched sequences indicates whether the novel virus is indeed novel (that is, outside an existing cluster), or an isolate of a known virus represented in the Viral Genomes list (that is, within an existing cluster). Virus names in Genbank must be treated with caution as a few, especially the old ones, are wrong or have been superseded.

Coda and the future

Virus systematics has had a rough ride during its first half-century. During that period it has become clear that a full genomic sequence is the 'gold standard' for describing a virus and determining its relationships. The sequences of even the largest virus genomes are now readily determined. The ICTV is currently revising its database, and as a result has a golden opportunity, perhaps the only one, to 'spring clean' its taxonomy of viruses. The resulting taxonomy should include only those isolates for which the full genomic sequences and their phenotypic and ecological features are known. All less well studied isolates should either be recorded as 'tentative' relatives of the fully described isolates, or be left unassigned. The attempt by the ICTV to make virus taxonomy look like the hierarchical taxonomy of cellular organisms should be abandoned as it confuses the unwary, and the polyphyletic origins of virus genera should be proclaimed in its published taxonomy; trees for 'natural' crown groups, and gene networks linking those groups. The ICTV database should include data of all known viruses, however imperfectly they are described. This revolution will, of course, not occur any time soon and virus systematics, like all products of evolution, will retain the vestigial evidence of its past!

Acknowledgement

I am indebted to Bryan Harrison and two anonymous referees for their very helpful comments on the penultimate version of this paper.

References

1. Gibbs, A.J. and Gibbs, M.J. (2006) A broader definition of 'the virus species'. *Arch. Virol.* 151, 1419–1422.
2. Fauquet, C.M. *et al.* (2005) *Virus Taxonomy: Classification and Nomenclature of Viruses. 8th Report of the International Committee on the Taxonomy of Viruses*, Elsevier-Academic Press.
3. Bao, Y. *et al.*(2008) in *Encyclopedia of Virology* Vol. 5, (Mahy, B.W.J. and Van Regenmortel, M.H.V., eds), pp. 342-348, Elsevier, Oxford.
4. Gibbs, A. and Ohshima, K. (2010) Potyviruses and the digital revolution. *Annu. Rev. Phytopath.* 48, 205–223.
5. Matthews, R.E.F. (1983) in *A critical appraisal of viral taxonomy* (Matthews, R.E.F., ed.), pp. 1–35, CRC Press.
6. Harrison, B.D. *et al.* (1971) Sixteen groups of plant viruses. *Virology* 45, 356–363.
7. Fenner, F. (1976) Classification and nomenclature of viruses. Second report of the International Committee on Taxonomy of Viruses. *Intervirol.* 7, 1–115.
8. Van Regenmortel, M.H.V. *et al.* (2010) A proposal to change existing virus species names to non-Latinized binomials. *Arch. Virol.* 155, 1909–1919.
9. Lemey, P. *et al.* (2009) *The Phylogenetic Handbook* (2nd edn) p. 9, Cambridge University Press.
10. Jeanmougin, F. *et al. (1998)* Multiple sequence alignment with Clustal X. *Trends Biochem. Sci.* 23, 403–405.
11. Martin, D. P. *et al.* (2005) RDP2: recombination detection and analysis from sequence alignments *Bioinform.* 21, 260–262.
12. Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425.
13. Guindon, S. and Gascuel, O. (2003) A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* 52, 696–704.
14. Fourment, M. and Gibbs, M. J. (2006) PATRISTIC: a program for calculating patristic distances and graphically comparing the components of genetic change. *BMC Evo. Biol.* 6, 1.
15. Drummond, A. J. and Rambaut, A. (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* 7, 214.
16. Koonin, E.V. *et al.* (2008) The Big Bang of picorna-like virus evolution antedates the radiation of eukaryotic supergroups. *Nature Rev. Microbiol.* 6, 925–939.
17. Vermeij, G.J. (1987) *Evolution and escalation: An ecological history of life*, Princeton University Press.
18. Boswell, K.F. *et al.* (1986) The VIDE (Virus Identification Data Exchange) project: a data base for plant viruses. *Rev. Plant Pathol.* 65, 221–231.
19. Dallwitz, M. (1980) A general system for coding taxonomic descriptions. *Taxon* 29, 41–46.

Biography

Adrian Gibbs is a true believer of Theodosius Dobzhansky's famous dictum that "Nothing in biology makes sense except in the light of evolution", and has been lucky enough to be able to apply this concept to the study of viruses all his working life. He worked from 1956 to 1966 on viruses of plants and bees at Rothamsted Experimental Station, UK, the world's oldest agricultural research station, and since then at the Australian National University, Canberra. He was elected a Life Member of the International Committee on Taxonomy of Viruses in 1996.