Van Regenmortel MHV (1990) Virus species, a much overlooked but essential concept in virus classification. *Intervirology* 31: 241–254.

Van Regenmortel MHV, Bishop DHL, Fauquet CM, Mayo MA, Maniloff J, and Calisher CH (1997) Guidelines to the demarcation of virus species. *Archives of Virology* 142: 1505–1518.

Relevant Website

http://www.ictv.ird.fr – ICTV; Taxonomic Proposal Management System.

Virus Classification by Pairwise Sequence Comparison (PASC)

Y Bao, Y Kapustin, and T Tatusova, National Institutes of Health, Bethesda, MD, USA

Glossary

Demarcation A mapping of ranges of pairwise distances into taxonomic categories.

Introduction

Virus classification is very important for virus research. It is also an extremely difficult task for many virus families. Traditionally, virus classification relied on properties such as virion morphology, genome organization, replication mechanism, serology, natural host range, mode of transmission, and pathogenicity. Yet viruses sharing the above properties can reveal tremendous differences at the genome level. For example, classification of many phages is currently based on presence, structure, and length of a tail, and this approach has been shown not to correlate with genomic information, leading to a very difficult situation and hundreds of unclassified phages.

Molecular virus classification based on virus sequences has been used increasingly in recent years, thanks to the growing number of viral sequences available in the public sequence databases. The most commonly used sequence comparison methods include multiple sequence alignment and phylogenetic analysis. Another molecular classification method that has drawn more and more attention from virologists is pairwise sequence comparison (PASC). In this article, we briefly describe various sequence comparison methods, introduce the PASC tool, and compare it with other methods.

Sequence Comparison Methods

A universal approach to compare biological sequences, in a sense of producing meaningful results at various levels of divergence, is in the realm of sequence alignment. An alignment is an arrangement of residues of two or more sequences in a way that reveals their possible relatedness, with space characters inserted into the sequences to indicate single-residue insertions and deletions. A variety of algorithms and programs are available to suit a wide range of problems requiring sequence alignments as parts of their solutions. Depending on the specifics of a problem, different types of algorithms or their combinations may work best. Most alignment algorithms can be broadly categorized by the scope of their application on sequences (local vs. global), or by the number of sequences involved (pairwise vs. multiple).

Each pairwise alignment can be viewed as an array of per-residue operations transforming one sequence to the other. These operations are substitutions (called matches and mismatches in nucleotide alignments), insertions, and deletions. A generalization of this concept to multiple alignments is possible. Alignments are scored using a scoring scheme appropriate for a biological context. The widely used affine scheme assigns substitution scores to substitutions and a penalty to each space, and an additional penalty to each gap defined as a maximal consecutive run of spaces. Given a set of sequences, an alignment is called optimal if it has the maximal score over all possible alignments. Optimal alignments are not necessarily unique; two or more alignments can be tied with the same score.

Local algorithms are capable of detecting similarities between arbitrary parts of sequences. Applications involving local alignments are numerous, including search for orthologous genes or conserved protein domains. Alignments produced with local algorithms are tractable to mathematical analysis, which allowed the building of tools that evaluate statistical significance of the alignments. The algorithm for computing optimal local alignments is known as Smith–Waterman and it runs in time proportional to the product of the sequences' lengths. Since this is too slow for large-scale searches, many heuristic methods (with BLAST being the most popular) have been developed, allowing matching typical queries against gigabase-sized archives of sequence in a matter of seconds.

Fast as they are, algorithms like BLAST are not suitable for all applications. Although they are capable of picking out segments of high similarity, no segment of input sequences is guaranteed to be a part of the resulting alignments, and some segments may belong to more than one individual alignment. Additional post-processing steps are often required in order to produce consistent sets of local alignments. This complicates the use of local alignments in applications involving uniform computing of identities.

When sequences in the set are expected to align end-to-end, global alignment algorithms are applicable. The strict algorithm for computing an optimal global alignment is known as Needleman–Wunsch. With the running time estimate being the same as in the Smith–Waterman, global alignments allow straightforward evaluation of identities as they provide unambiguous mapping for every residue. An important end-space free variant is used when one of the sequences is expected to align in the interior of another. Global algorithms are normally not suitable for applications aiming to capture rearrangement events.

Multiple sequence alignments can technically be viewed as a generalization of the pairwise case. However, they often serve different goals, such as a detection of weak and/or dispersed similarities over a set of sequences known to share a common function or structure. Computing an optimal multiple sequence alignment is a computationally costly task, and most implementations use various heuristics to approximate the alignment in a reasonable time. Note that even when the optimal multiple alignment is available, pairwise alignments inferred from it are not guaranteed to be optimal. There are many multiple sequence alignment tools available: CLUSTALW, DIALIGN, MAFFT, MUSCLE, PROBCONS, ProDA, and T-COFFEE, etc.

Phylogenetic analysis is probably the most frequently used molecular virus classification tool. A phylogeny or evolutionary tree is a mathematical structure which is used to model the historical relationships between groups of organisms or sequences. Main types of methods used to construct phylogenetic trees include distance-based methods (such as neighbor-joining), parsimony, maximum likelihood, and other probabilistic inference techniques. The most common distance-based methods utilize multiple sequence alignments to estimate the evolutionary distance between each pair of sequences and reconstruct the tree from the distances. Either protein sequences or DNA sequences can be used.

Phylogenetic analysis was used in the vast majority of virus families described in the Eighth Report of the International Committee on the Taxonomy of Viruses (ICTV) to support their classification. It has also been applied to the classification of a large group of distantly related viruses. For example, phages consist of many different families. Therefore, the conventional phylogenetic analysis that uses genomic sequences or individual protein sequences would not work for the classification of phages as a whole group. A 'phage proteomic tree' was developed to classify phages by the overall similarities of all protein sequences present in the phage genomes.

Although phylogenetic analysis is well established as a tool for virus classification, it is usually computationally intensive, and requires expertise to perform the analysis and to interpret the results. A more robust method is preferred so that researchers without an advanced computer system and advanced knowledge about the phylogenetic analysis can also use it. Also, as discussed below, despite the fact that some of the sequence alignment methods (such as BLAST) are very fast and easy to carry out, their results will not reveal the taxonomic relationships between the two viruses. A method that can place a new virus in the appropriated taxonomic position is desired. PASC is a good combination of the two methods.

The Principle of PASC

In the PASC system, pairwise global alignment is performed on complete genomes or particular protein sequences for each viral family, and their percentage of identity is calculated. The number of virus pairs at each percentage is plotted. The distribution of the identities is not evenly spread, but rather clustered into groups of peaks for viruses at strain, species, genus, and subfamily levels. The percentage range of each peak serves as a good reference for taxonomic classification based on sequence similarities. This method has been applied to polioviruses using the protein and nucleotide sequences of the VP1 gene, as well as the whole genome sequence, coronaviruses using the protein sequences of the polymerase and helicase, potyviruses using the protein and nucleotide sequences of the complete ORF and the coat protein gene, geminiviruses using the complete sequences of DNA-A, flexiviruses using the protein and nucleotide sequences of the three major viral gene products (replication protein, triple gene block and coat protein), papillomaviruses using the nucleotide sequences of the L1 gene, and poxviruses using the protein sequences of the DNA polymerase. There is an increasing interest to expand PASC to other virus families.

In order to apply PASC to a larger number of virus families and be used by a wide range of virologists, the following should be considered:

1. The same algorithm and sequence dataset should be used to determine the demarcations and to place new viruses in the right taxonomic position.
2. The sequence dataset used for demarcation determination and the virus taxonomy database should be updated frequently to reflect the most recent status.

3. The algorithm should be robust and fast enough for large sequence sets.
4. The system should be readily accessible to researchers worldwide and easy to use.

PASC Implementation at NCBI

The National Center for Biotechnology Information (NCBI) has developed a web-based PASC system that meets all of the criteria mentioned above. In this implementation, complete viral genomes are organized into groups corresponding to broad taxonomic entities such as family or floating genus. Within each group, alignments are pre-computed and stored in a database for each pair of the genomes. The alignments are used to evaluate identities, defined as the ratio of matching residues over the total alignment length. The alignments are computed using the pairwise global algorithm with the affine scoring scheme assigning one to matches and minus one to mismatches and nonterminal spaces and gaps. Since the genomes vary in lengths, terminal spaces are not penalized during the alignment computing but taken into account when computing the identity.

PASC interface is built around a histogram of pairwise identities. The primary feature of the interface is the comparison of an external sequence such as a newly sequenced viral genome, with genomes in a user-selected group. After the sequence is submitted, PASC will start computing the alignments, or extracting them from the database if the query is a member of the group. At the end of the process, a user is presented with a list of closest matches. Matches can be selected to visualize their positions on the identity distribution chart.

The PASC system at NCBI not only reproduced results for virus families for which PASC had been applied to, but also generated data with well-separated identity distributions that can be used as taxonomy demarcations for other virus families such as *Caliciviridae*, *Flaviviridae*, and *Togaviridae* among others.

Applications of PASC

PASC can be used to define taxonomy demarcations for many viral families. Two examples are shown here. In the first example, 45 complete genomes from the family *Luteoviridae* were used to construct the distribution of the pairwise identities among the genomes (**Figure 1**). Pairs located at 88% and up are all from different strains in the same species; pairs between 53% and 85% are mostly from different species in the same genus; and pairs located at 52% and lower are all from different genera. These percentages can therefore serve as demarcations

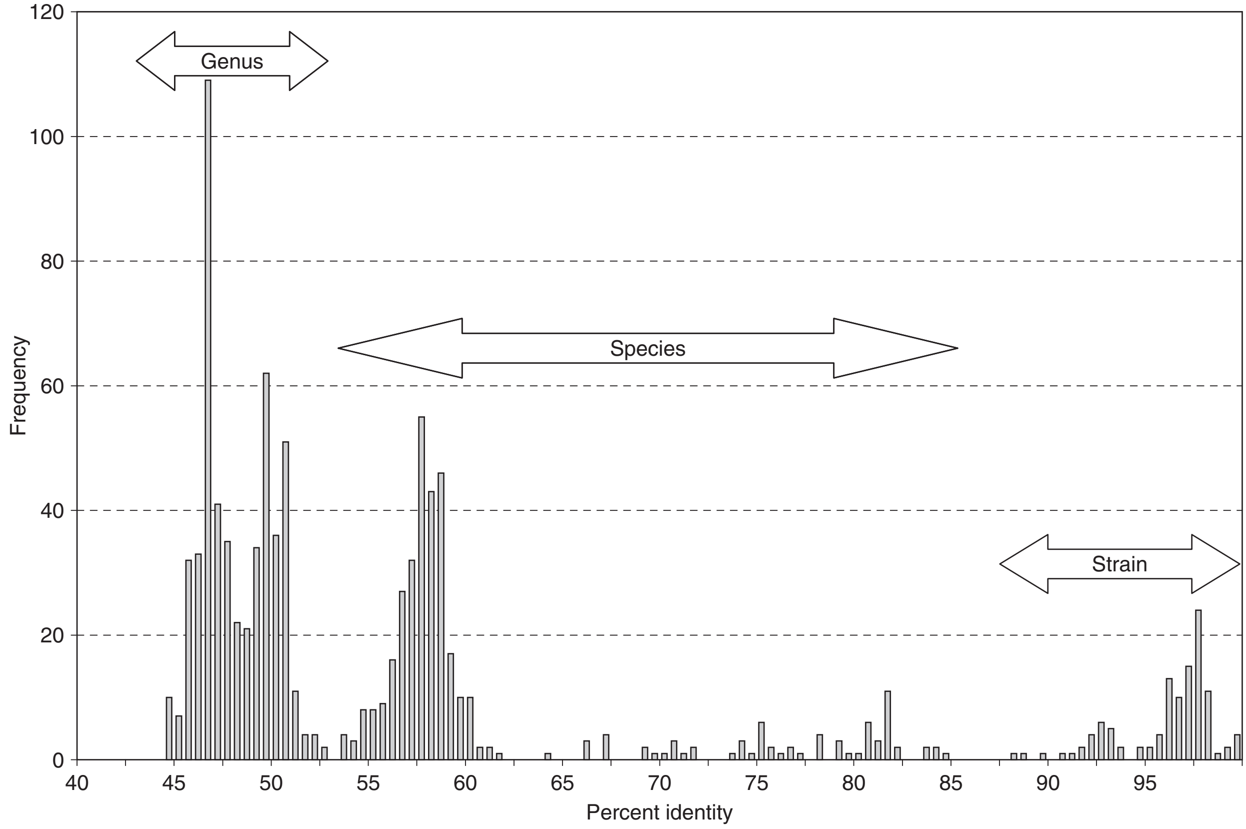

Figure 1 Frequency distribution of pairwise identities from the complete nucleotide sequence comparison of 45 luteoviruses.

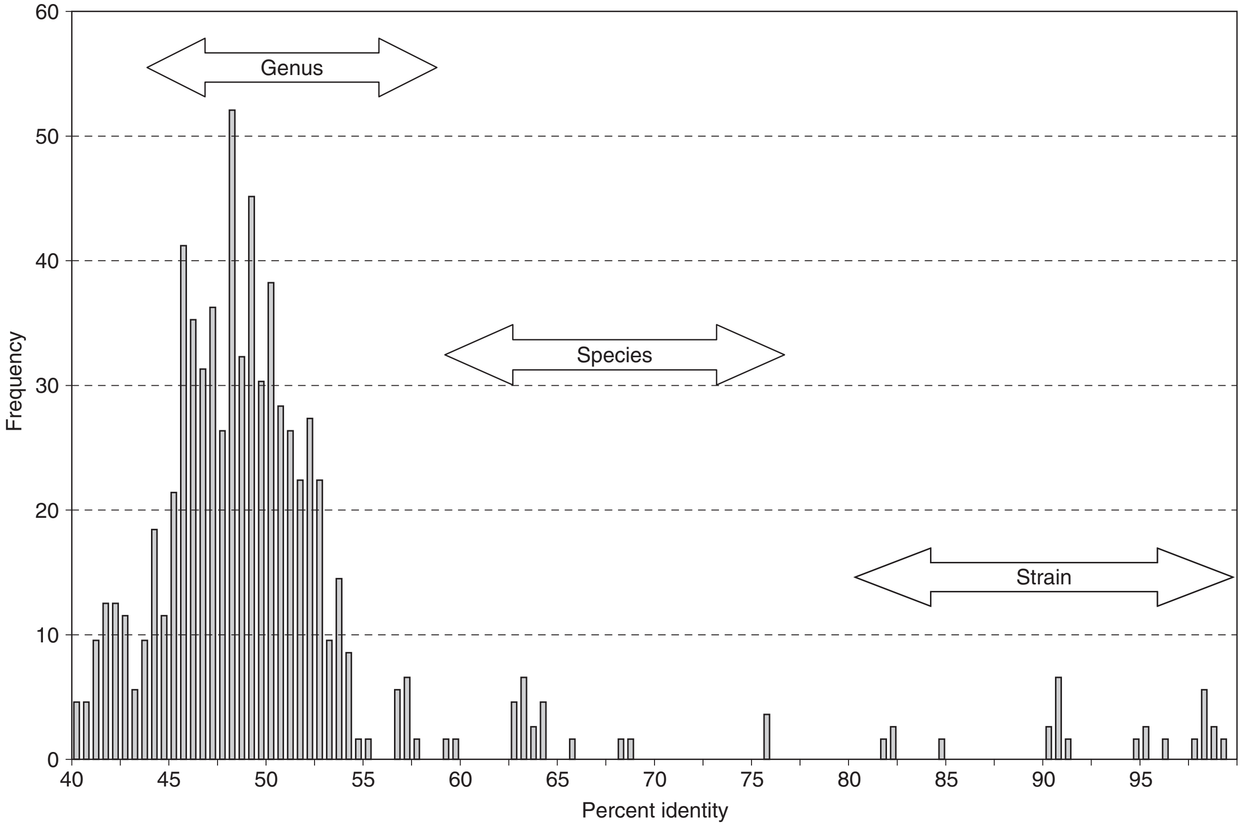

Figure 2 Frequency distribution of pairwise identities from the nucleotide sequence comparison of the RdRp segments of 38 reoviruses.

for classification of luteoviruses. In the second example, the pairwise identities of 38 RNA-dependent RNA polymerase (RdRp) gene segment of viruses in the family *Reoviridae* were calculated and plotted (**Figure 2**). Similar to the above mentioned example, 81% and up, between 59% and 76%, and 57% and below can be used as boundaries for strains, species, and genera of reoviruses based on the identities between their polymerase genes. It should be noted that the determination of demarcation using PASC is not always straightforward. For some virus families, phenotypic characteristics of the viruses are required to be taken into account.

PASC can place newly sequenced viruses into the correct taxonomy group. For example, a genomic sequence of a luteovirus (GenBank accession number AY956384) appeared recently in the international sequence databases with the name chickpea chlorotic stunt virus, which is not an official ICTV species name. When this sequence was tested with the luteoviruses in the PASC system, it was found that the virus with the highest sequence similarity to it is cucurbit aphid-borne yellows virus in the genus *Polerovirus*. The similarity is 61.7%, which is in the demarcation of different species in the same genus. It can thus be suggested that this virus is a member of a new species in the genus *Polerovirus*.

Finally, PASC can identify possible questionable classifications in the existing groups when the peaks on the graph are very well separated. **Table 1** lists the pairs of the RdRp segment of reoviruses whose identities are between 54% and 54.5% in **Figure 2**. From the description above, this region represents pairs of viruses from different genera. This is true for most of the pairs in the table. However, the group also includes a pair containing St. Croix river virus and palyam virus, which both currently belong to the genus *Orbivirus*. Further investigation using PASC revealed that palyam virus has identities higher than 59% with many other viruses in the family, and therefore is indeed an orbivirus. The highest identity of St. Croix river virus with other viruses is only about 54%, which is lower than the demarcation for species in the same genus. This suggests that the species *St. Croix river virus* should probably be placed in a new genus in the *Reoviridae* family.

Advantages of PASC Compared to Other Methods

Unlike other classification methods based on the phenotypic properties of viruses, PASC is a quantitative tool. For those virus families that are suitable for PASC analyses,

Table 1 Pairs of the RdRp segment of the reoviruses with identities between 54% and 54.5%

Identity	*Same genus?*	*Same species?*	*Genome 1*[a]	*Genome 2*
0.544895	No	No	20279540 Coltivirus|Eyach virus	37514915 Mycoreovirus|Mycoreovirus 3
0.542353	No	No	8574569 Seadornavirus|Banna virus	24286507 Cypovirus|Cypovirus 1|Dendrolimus punctatus cypovirus 1
0.541812	No	No	8574569 Seadornavirus|Banna virus	32470626 Orthoreovirus|Mammalian orthoreovirus|Mammalian orthoreovirus 3
0.541744	No	No	8574569 Seadornavirus|Banna virus	25808995 Orbivirus|Bluetongue virus|Bluetongue virus 2|Corsican bluetongue virus
0.541257	Yes	No	50253405 Orbivirus|St. Croix River virus	50261332 Orbivirus|Palyam virus
0.541096	No	No	22960700 Coltivirus|Colorado tick fever virus	32349409 Mycoreovirus|Mycoreovirus 1|Cryphonectria parasitica mycoreovirus-1 (9B21)
0.540999	No	No	8574569 Seadornavirus|Banna virus	14993633 Cypovirus|Cypovirus 1
0.54089	No	No	14993610 Cypovirus|Cypovirus 14	20177438 Fijivirus|Nilaparvata lugens reovirus

[a] The numbers correspond to sequences in GenBank. The taxonomy lineages from the genus level of the viruses are also shown.

demarcations can be easily determined and new viruses can be clearly placed into the correct taxonomy. However, there are times when PASC alone cannot give a definite classification, and other viral properties have to be considered.

Compared with another quantitative approach, phylogenetic analysis, PASC is less computationally intensive and can be easily updated with new sequence data. In addition, PASC results are relatively easier to interpret, which can be potentially done by a computer program without human intervention. It would therefore be possible to set up an automatic system for high throughput classification.

Many researchers use BLAST to search sequence databases to find best matches for viral sequences of interest. Although BLAST is readily available and easy to run, it is not the best tool for virus classification. This is because BLAST is a local alignment program, and, as discussed above, it may not take highly variable regions into account when calculating identities. Even when an output from BLAST covers the whole sequence as a single alignment, information about the taxonomy relationship between the query virus and the virus closest to it is not immediately available. For example, an identity of 75% could be within the range of the same species in one viral family, but in the range of different species in another family. On the contrary, PASC suggests explicitly whether the query virus is in the same species as some existing viruses, or if it should be assigned within a new species or genus in the family.

The total number of virus sequences in the GenBank/EMBL/DDBJ databases is more than 4 times now than 5 years ago (from about 109 000 to about 446 000 in October 2006). It is possible to test the PASC system on many virus families now. New sequencing technology makes it possible to generate large amounts of virus sequences from environmental samples without the need to isolate and purify virus particles. In such cases, a molecular based method is the only way to classify the viral sequences, and PASC can be very useful for this purpose.

Limitations of PASC

Although PASC has been applied successfully to several virus families, the approach has some limitations.

First of all, PASC is not suitable for virus families whose current classification is largely based on virus morphologies, such as phages in the families *Siphoviridae* and *Podoviridae*. The whole genome PASC may not work well for virus families with highly diverse sequences. This includes viruses with low overall sequence similarities or large differences in genome sizes and organization. For example, in the family *Herpesviridae*, the percentage of identity between different species in the genus *Varicellovirus* ranges from 39% (between cercopithecine herpesvirus 9 and suid herpesvirus 1) to 83% (between bovine herpesvirus 1 and bovine herpesvirus 5), while the percentage of identity between a virus in the genus *Simplexvirus* and one in the genus *Varicellovirus* could be as high as 54% (between cercopithecine herpesvirus 2 and bovine herpesvirus 5). The overlap of such identities makes it impossible to determine the species and genus demarcations for herpesviruses. In the family *Poxviridae*, the largest genome (canarypox virus) is almost 3 times as big as the smallest one (bovine papular stomatitis virus). The huge differences in the genomes sizes will introduce large artifacts when calculating the identities. In such cases, single gene or a cluster of genes needs to be used in PASC instead of whole genomes. The polymerase protein sequences of poxviruses have been used to perform PASC and a good result was obtained. However, not every single gene can be used for PASC. The genes must be present in all viruses of a family and be very conserved. They need to be tested extensively and accepted by the research community as a useful taxonomic criterion before being applied to the PASC system.

Second, for those families whose PASC were constructed with whole genomes, the query sequences to be tested on PASC to determine their taxonomic positions have to be complete genome sequences as well in order to get an accurate prediction. Although a percentage of identity can be obtained when a partial genome is used, the value will be smaller than what it really should be if a complete genome were used because of the way percentage of identity is calculated in this system. In addition, this value obtained with a partial sequence may not reflect the real taxonomic position of this virus, if, for example, recombination is frequent and important for this virus. This will reduce the number of sequences that can be tested by PASC.

Last, it is almost impossible to get identical PASC results when different methods are used. As mentioned above, there are many pairwise sequence alignment programs available. Even within a single program, variation of parameters can affect alignments. After an alignment is obtained, there can be various ways to calculate the distance. For demarcations computed using different types of distances, it may be difficult to choose a rational reason to privilege one demarcation threshold over another. Moreover, once a demarcation is adopted and a researcher uses a different definition to measure the distance between a new virus and an existing one, the comparison with the histogram will sometimes be misleading. These issues can be overcome by using a centralized PASC system where the alignment identities of new viruses are computed using the same algorithm and the same parameter set that were used to compute identities within the family and to create the demarcation.

Conclusion

PASC is a molecular classification tool for many virus families. It calculates the pairwise identities of virus sequences within a virus family and displays their distributions, and can help determine the demarcations at strains, species, genera, and subfamilies level. PASC has many advantages over conventional virus classification methods. The tool has been successfully applied to many virus families, although it may not work well for virus families with highly diverse sequences. The PASC tool at NCBI established distributions of identity for a number of virus families. A new virus sequence can be tested with this system within a few minutes to suggest the taxonomic position of the virus in these families. This system eliminates the potential discrepancies in the results caused by different algorithms and/or different data used by the virology community. Data in the system can be updated automatically to reflect changes in virus taxonomy and additions of new virus sequences to the public database. The web interface of the tool makes it easy to navigate and perform analyses.

See also: Taxonomy, Classification and Nomenclature of Viruses; Virus Databases; Virus Species.

Further Reading

Edgar RC and Batzoglou S (2006) Multiple sequence alignment. *Current Opinion in Structural Biology* 16: 368–373.

Fauquet CM, Mayo MA, Maniloff J, Desselberger U and Ball LA (eds.) (2005) *Virus Taxonomy, Classification and Nomenclature of Viruses: Eighth Report of the International Committee on the Taxonomy of Viruses.* London: Academic Press.

Felsenstein J (2004) *Inferring Phylogeny.* Sunderland, MA: Sinauer Associates.

Gusfield D (1997) *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology.* Cambridge: Cambridge University Press.

Page RD and Holmes EC (1998) *Molecular Evolution: A Phylogenetic Approach.* Oxford: Blackwell Science.

van Regenmortel MHV (2007) Virus species and virus identification: Past and Current Controversies. *Infection, Genetics and Evolution* 7: 133–144.

van Regenmortel MHV, Bishop DH, Fauquet CM, *et al.* (1997) Guidelines to the demarcation of virus species. *Archives of Virology* 142: 1505–1518.

Relevant Website

http://www.ncbi.nlm.nih.gov – NCBI, PASC.