Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Introduction: Evolution, Phylogeny, and Viruses

Biological species, including viruses, change through generations and over time in the process known as evolution. These changes are first fixed in the genome of successful individuals that give rise to genetic lineages. Due to either limited fidelity of the replication apparatus copying the genome or physico-chemical activity of the environment, nucleotides may be changed, inserted, or deleted. Genomes of other origin may also be a source of innovation for a genome through the use of specially evolved mechanisms of genetic exchange (recombination). Accepted changes, known as mutations, may be neutral, advantageous, or deleterious, and depending on the population size and environment, the mutant lineage may proliferate or go extinct. Overall, advantageous mutations and large population size increase the chances for a lineage to succeed. The lineage fit is constantly reassessed in the ever-changing environment and lineages that, due to mutation, became a success in the past could be unfit in the new environment. Due to the growing number of mutations accumulating in the genomes, lineages diverge over time, although occasionally, due to stochastic reasons or under similar selection pressure, they may converge.

The relationship between biological lineages related by common descent is called phylogeny; the same term also embodies the methodology of reconstructing these relationships. Phylogeny deals with past events and, therefore, it is reconstructed by quantification of differences accumulated between lineages. Due to the lack of fossils and (relatively) high mutation rate, viruses were not considered to provide a recoverable part of phylogeny until the advent of molecular data proved otherwise. Comparison of nucleotide and amino acid sequences, and, occasionally, other quantitative characteristics such as distances between three-dimensional structures of biopolymers, have been used to reconstruct virus phylogeny. Results of phylogenetic analysis are commonly depicted in the form of a tree that may be used as a synonym for phylogeny. For instance, all-inclusive phylogeny of cellular species is depicted as the Tree of Life (ToL).

With few exceptions, virus phylogeny follows the theory and practice developed for phylogeny of cellular life forms. For inferring phylogeny, differences between the sequences of species members, assumed to be of a discernable common origin, are analyzed. If species in all lineages evolve at a uniform constant rate, like clock ticks, their evolution conforms to a molecular clock model. The utility of this model in relation to viruses may be very limited. Rather, related virus lineages may evolve at different and fluctuating rates and some sites may mutate repeatedly with each new mutation erasing a record about the prior change. As a result, the accumulation of inter-species differences may progress non-linearly with the time elapsed. At present, our understanding of these parameters of virus evolution is poor and this limits our ability to assess the fit between a reconstructed phylogeny and the true phylogeny, with the latter practically remaining unknown for most virus isolates. This gap in our knowledge does not eliminate the conceptual strength of phylogenetic analysis for reconstructing the relationships between biological species.

The ultimate goal of virus phylogeny is reconstructing the relationships between ‘all’ virus isolates and species. In contrast to cellular species, which form three compact domains (kingdoms) and whose origin is traced back to a common ancestor in the ToL, major virus classes may combine species that have originated from different ancestors. Thus, reconstructing the comprehensive virus phylogeny requires comparisons that involve genomes of virus and cellular origins. This formidable task remains largely ‘work in progress’. In fact, most efforts in virus phylogeny are invested in reconstructing the relationships at the micro, rather than grand, scale and they focus on well-sampled lineages that have practical (e.g., medical) relevance. Phylogeny itself or in combination with other data may provide a deep insight into virus evolution and diverse aspects of virus life cycles, including virus interactions with their hosts.

Our knowledge about contemporary virus diversity has been steadily advancing with new viruses being
constantly described by systematic efforts as well as occasional discoveries. These developments indicate that only a small part of virus diversity has so far been unraveled and has become available for phylogenetic studies. It is also likely that many more lineages existed in the past; some of these lineages are likely to have ancestral relationships with contemporary lineages.

**Tree Definitions**

Species share similarity that varies depending on the rate of evolution and time of divergence. The entire process of generating contemporary species diversity from a common ancestor is believed to proceed through a chain of intermediate ancestors specific for different subsets of the analyzed species. The relationship between the common ancestor, intermediate ancestors, and contemporary species may be likened to the relationship between, respectively, root, internal nodes, and terminal nodes (leaves) of a tree, an abstraction that is widely used for the visualization of this relationship. Trees are also part of graph theory, a branch of mathematics, whose apparatus is used in phylogeny. Formally and due to a strong link between phylogeny and taxonomy, leaves may be called operational taxonomy units (OTUs) and internal nodes and roots, since they have not been directly observed, are known as hypothetical taxonomy units (HTUs). Nodes are connected by branches or edges.

The tree may be characterized by topology, length of branches, shape, and the position of the root. The topology is determined by relative position of internal and terminal nodes; it defines branching events leading to contemporary species diversity. If two or more trees obtained for different data sets feature a common topology, these trees are called congruent. The branch length of a tree may define either the amount of change fixed or the time passed between two nodes connected in a tree, and is known as ‘additive’ or ‘ultrametric’, respectively. The tree shape may be linked to particulars of evolutionary process and reflect changes in the population size and diversity due to genetic drift and natural selection. The position of the root at the tree defines the direction of evolution. Species that descend from an internal node in a rooted tree form a cluster and the node is called most recent common ancestor (MRCA) of the cluster that thus has a monophyletic origin. The branch lengths and the root position may be left undefined for a tree that is then called ‘cladogram’ and ‘unrooted tree’, respectively.

**Phylogenetic Analysis**

Multiple alignment of polynucleotide or amino acid sequences representing analyzed species and maximized for similarity is traditionally used as input for phylogenetic analysis. The quality of alignment is among the most significant factors affecting the quality of phylogenetic inference. Due to the redundancy of the genetic code, changes in polynucleotide sequences are accumulated at a higher rate than those in amino acid sequences. In viruses, including RNA viruses, this difference is not counterbalanced by other constraints linked to dinucleotide frequency and RNA secondary (tertiary) structure. Because of these differences, phylogeny of closely related species is commonly inferred using polynucleotide sequences, while protein sequences, preserving better phylogenetic signal, may be used to infer phylogeny of distantly related species.

Differences between species, as calculated from alignment, may be quantified as either pairwise distances forming a distance matrix or position-specific substitution columns (discrete characters of states of alignment), the latter preserving the knowledge about location of differences. The respective methods dealing with these quantitative characteristics are known as distance and discrete (character state). The distance methods are praised for their speed and are considered a technique of choice for analysis of large data sets. They are often designed to converge on a unique phylogeny, with none others being even considered. The unweighted pair group method with arithmetic means (UPGMA) in which a constantly recalculated distance matrix is used to define the hierarchy of similarities through systematic and stepwise merging of most similar pairs at a time was the first technique introduced for clustering. The neighbor-joining (NJ) method uses a more sophisticated algorithm of clustering that minimizes branch lengths, and is the most popular among distance methods. Although different trees may be compared in how they fit a distance matrix, it is distance character-based methods that are routinely used to assess numerous alternative phylogenies in search for the best one in a computationally very intensive process. Due to the calculation time involved, assessing all possible phylogenies is found to be impractical for data sets including more than 10 sequences; for larger data sets different heuristic approximations are used that may not guarantee a recovered phylogeny to be the best overall. There are two major criteria for selecting the best phylogeny using character-state based information through either maximum parsimony (MP) or maximum likelihood (ML). In MP analysis, a phylogeny with a minimal number of substitutions separating analyzed species is sought. The ML analysis offers a statistical framework for comparing the likelihood of fitting different trees into the data in search for one with the best fit. The latter approach is mathematically robust and its statistical power may also be used in combination with other techniques of tree generation. Most recently, a Bayesian variant of the ML
approach has gained popularity. It utilizes prior knowledge about the evolutionary process in combination with repeated sampling from subsequently derived hypotheses.

After a tree is chosen, it is common to assign support for internal nodes through assessing nodes’ persistence in trees related to the chosen tree. One particular technique, called bootstrap analysis, in which trees are generated for numerous randomly modified derivatives of the original data set, is most frequently used. Each internal node in the original tree is characterized by a so-called bootstrap value that is equal to the number of nodes appearing in all tested trees. Although the relationship between bootstrap and statistical values is not linear, nodes with very high bootstrap values are considered to be reliable. If species evolve according to a molecular clock model, the root position in a tree could directly be calculated from the observed inter-species differences as a midpoint of cumulative inter-species differences. Alternatively, the root position may be assigned to a tree from knowledge about analyzed species that was gained independently from phylogenetic analysis. Commonly, this knowledge comes in the form of a single or more species which are assumed (or known) to have emerged before the ‘birth’ of the analyzed cluster. These early diverged species are collectively defined as ‘outgroup’, while the analyzed species may be called ‘in-group’. Also, a tree may be generated unrooted, a common practice in phylogenetic analysis of viruses for which the applicability of the molecular clock model remains largely untested and reliable outgroups may not be routinely available.

In the unrooted tree, grouping of species in separate clusters may be apparent, although these clusters may not be treated as monophyletic as long as the direction of evolution has not been defined. These challenges are addressed by the development of new approaches that infer rooted trees without artificially restricting species evolution to a constant rate (known as relaxed molecular clock models).

Virus phylogeny can be inferred using genomes or distinct genes and each of these approaches, standard in phylogenomics, may be considered as complementary. Under the first approach, genome-wide alignments are used for analysis. Due to complexities of the evolutionary process that may be region specific, reliable genome-wide alignments can routinely be built only for relatively closely related viruses whose analysis, however, may be further complicated by recombination events (see below). Using the second approach, genes with no evidence for recombination may be merged (concatenated) in a single data set that may be used to produce a superior phylogenetic signal compared to those generated for distinct genes or entire genomes. For viruses with small genomes or for a diverse set of viruses, it is common practice to use a single gene to infer virus phylogeny. Although the results produced may be the best models describing evolutionary history of a group of viruses, the validity of this gene-based approach for the genome-wide extrapolation remains a point of debate.

When the tree is reconstructed for (part of) genomes, an underlying assumption is that the analyzed data set has a uniform phylogeny. This condition may be violated due to homologous recombination between (closely) related viruses. In phylogenetic terms, recombination may be revealed through incongruency of trees built for a genome region, where recombination occurs, and other regions. Trees may also become incongruent due to various technical reasons related to the size and diversity of a virus data set and deviations of the evolutionary process among lineages. These characteristics complicate interpretation of the congruency test, which is widely used in different programs to identify recombination in viruses.

Applications of Phylogeny in Virology

Phylogenetic analysis is used in a wide range of studies to address both applied and fundamental issues of virus research, including epidemiology, diagnostics, forensic studies, phylogeography, origin, evolution, and taxonomy of viruses. First question to be answered during an outbreak of a virus epidemic concern the virus identity and origin. Answers to these questions form the basis for implementing immediate practical measures and prospective planning enabling specific and rapid virus detection and epidemic containment, which may include the use and development of antiviral drugs and vaccines. Among different analyses performed for virus identification at the early stage of a virus epidemic, the phylogenetic characterization is used for determining the relationship of a newly identified virus with all other previously characterized and sequenced viruses.

Results of this analysis may be sufficient to provide answers to the questions posed, as regularly happens with closely monitored viruses that include most human viruses of high social impact, for example, influenza, human immunodeficiency virus (HIV), hepatitis C virus (HCV), poliovirus, and others. For these viruses, there exist large databases of previously characterized isolates and strains that comprehensively cover the natural diversity. Should a newly identified virus belong to one of these species, chances are that it has evolved from a previously characterized isolate or a close variant and this immediately becomes evident in the clustering of these viruses in the phylogenetic tree. Combining the results of gene-specific and genome-wide phylogenetic analysis allows one to determine whether recombination contributed to the isolate origin. For instance, recombination was found to be extremely uncommon in the evolution
of HCV, but not for poliovirus lineages that recombine promiscuously, also with closely related human coxsackie A viruses, both of which belong to human enteroviruses.

When an emerging infection is caused by a new never-before-detected virus, the phylogenetic analysis is instrumental for classification of this virus and in the case of a zoonotic infection, for determining the dynamic of virus introduction into the (human) population and initiating the search for the natural virus reservoir. This was the case with many emerging infections including those caused by most recently introduced Nipah virus, a paramyxovirus, and SARS coronavirus (SARS-CoV). With the latter virus, poor sampling of the coronavirus diversity in the SARS-CoV lineage at the time, some uncertainty over the relationship between phylogeny and taxonomy of coronaviruses, and the complexity of phylogenetic analysis of a virus data set including isolated distant lineages led to considerable controversy over the exact evolutionary position of SARS-CoV among coronaviruses. Since then, the matter has largely been resolved but this experience illustrates some challenges in inferring virus phylogeny.

The search for a zoonotic reservoir of an emerging virus may involve a significant and time-consuming effort that requires numerous phylogenetic analyses of ever-expanding sampling of the virus diversity generated in pursuit of the goal. In this quest, phylogenetic analysis canalizes the effort and provides crucial information for reconstructing parameters of major evolutionary events that promoted the virus origin and spread. For instance, intertwining HIV and simian immunodeficiency virus (SIV) lineages in the primate lentivirus tree led to postulation that the existing diversity of HIV in the human population originated from several ancestral viruses independently introduced from primates over a number of years. Similar phylogenetic reasoning was used to trace the origin of a local HIV outbreak to a common source of HIV introduction through dental practice (known as ‘HIV dentist’ case). These are typical examples illustrating the utility of phylogenetic analysis for epidemiological and forensic studies.

Geographic distribution of places of virus isolation is another important characteristic relative to which virus phylogeny may be evaluated. This field of study belongs to phylegography. The evolution of human JC polyomavirus provides an example of confinement of circulation of virus clusters to geographically isolated areas, represented by three continents. Recent identification of West Nile virus in the USA illustrates geographical expansion of an Old World virus into the New World. Analysis of phylogenies of field isolates of rabies virus of the family *Rhabdoviridae* sampled from different animals over Europe led to the recognition that interspecies virus expansion is occurring faster when compared to geographical expansion.

Phylogenies also reveal information about the relative strength of the virus–host association over time. In some virus families (e.g., the *Coronaviridae*) host-jumping events may be relatively frequent, including the emergence of at least two human viruses, dead-end SARS-CoV and successfully circulating human coronavirus OC43 (HCoV-OC43). At the other end of the spectrum one finds the family *Herpesviridae*. Extensive phylogenetic analysis of herpesviruses and their hosts showed a remarkable congruency of topologies of trees indicating that this virus family may have emerged some 400 million years ago and that herpesviruses coevolve with their hosts.

Phylogenetic analysis becomes increasingly important in virus classification (taxonomy) and relies on complex multicharacter rules applied to separate virus families by respective ‘study groups’. For viruses united in high-rank taxa above the genus level, phylogenetic clustering for most conserved replicative genes is commonly observed and used in the decision making process. For instance, human hepatitis E virus, originally classified as a calicivirus using largely virion properties, was eventually expelled from the family due to poor fit of genome characteristics, including results of phylogenetic analysis. Phylogenetic considerations also played an important role in forming recently established families, for example, the *Marnaviridae* and *Dicistroviridae*. In contrast, phylogenetic analysis has been of relatively little use in the taxonomy of large DNA phages which has been developed in such a way that existing families may unite phages with different gene layouts and phylogenies. The relationship between phylogeny and taxonomy is evolving and in future one might hope for important advancements that improve cross-family consistency in relation to phylogeny.

See also: Emerging and Reemerging Virus Diseases of Vertebrates; Evolution of Viruses; Origin of Viruses; Taxonomy, Classification and Nomenclature of Viruses; Virus Classification by Pairwise Sequence Comparison (PASC); Virus Databases; Virus Evolution: Bacterial Viruses; Virus Species.

**Further Reading**

Dolja VV and Koonin EV (eds.) (2006) Comparative genomics and evolution of complex viruses. *Virus Research* 117: 1–184.

Domingo E (2007) Virus evolution. In: Knipe DM, Howley PM, Griffin DE, et al. (eds.) Fields Virology, 5th edn., pp. 389–421. Philadelphia, PA: Wolters Kluwer, Lippincott Williams and Wilkins.

Domingo E, Webster RG, and Holland JJ (eds.) (1999) Origin and Evolution of Viruses. San Diego: Academic Press.

Fauquet CM, Mayo MA, Maniloff J, Desselberger U, and Ball LA (2005) *Virus Taxonomy: Eighth Report of the International Committee on Taxonomy of Viruses*. San Diego, CA: Elsevier Academic Press.

Felsenstein J (2004) Inferring Phylogenies. Sunderland, MA: Sinauer Associates, Inc.

Gibbs AJ, Calisher CH, and Garcia-Arenal F (eds.) (1995) *Molecular Basis of Virus Evolution*. Cambridge: Cambridge University Press.
**Glossary**

**Cre** (cis-acting replication element) First found in human rhinovirus genomic RNA and subsequently identified in other picornavirus genomes, the cre acts as template for the viral RNA polymerase 3Dpol to uridylylate VPG to VPg-pU-pU. Evidence suggests that this cre can function in trans as well.

**IRES** (internal ribosome entry site) An RNA sequence typically characterized by extensive nucleic acid secondary structure. The 40S ribosomal subunit of the cellular translation machinery interacts with RNA stem loops/sequence, and subsequently allows translation of downstream RNA sequence of the IRES. Translation, therefore, proceeds without recognition of a 5’ cap. Utilized by some virus families (including Picornaviridae) and cellular messenger RNAs.

**Polyprotein** In the context of the discussion of Picornaviridae, this refers to the long protein resulting from translation of the single open reading frame of picornavirus RNA. The polyprotein is processed by viral proteinases to yield mature viral proteins.

**Positive-strand RNA** A single molecule of picornavirus RNA that encodes functional viral protein when translated in the 5’–3’ direction. This is the ‘sense’ orientation of picornavirus RNA as it enters the cell that is also encapsidated in progeny virions.

**RNP complex** Ribonucleoprotein complex. Describing a stable interaction of RNA and protein(s), either in vivo or in vitro.

**Uridylation** Refers to the addition of two uridylic residues to the VPg molecule by the picornavirus RNA-dependent RNA polymerase 3Dpol (and other viral proteins) using a viral RNA template.

**VPg** Virus protein, genome-linked. Also known as 3B, the function of VPg is to act as a protein primer for the picornavirus RNA-dependent RNA polymerase. Following uridylation, VPg-pU-pU is covalently attached to the 5’ end of picornavirus RNAs (positive- and negative-strands).

**Overview of Virus Family Picornaviridae**

Introduction

Viruses belonging to the family Picornaviridae are small (Latin *Pico*) RNA (*rna*) viruses whose host range is typically restricted to mammals. Genera associated with Picornaviridae include erbovirus, teschovirus, kobuvirus, aphthovirus, cardioivirus, enterovirus, hepatovirus, parechovirus, and rhinovirus. The first three of these genera are relatively recent additions to the picornavirus family, and the last four contain pathogens that are the most extensively studied picornaviruses capable of infecting humans. Particularly, poliovirus of the enterovirus family is widely considered to be the ‘prototypical’ picornavirus, and perhaps the most feared by humans due to the potential for poliovirus infection to result in paralytic poliomyelitis. This debilitating affliction can result in paralysis of one or more limbs in an infected individual, and in rare cases even death. Human rhinovirus infection in humans results in the common cold, and is one of the most prevalent diseases throughout the world. While in the developed world the common cold is often seen merely as an inconvenience at worst, it is the most important cause of asthma exacerbations, and there is no effective vaccine for the virus nor any effective medical treatment for infection.

Due to the intense interest in Picornaviridae based upon the diseases associated with picornavirus infections, the virus family has received extensive scientific attention to understand the mechanisms of gene expression and replication of its members. In turn, insights into the manner in which members of the Picornaviridae propagate have allowed a better understanding of how these viruses cause disease and also how other unrelated virus families