Building (Viral) Phylogenetic Trees Using a Maximum Likelihood Approach

Kelly M. King[1] and Koenraad Van Doorslaer[1,2,3]

[1]School of Animal and Comparative Biomedical Sciences, University of Arizona, Tucson, Arizona
[2]Department of Immunobiology, Cancer Biology Graduate Interdisciplinary Program, Genetics Graduate Interdisciplinary Program, BIO5 Institute, and the University of Arizona Cancer Center, University of Arizona, Tucson, Arizona
[3]Corresponding author: *vandoorslaer@email.arizona.edu*

Phylogenetic analyses allow for inferring a hypothesis about the evolutionary history of a set of homologous molecular sequences. This hypothesis can be used as the basis for further molecular and computational studies. In this unit, we offer one specific method to construct a Maximum Likelihood phylogenetic tree. We outline how to identify homologous sequences and construct a multiple sequence alignment. Following alignment, sequences are screened for potentially confounding factors such as recombination and genetic saturation. Finally, a Maximum Likelihood phylogenetic tree can be constructed implementing a rigorously tested model of evolution. The workflow outlined in this unit provides sufficient background for inferring a robust phylogenetic tree starting from a particular gene of interest. © 2018 by John Wiley & Sons, Inc.

Keywords: evolution • maximum likelihood • phylogeny • saturation • virus

How to cite this article:
King, K. M., & Van Doorslaer, K. (2018). Building (viral) phylogenetic trees using a maximum likelihood approach. *Current Protocols in Microbiology*, *51,* e63. doi: 10.1002/cpmc.63

INTRODUCTION

In (molecular) phylogenetic inference (a.k.a. tree building), scientists attempt to deduce the evolutionary relationship between genes/nucleotide sequences or proteins to gain insight into how these molecular sequences evolved. By understanding how extant sequences came to be, we may be able to predict how these genomes could evolve in the future. These phylogenetic trees are a visual representation of mathematical constructions depicting evolutionary relationships under a given set of parameters. Importantly, there is no cookbook approach to phylogenetic analysis. The optimal strategies and settings are highly dependent on the available data and the biological question. It is, however, possible to divide a typical phylogenetic tree-building experiment into several distinct steps (Fig. 1). Following multiple sequence alignment (Basic Protocol 2), the alignments are screened for factors that may confound the tree building process [e.g., genetic saturation (Support Protocol 4) or recombination (Basic Protocol 3)]. Before estimating a molecular phylogeny, one must determine whether amino acid or nucleotide sequences are more suitable for the analysis. To select the appropriate type of sequence, the evolutionary distance between the genes or species of interest should be taken into consideration. DNA sequence alignments are better suited to evaluate shorter evolutionary distances (i.e., genes from closely related or same species). At longer timescales,

CURRENT PROTOCOLS
A Wiley Brand

Current Protocols in Microbiology e63, Volume 51
Published in Wiley Online Library (wileyonlinelibrary.com).
doi: 10.1002/cpmc.63

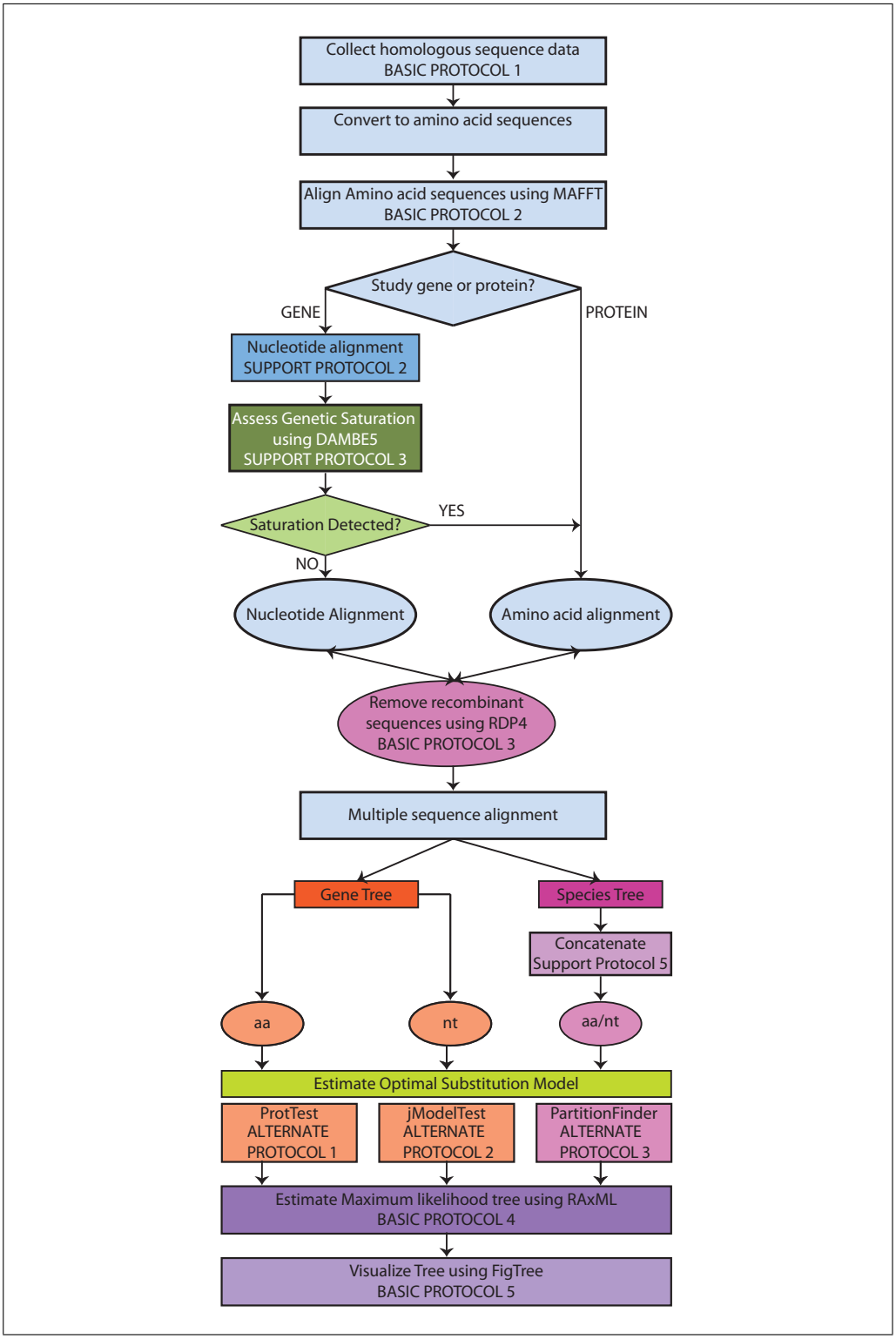

Figure 1 Flowchart describing the protocol for constructing a phylogenetic tree using a maximum likelihood approach. A typical phylogenetic tree building experiment can be divided into distinct steps. The different protocols in this unit are indicated

sequences in a multiple alignment may have undergone so many repeated substitutions that the real genetic distances could be underestimated (i.e., genetic saturation; Philippe et al., 2011). Since nucleotide sequences saturate more rapidly than protein sequences (Philippe et al., 2011), amino acid sequences are a better choice when analyzing evolutionarily more distantly related viruses. In fact, we recommend aligning all

protein-coding sequences at the amino acid level. If closely related viral sequences are being studied, it is possible to convert the protein sequences back into the associated nucleotide sequences (Support Protocol 3). However, if nucleotide sequences are to be used, we recommend that the sequences be tested for genetic saturation before continuing (Support Protocol 4). Next, the optimal model of evolution is estimated. Finally, gene or species trees are estimated under the Maximum Likelihood criterion. Throughout this unit, we assume that the sequence information is derived from protein-coding genes. While similar steps apply when constructing trees of non-coding sequences, all the analysis steps would need to be performed at the nucleotide level. For each of the above steps, we will highlight one of several excellent programs and algorithms available (see *https://evolution.genetics.washington.edu/phylip/software.html* for a list of phylogenetic programs). A deciding factor in choosing some of the programs was their availability on the CIPRES Science Gateway Toolkit (Miller et al., 2015). Importantly, all the software described in this unit is freely available.

COLLECT SEQUENCE DATA

BASIC PROTOCOL 1

This protocol describes the crucial first step of collecting the sequence data to be analyzed. The biological question will typically dictate the type of sequence data to be collected. Sequence data collection is an iterative project, and the effect of adding new data to an alignment should be carefully determined (see Basic Protocol 2). A typical approach to detect homologous genes or proteins uses a BLAST-based database search approach (Altschul, Gish, Miller, Myers, & Lipman, 1990). Importantly, most phylogenetic programs explicitly assume homology (Thornton & DeSalle, 2000). Therefore, if homology (i.e., shared descent) is uncertain, the results of these programs should be treated with caution. Database searches using BLAST attempt to quantify whether two sequences share more similarity than would be expected by chance (Pearson, 2013). When excess similarity is detected, it is assumed that common origin is the most likely explanation, and homology is considered. Importantly, the BLAST algorithm minimizes type I errors (false positives) but cannot control for false negatives (type II errors). The identified homologous sequences can be used to infer phylogenies. A phylogenetic tree based on a single gene or protein will provide insight into the evolution of that genetic element, but this may not be representative of the evolution of the entire virus (Maddison & Wiens, 1997). If the history of the entire virus is of interest, a species tree approach, in which multiple gene (products) are analyzed, may be more appropriate.

Please note that many alternatives to BLAST-based homology searches are available, e.g., FASTA (Pearson, 2016) and HMMER (Finn, Clements, & Eddy, 2011). In addition, many virus-specific databases exist where homologous genes have been curated (e.g., *https://www.oxfordjournals.org/nar/database/subcat/5/18*). If sequences are obtained from these public resources, this protocol can be skipped.

Materials

Nucleotide sequence of gene of interest. Examples of this could be:

(1) New sequence(s) identified by a research group
(2) Sequence(s) downloaded from a curated database (e.g., *https://www.oxfordjournals.org/nar/database/subcat/5/18*)
(3) Computer with Internet connection

Alternatively, a copy of the BLAST database can be installed locally

1. Open the BLAST page on the NCBI website at *https://blast.ncbi.nlm.nih.gov/*.
2. Select the tBLASTx portal.

   tBLASTx will use a translated nucleotide query to search the translated nucleotide database.

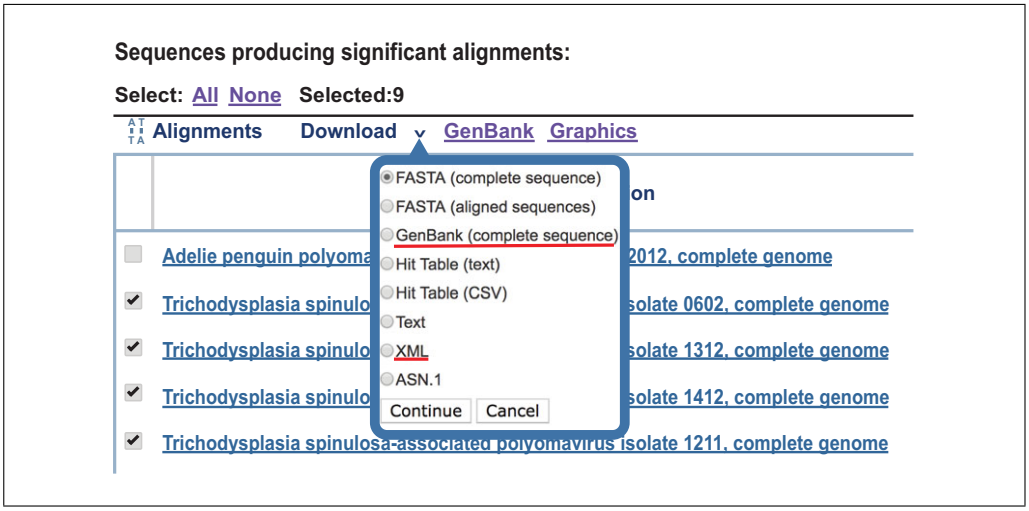

Figure 2 Selection of BLAST hits. Screenshot illustrating the processes involved in downloading the BLAST hits. Use the checkboxes to select the sequences, and download the files in XML and GenBank format (highlighted in red).

3. Copy and paste the nucleotide sequence of your gene of interest (GOI) into the query box or upload the sequence file in FASTA format.

   The first line in a FASTA file starts with a ">" followed by a unique description of the sequence. The next lines contain the actual sequence itself in standard one-letter code.

4. Perform the BLAST search using the default settings.

   Additional information on how to configure searches are beyond the scope of this protocol. However, a recent Current Protocols article describes the use of Blast to find homologs (Ladunga, 2017). Additional details are available through the NCBI website (https://blast.ncbi.nlm.nih.gov/).

5. Filter the list based on expected similarity using a similarity cut-off of 50 bits.

   Percent identity is not a reliable measure of sequence homology. For example, if sequence similarity is used as a proxy for homology, a 30% identity threshold underestimates the number of homologs between humans and yeast by 33%. We recommend using bits to predict homology; for average-length proteins, a bit score of 50 is almost always significant. The NCBI Blast website color codes bit scores between 50 and 80 bits green. Unless a very long protein is being used (or a very large database), 50 bits of similarity is a good rule of thumb for inferring homology in protein alignments (Pearson, 2013).

6. Download selected matches to your computer.

   a. Scroll down to "Sequences producing significant alignments:".
   b. Select the boxes next to the sequences of interest.
   c. Click Download and select XML (Fig. 2).

      Save the file as `Alignment.xml`.

   d. Click Download and select "GenBank (complete sequence)" (Fig. 2).

      Save the file as `sequence.gb`.

7. Convert sequence matches to multi-sequence FASTA file.

   a. Execute the included python script to convert `Alignment.xml` and `sequence.gb` into FASTA-formatted files.

      This python script is included in the `Blast` *subfolder of the* `FigShare` *data folder associated with this protocol. This script has been tested in Python 2.7 and requires Biopython.*

8. If desired, this search can be repeated iteratively to identify a more complete set of more divergent sequences.

Examples in this unit are based on papillomavirus E6 and E7 genes. The E6 and E7 nucleotide sequences for all known human Alphapapillomavirus (de Villiers, Fauquet, Broker, Bernard, & zur Hausen, 2004) and the Bovine papillomavirus (BPV1) were downloaded from the Papillomavirus Episteme (Van Doorslaer et al., 2013, 2017) (https://pave.niaid.nih.gov/; accessed on 1/15/2018). All the input and (intermediate) output files associated with this unit can be found at FigShare, DOI: 10.6084/m9.figshare.6365201.

CREATE USER ACCOUNT ON THE CIPRES SCIENCE GATEWAY

SUPPORT PROTOCOL 1

This protocol provides instructions for creating a user account in the CIPRES Science Gateway V. 3.3 (Miller et al., 2015). Within this portal, many sequence analysis programs used to align sequences and estimate phylogenic trees are freely accessible without the need for download or local installation. Those utilized in this paper include MAFFT (Multiple Alignment using Fast Fourier Transform), v7.3 (Katoh, 2002; Katoh & Toh, 2010), RAxML-HPC v.8 (Randomized Accelerated Maximum Likelihood for High-Performance Computing, version 8) (Pfeiffer & Stamatakis, 2010; Stamatakis, 2014), jModelTest2 (Darriba, Taboada, Doallo, & Posada, 2012; Posada, 2009), and PartitionFinder2 (Lanfear, Calcott, Ho, & Guindon, 2012, 2017).

Materials

Web browser
The CIPRES Science Gateway website at *https://www.phylo.org/*

1. Access the CIPRES Science Gateway (Miller et al., 2015).
2. Click ‘Use the CIPRES Science Gateway’ link on the homepage.
3. Select Register under the CIPRES login.
4. Fill out the required fields and click Register.

Notifications will be sent to the submitted e-mail address.

GENERAL USE OF THE CIPRES SCIENCE GATEWAY

SUPPORT PROTOCOL 2

This protocol walks through the basic steps of running an analysis on the CIPRES Science Gateway. These general steps will apply to all further protocols using the CIPRES Science Gateway.

Materials

Web browser
Functional CIPRES user account (Support Protocol 1)
Protocol-specific data files

1. Login to the CIPRES gateway website (*https://www.phylo.org/*) using the credentials created in Support Protocol 1.
2. Assign a project folder by clicking Create New Folder on the user homepage.
   a. Each project folder contains a `Data` and `Tasks` subfolder (Fig. 3).
   b. Additional folders can be created by clicking the Create Subfolder button.
3. Add files to the `Data` subfolder:
   a. Add data from your local computer.
      i. Click on the Data folder link.

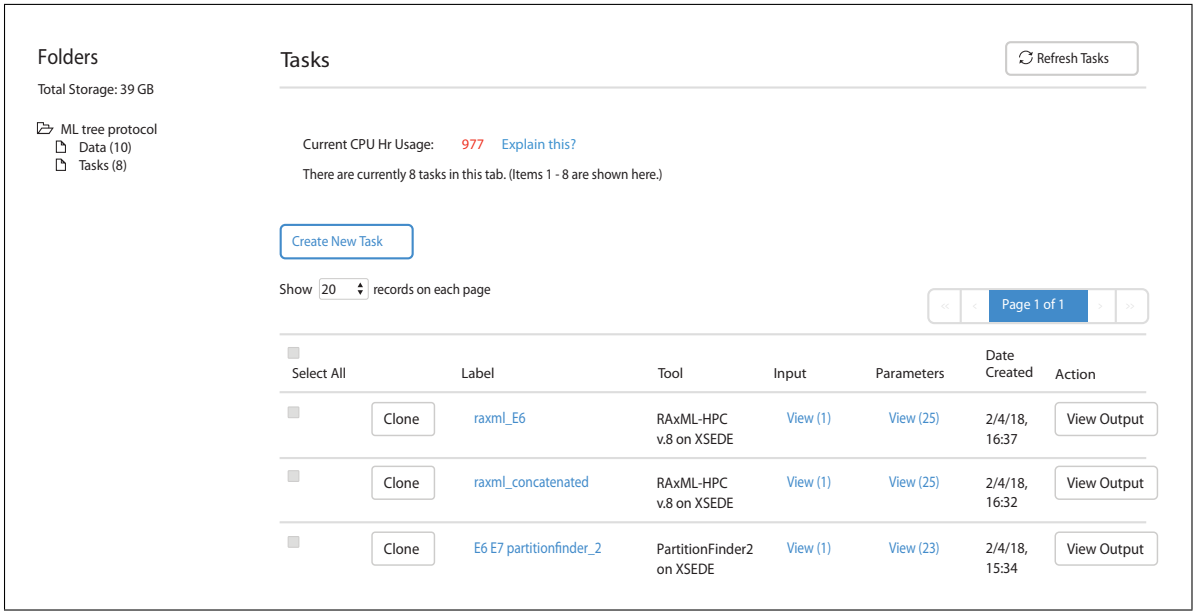

Figure 3 Screenshot demonstrating the general use of the CIPRES Science Gateway.

ii. Click Upload Data.
    1. Provide a name and Description for the file.
    2. Enter the protocol-specific data using one of the following options.

a. Manually copy and paste the data.
b. Click Choose Files to upload data files from your computer.
c. Following conclusion of a CIPRES 'Task' the resulting data can be used for the next analysis.
    i. Navigate to the `Tasks` subfolder and select `View Output`.
    ii. View the files you wish to transfer.
    iii. Click Save To Current Folder → Save.

4. Set up and run a new task:

a. Click on the Tasks folder link.
b. Click Create New Task.
c. Under the 'Task Summary' tab:
    i. Enter a Description of the task.
    ii. Click 'Select input Data' to select the data to be used in the analysis.
    iii. Click Select Tool to select the tool used during the analysis.
    iv. Click Set Parameters to set the tool specific parameters.
        1. The Simple Parameters are displayed by default.

           Many option titles are hyperlinks. When clicked, they will provide more details on each option.

        2. Click the Advanced Parameters for more options.
        3. Click Save Parameters after making changes.
    v. Click Save and Run Task.

       The 'clone' button (Fig. 3) makes it possible to exactly duplicate a task based on a previous analysis. Following cloning, the analysis can be executed, or input data and other parameters can be altered.

5. Track progress of a running task.

a. Click on the Tasks folder link.
b. Click View Status next to the task of interest.

i. The input file can be viewed by clicking the hyperlink.
ii. A list of parameters is provided by clicking the hyperlink.
iii. Intermediate results for each analysis are available through the List hyperlink.
    1. The `stderr.txt` file will contain error messages if anything went wrong with the run.
    2. The `stdout.txt` file will contain information regarding the analysis. This information is similar to that available when the program is run on the local machine.

6. View output files.

    a. At the completion of the task, an e-mail will be sent to the email address on file.
    b. Click on the Tasks folder link.
    c. Click View Output next to the task of interest.
    d. Select the box next to 'Select all' and click Download Selected.

    This will download a zip archive file containing all output data.

    Individual results can be viewed or downloaded directly through CIPRES. Clicking View will allow you to save the output file to the current Data folder, so the output can be used for future analyses.

7. We recommend renaming the output file to a more descriptive file name.

BUILD A MULTIPLE SEQUENCE ALIGNMENT

BASIC PROTOCOL 2

Multiple alignment methods are used in phylogenetic analyses to create initial assessments of sequence homology. At their core, multiple sequence alignments represent a hypothesis on the homology and evolutionary analysis of the sequences. As such, the method used for sequence alignment can dramatically affect the final results. Several methods for multiple sequence alignment are available. A list of examples is available at *https://evolution.genetics.washington.edu/phylip/software.html*. In this protocol, we use MAFFT (Katoh, 2002; Katoh & Toh, 2010). MAFFT has several options for multiple alignment strategies. A more detailed description of alignment algorithms implemented in MAFFT can be found in Table 1.

Table 1 MAFFT Algorithm Options

Algorithm	Method	Description
Auto	NA	Automatically selects from L-INS-I, FFT-NS-I, and FFT-NS-2
E-INS-i	Iterative Refinement	Very slow; best with ≤200 sequences containing multiple conserved domains and large gaps
L-INS-i	Iterative Refinement	Very slow; best with ≤200 sequences containing one conserved domain and large gaps
G-INS-i	Iterative Refinement	Very slow; best with ≤200 sequences with global homology
FFT-NS-I x1000	Iterative Refinement	Slow; 1000 iterations maximum
FFT-NS-I x2	Iterative Refinement	Medium; 2 iterations only
FFT-NS-2	Progressive	Fast; progressive method
FFT-NS-1	Progressive	Very fast; best with >2000 sequences

Materials

CIPRES user account (Support Protocol 1)
File containing amino acid sequences in FASTA format (Basic Protocol 1)

1. Access MAFFT on XSEDE (7.305) through CIPRES.
2. Click 'Select data' and select the amino acid FASTA file (Basic Protocol 1) of the protein of interest.
3. Set the analysis parameters.
   a. Run time can be adjusted based on the size of the input dataset and alignment algorithm.

      The E6 amino acid FASTA file contains 65 sequences with lengths varying between 136 and 214 residues. Set runtime to 30 min.

   b. We recommend outputs in FASTA format, which is the preferred input format for the majority of analyses performed in this protocol.

      Due to a bug in CIPRES, PartitionFinder2 (Alternate Protocol 3) requires a phylip-formatted file. All other protocols in this unit will use FASTA formatted files.

   c. The default `-auto setting` will automatically attempt to select an appropriate alignment strategy based on data size. It is the users' responsibility to confirm that the automatic option is appropriate.

      Use Table 1 to assist in selecting algorithms.
4. CIPRES will name the resulting alignment `output.mafft`. We recommend renaming this file with a unique name.

   We renamed the `output.mafft` files to `PaVE_E6_clean_aa_mafft.fas` and `PaVE_E7_clean_aa_mafft.fas`.
5. If multiple genes will be analyzed, steps 2 to 5 should be repeated for each amino acid alignment.

   If you are building a species tree, make sure that sequences for each alignment are entered in the same order. Furthermore, check the 'maintain input order' checkbox. This step is crucial when concatenating alignments for PartitionFinder 2 (Support Protocol 6 and Alternate Protocol 3).
6. The alignment is a key step in the tree-building process. Please carefully check the quality of the alignment. If available, use biological information during this step. For example, confirm that conserved motifs are aligned properly.

   For example, the papillomavirus E6 protein contains highly conserved cysteine residues. Ensure that these residues are aligned.

SUPPORT PROTOCOL 3

CONVERT AMINO ACID ALIGNMENT TO A NUCLEOTIDE ALIGNMENT

When analyzing protein-coding genes, we recommend aligning amino acid sequences. By aligning proteins, we ensure that codons are not split during the alignment process. However, if nucleotide sequences are of interest, the amino acid sequence needs to be 'back-translated' into DNA. This support protocol uses RevTrans (Wernersson & Pedersen, 2003) to convert a protein sequence alignment into the corresponding DNA sequence alignment.

Materials

Nucleotide sequence file (Basic Protocol 1)
Protein multiple sequence alignment file (Basic Protocol 2)

RevTrans website: *https://www.cbs.dtu.dk/services/RevTrans/* (Wernersson & Pedersen, 2003)

1. Access the RevTrans website (*https://www.cbs.dtu.dk/services/RevTrans/*).
2. Upload FASTA-formatted, unaligned DNA sequences in the first text box.
3. Upload FASTA-formatted, aligned protein sequences in the second text box.
4. Click the 'translate only' button on the far right of the screen.
5. Save the output FASTA file to your local computer.
6. Upload the FASTA-formatted nucleotide sequence to the CIPRES gateway.

USE DAMBE TO TEST FOR GENETIC SATURATION

SUPPORT PROTOCOL 4

Phylogenetic analysis of nucleotide sequences from highly divergent species increases the risk of genetic saturation, leading to an underestimation of actual genetic distances between sequences (Arbogast, Edwards, Wakeley, Beerli, & Slowinski, 2002). Xia and colleagues (Xia, Xie, Salemi, Chen, & Wang, 2003a) developed a method to test whether the DNA sequences have undergone (excessive) substitution saturation. If excessive saturation is identified, nucleotide sequences must be translated to amino acid sequences to proceed with the analysis. This protocol will provide a method to visually appraise saturation, as well as a more robust, statistical method (Xia & Lemey, 2009; Xia, Xie, Salemi, Chen, & Wang, 2003b). The method by Xia is based on entropy in information theory (Xia et al., 2003b). The method compares an observed substitution saturation index (*Iss)* to a critical saturation index (*Iss.c*). Since the *Iss.c* value is based on simulations where the number of operational taxonomic units (numOTUs) is less or equal to 32, DAMBE automatically creates subsamples of 4, 8, 16, and 32 OTUs and tests whether saturation exists for these subsets. In most cases, the *Iss* index will increase with numOTU.

Materials

Functioning installation of DAMBE, available at *https://dambe.bio.uottawa.ca/DAMBE/dambe.aspx*
Nucleotide multiple sequence alignment (Support Protocol 3).
Text editor

NOTE: Do not use Microsoft Word or similar as a text editor. Word processing programs add hidden data and characters.

1. Access the DAMBE website and download the software package for either MacOS, Linux or Windows.
2. Open DAMBE 6 (Xia, 2017) and import sequences by clicking File → Open standard sequence file.

   When browsing for sequence files, set the file type to 'Unknown (*)' in order to view and select MAFFT output files.

3. Select "Protein-Coding Nuc. Seq."
4. Select "Standard (Trans-Table=1)."
5. Click "Go!".

   If the only stop codons are terminal, you can ignore the warning about embedded stop codons. Otherwise, check your sequences.

6. Select Seq. Analysis → Substitution rate over site → Estimate proportion of invariant sites.

Table 2 Human Papillomavirus E6 Nucleotide Sequences are Saturated[a],[b]

NumOTU	Iss	Iss.cSym	T	DF	P	Iss.cAsym	T	DF	P
4	0.962	0.806	5.725	646	0.0000	0.775	6.863	646	0.0000
8	1.057	0.766	7.313	646	0.0000	0.657	10.07	646	0.0000
16	1.184	0.746	8.052	646	0.0000	0.536	11.896	646	0.0000
32	1.337	0.719	8.72	646	0.0000	0.394	13.322	646	0.0000

[a]Analysis performed on all sites. Testing whether the observed Iss is significantly lower than Iss.c. IssSym is Iss.c assuming a symmetrical topology. IssAsym is Iss.c assuming an asymmetrical topology.
[b]Note: Two-tailed tests are used.

Table 3 Guide to Interpreting DAMBE Output

Interpretation of results:

	Significant Difference	
	Yes	No
Iss < Iss.c	Little saturation	Substantial saturation
Iss > Iss.c	Useless sequences	Very poor for phylogenetics

a. Choose the 'Use new tree' option.
b. Select "Neighbor-Joing" (sic) as the Tree-building algorithm.
c. Change the genetic distance to 'GTR'.
d. Choose an outgroup for the dataset and click 'Run'.

If you are following along with our example, select BPV1 as the outgroup.

e. Take note of invariant site proportion.

For the E6 alignment file P(invariant) = 0.03009 (since this is an estimate, this value may differ in your analysis).

7. Select Seq. Analysis → Measure Substitution Saturation (Nuc. seq. only) → Test by Xia et al. (Xia et al., 2003a).

a. Note the warning with regard to the effects of sites containing gaps on the sensitivity of the method.
b. Click "ok".
c. Enter the invariant site proportion calculated above.

8. Select Edit → Copy to excel.
9. Open your favorite text editor and paste the data.

The example output for the HPV E6 alignment is shown in Table 2.

Use Table 3 (Xia et al., 2003a) for assistance interpreting the results.

In our HPV E6 example, the Iss (0.394) of a symmetrical tree with 32 taxa is statistically greater than the Iss.c (0.719) (Table 2). These sequences appear to have undergone significant genetic saturation and should not be used for phylogenetic inference.

10. Select Graphics → Transitions and transversions versus divergence.
11. Select GTR under Genetic Distance.
12. Click 'Go!'.
13. Select Edit → Copy to excel.
14. Open your favorite text editor and paste the data.

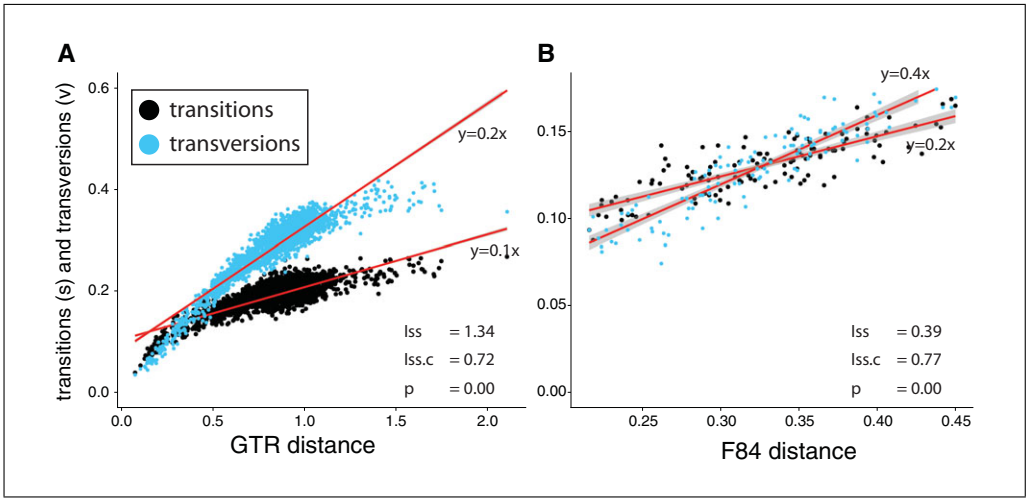

Figure 4 Example of genetic saturation curves. Plotting the observed transitions (black) and transversions (blue) against a model corrected genetic distance allows for visual analysis of genetic saturation. The slope of the linear regression through these plots approaches 1 for unsaturated samples. The linear regression was performed using the "lm" function in R. The results of the formal saturation test by Xia et al. (2003a) are also shown. (**A**) Compares the transition and transversion curves (with linear regression) of the Alphapapillomavirus E6 nucleotide sequences against the GTR corrected distance. (**B**) Example of an analysis based on an alignment of 16 invertebrate EF1a sequences (Xia et al., 2003a). This dataset is included with DAMBE. Based on this analysis, the EF1a sequences do not appear to be genetically saturated.

15. Plot and visualize the data using your favorite graphing software package (Fig. 4).

    The slope of the regression line can be used to estimate the extent of saturation. Alignments without saturation approach 1, while sequences with extensive saturation have a slope of zero. Figure 4A shows the HPV E6 data, while an example of a non-saturated sequence is shown in Figure 4B.

TEST FOR RECOMBINATION

BASIC PROTOCOL 3

While genetic recombination is an important evolutionary process for many viruses, recombination within and between a set of molecular sequences can dramatically affect the reliability and accuracy of the resulting phylogenetic trees (Posada & Crandall, 2002; Schierup & Hein, 2000). RDP4 (Recombination Detection Program, version 4) attempts to detect and visualize recombination in viral genome sequence alignments (Martin, Murrell, Khoosal, & Muhire, 2017). Like other recombination detection programs, RDP4 analyzes nucleotide sequences for the presence of recombinant sequences. We recommend removing these recombinant sequences or regions from the multiple sequence alignments to minimize the overall impact on the downstream tree building process.

Materials

RDP4 homepage at *https://web.cbio.uct.ac.za/~darren/rdp.html*
Nucleotide multiple sequence alignment file (Support Protocol 3)
RDP4 requires a nucleotide sequence alignment as input

1. Access the RDP homepage at *https://web.cbio.uct.ac.za/~darren/rdp.html*. Download RDP4 for either MacOS (with PlayOnMac), Linux (with Wine), or Windows.
2. Follow the installation instructions.
3. Open RDP4.

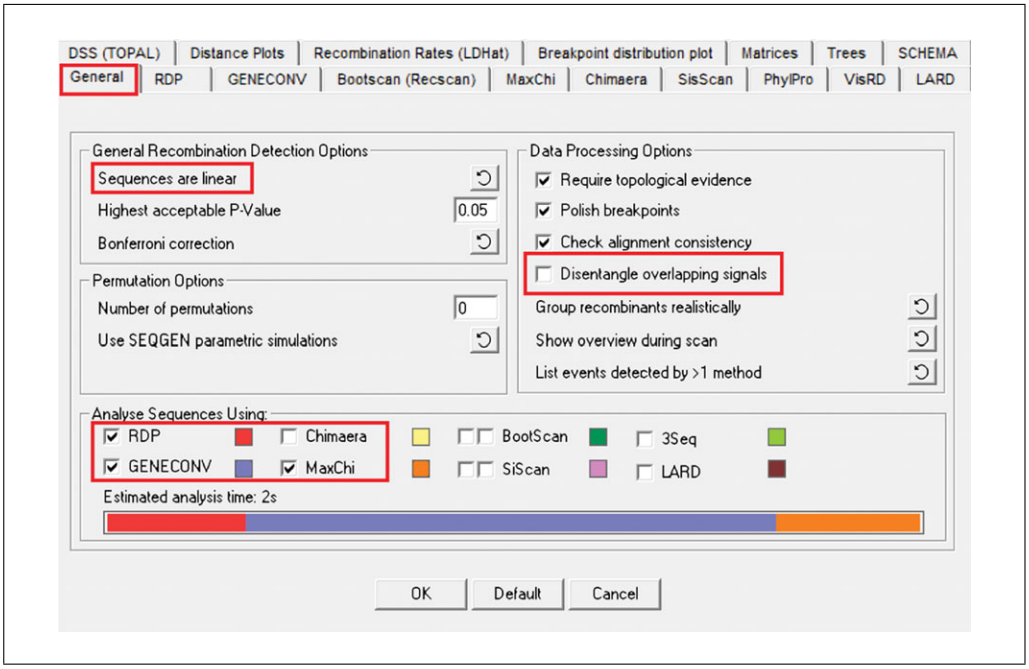

Figure 5 Setting basic parameters for the initial recombination screen using RDP4. Screenshot of analysis options in RDP4. The options that differ from default setting are highlighted in red.

4. Import the generated nucleotide multiple sequence alignment from Support Protocol 3. Select Open → choose downloaded multiple sequence alignment file.

5. Detect Recombination (see Figure 5 for details on parameter settings).

   a. Select Options → General → General Recombination Detection Options.
   b. Specify that the sequences are linear or circular—depending on genome type.
   c. "Analyse Sequences using:" RDP (Martin & Rybicki, 2000), GENECONV (Padidam, Sawyer, & Fauquet, 1999), and MAXCHI (Smith, 1992).
   d. Under Data Processing Options, unselect "Disentangle overlapping events"; leave the other settings as default.

   i. If desired, the "window and step sizes" for specific methods can be changed. Please refer to the RDP4 manual for more details on this setting.

   e. Select OK to return to the main window.
   f. Run the analysis for a full exploratory recombination scan using all methods selected in the General options tab by clicking on Run.

6. It is important that you carefully analyze the recombination hypothesis that the program provides. For example, RDP4 will provide a warning if it cannot determine which sequence is the recombinant and which are the parental donor sequences. Please refer to the RDP4 manual for detailed instructions on how to screen the initial recombination hypothesis.

   A detailed step-by-step method is also available in a recent protocol published by Salminen and Martin (2010).

7. If the proposed recombination hypothesis was updated in step 6, it is important that the alignment be reanalyzed by clicking "rescan."

8. Remove the identified recombinant sequences from the alignment.

   RDP4 will identify the recombinant sequence(s) in the nucleotide alignment. If the phylogenetic tree will be estimated based on the protein sequences, make sure to also remove the identified sequence from the amino acid alignment if needed. If you are interested in

building a species tree, the viral sequence should be removed from each partition prior to concatenation.

No recombination was detected in the example papillomavirus E6 or E7 sequences.

9. Realign the sequences as described in Basic Protocol 2.

INSTALLATION OF PROTTEST3

SUPPORT PROTOCOL 5

Materials

Java v.8 (or better), available at *https://www.java.com/en/download/win10.jsp*
ProtTest 3, available at *https://github.com/ddarriba/prottest3/releases*

1. Access the Java website and follow the links to download the latest version (containing Java SE).
2. Access the website containing the ProtTest 3.4.2 file.
   a. Download the TAR file: `prottest-3.4.x-xxxxxxxx.tar.gz`.
   b. Extract files to a local folder.
3. Within the extracted folder:
   a. Extract the ProtTest Java file.
   b. Access the `prottest-3.4.x` folder.
   c. Extract the GZ file named `mpj.tar`.
   d. Return to the `prottest-3.4.x folder`.
4. Run the program:
   a. Windows users:
      • Double click 'runXProtTestHPC' to open ProtTest.
   b. MacOS/Linux users
      • open Terminal.
      • cd into the extracted prottest folder.
      • type `./prottest3`.

CONCATENATE MULTIPLE SEQUENCE ALIGNMENTS

SUPPORT PROTOCOL 6

When building a species tree from either nucleotide or amino acid sequences, individual multiple sequence alignments (Basic Protocol 2) must be concatenated and used as input for PartitionFinder2 (Alternate Protocol 3). This protocol uses the freely accessible FASTA alignment joiner, but many other options are available.

Materials

Amino acid or nucleotide multiple sequence alignment files (Basic Protocol 2) that have been tested for recombination (Basic Protocol 3)

Ensure that the sequences occur in the exact same order in each alignment!

FASTA alignment joiner (*https://users-birc.au.dk/biopv/php/fabox/alignment_joiner.php*)

1. Access the FASTA alignment joiner website.
2. Upload the first multiple sequence alignment into the first textbox.
3. Upload the second multiple sequence alignment into the remaining box.
4. Select 'Join Alignments'.

a. Download the concatenated sequence file (in FASTA format).
b. Upload concatenated alignment file into the CIPRES data folder.
c. If more than two alignments will be concatenated, repeat steps 1 to 4a by adding new alignments to the previously concatenated version.

IDENTIFYING THE OPTIMAL EVOLUTIONARY SUBSTITUTION MODEL

The maximum likelihood method requires a substitution model to statistically estimate the probability of particular mutations within a multiple sequence alignment. These substitution models represent a specific hypothesis about relative mutation rates of the aligned sequences. It is important that the correct model be chosen. The assumptions of overly restrictive models are easily violated, while over-parameterized models may be too computationally expensive (Sullivan & Joyce, 2005). The following protocols use automated model selection programs to choose the simplest model that is not significantly worse than a more complex option. ProtTest (Alternate Protocol 1; Darriba, Taboada, Doallo, & Posada, 2011) and jModelTest2 (Alternate Protocol 2; Darriba et al., 2012; Posada, 2009) are used to identify the best substitution model for a single protein or DNA sequence, respectively. However, when constructing species trees, multiple gene or protein alignments are considered. Each of these partitions may be evolving under a different substitution model. *A priori*–specified data blocks (typically representing different genes in the concatenated alignment) can be individually analyzed to generate a partitioning scheme containing the best substitution model for each data block. Partitioning a concatenated alignment into individual proteins is common. Similarly, a codon alignment can be partitioned allowing each codon position to evolve according to a different model. By selecting the optimal partitioning scheme, PartitionFinder (Alternate Protocol 3) aims to increase the accuracy of phylogenetic trees when building species trees (Lanfear, Frandsen, Wright, Senfeld, & Calcott, 2012, 2017). By optimizing the model for each partition, the algorithm accounts for variation in substitution that is observed across different genes/proteins in the multiple sequence alignment, increasing the accuracy of phylogenetic trees.

All three programs implement maximum likelihood criteria in their analysis to find the optimal evolutionary model. For more background information and details on the different models, we refer to the phylogenetic handbook (Lemey, Salemi, & Vandamme, 2009).

ALTERNATE PROTOCOL 1

DETERMINE THE AMINO ACID SUBSTITUTION MODEL USING PROTTEST3

Materials

Amino acid multiple sequence alignment (Basic Protocol 2)
Functioning copy of ProtTest 3 (Support Protocol 5)
Text editor

1. In ProtTest select File → Load Alignment. Upload multiple sequence alignment.
2. Select 'Analysis' → 'Compute likelihood scores'.
   a. Do not alter the default options.
   b. Click Compute and a Running status window will appear.

      You can observe the progress of the analysis in the main ProtTest window under the Phyml-log tab.
3. Following completion of the analysis, select 'Selection' → 'Results'.

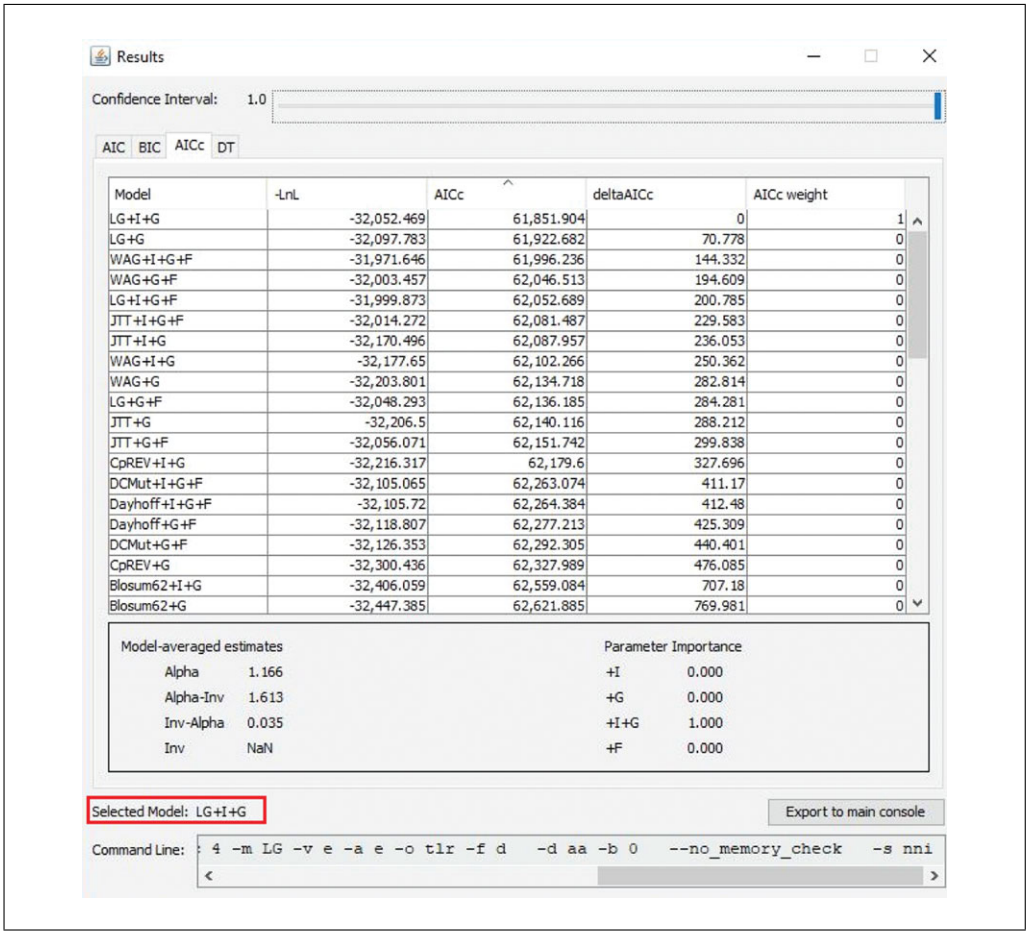

Figure 6 Example of ProtTest analysis. The Alphapapillomavirus E6 amino acid alignment was analyzed using ProtTest3 as described in Alternate Protocol 1. The table lists the best models of substitution ranked according to the AICc (corrected Akaike information criterion) estimator. The selected model is highlighted in red.

4. Four criteria (AIC, AICc, BIC, DT) (Luo et al., 2010) can be used to estimate the relative quality of statistical models for a given set of data. The AICc (corrected Akaike information criterion) corrects for small sample size, decreasing the risk of overfitting the model (Brewer et al., 2016; Anderson & Burnham, 2002). As the sample size gets larger, AICc approaches AIC. We therefore recommend the use of the AICc criterion to select the optimal model. Models are sorted into four tabs corresponding to each type of analysis criterion (AIC, AICc, BIC, DT). The optimal model is listed at the top of the list (Fig. 6).

   Note that the BIC (Bayesian information criterion) should be used if the eventual phylogenetic tree will be based on Bayesian statistics.

5. Click ‘Export to main console’.

6. Save the output of the run by selecting ‘File’ → ‘Save Console’.

7. Open the saved log file in a text editor and locate the line “Best model according to AICc:”. This will identify the optimal model.

8. Open a text editor and enter the selected model using the following scheme (see Alternate Protocol 3 for more details).

   For example, if the viral E6 protein alignment is 122 amino acids in length and evolves according to the LG model, the text file should contain:

   ```
   LG, E6 = 1-222
   ```

When entering the model information, ensure that RAxML-compatible models are entered. If Invariant or discrete gamma models (+I,+G respectively) are selected, these options should be manually entered using CIPRES (Basic Protocol 4).

9. Save the text file. This file will be used in Basic Protocol 4.

ALTERNATE PROTOCOL 2

DETERMINE THE NUCLEOTIDE SUBSTITUTION MODEL USING jModelTest2

Materials

CIPRES user account (Support Protocol 1)
Nucleotide multiple sequence alignment (Support Protocol 3)
Text editor

1. Select 'jModelTest2 on XSEDE'.
2. Select the Nucleotide multiple sequence alignment (Support Protocol 3) as data.
3. We recommend the following analysis parameters:
   a. "Set the number of substitution schemes (-s)": to 11.
      This selects all named substitution schemes.
   b. "Select Information Criterion (choose all needed)": Hold down Shift to select all options.
4. Download the output files.
5. The results from the analysis are detailed in the `stdout.txt` file.
6. Open the `stdout.txt` file in a text editor and search for "Model selected:"

   For the papillomavirus E6 gene, the GTR+I+G model of evolution was selected.

7. Open a text editor and enter the selected model using the following scheme (see Alternate Protocol 3 for more details).

   For example, the viral E6 alignment is 669 bp long and evolves according to the GTR+I+G model. Therefore, the text file should contain:

   ```
   DNA, E6 = 1-669
   ```

   When entering the model information, ensure that RAxML compatible models are entered. If Invariant or discrete gamma models (+I,+G respectively) are selected, these options should be manually entered using CIPRES (Basic Protocol 4).

8. Save the text file. This file will be used in Basic Protocol 4.

ALTERNATE PROTOCOL 3

DETERMINE THE OPTIMAL PARTITIONING SCHEME USING PartitionFinder2

Materials

CIPRES user account (Support Protocol 1)
Concatenated amino acid or nucleotide multiple sequence alignment (Support Protocol 6)
Start and end positions of each partition (i.e., gene or protein) in the concatenated alignment
Text editor

1. At the time of writing, a bug in CIPRES does not convert an input FASTA file to a properly formatted phylip file needed for PartitionFinder2.

   The NCLconverter app on CIPRES can perform the conversion from FASTA to phylip. Save the phylip-formatted file to perform the analysis.

2. Select 'PartitionFinder2 on XSEDE'.

3. Select the Concatenated amino acid or nucleotide multiple sequence alignment as data.

4. Analysis parameters:

   a. Select whether you are analyzing DNA, protein, or morphology data.
   b. Create the partition file under Advanced Parameters:
      i. "Linked branch lengths?". Select Unlink.
      ii. "Select the model": All.
      iii. "Select the metric for model (model selection)": AICc.
      iv. "Select the search algorithm" All.
      v. "How many data blocks do you have?" Enter the number of genes or proteins.
      vi. For each data block:
         1. Enter the name of your first data block.
         2. Enter the beginning of the range.

            This is the position of the first nucleotide (amino acid) of the gene (protein) in the concatenated alignment.

         3. Enter the end of the range.

            This is the position of the last nucleotide (amino acid) of the gene (protein) in the concatenated alignment.

         4. If working with codon data, checking the "This a codon analysis (will repeat the range/1,/2,and/3)" option will further partition each gene into distinct codon positions.
         5. Alternatively, you can manually create a partitioning file (cgf text file). Check the PartitionFinder2 Manual for directions.

            Note that the alignment file must be named `infile.phy`.

5. The results from the analysis are stored in the `analysis.zip` folder.

   a. Download the folder and extract files to a local folder.
   b. Open `best_schemes.txt` to view the best partitioning scheme determined for the data.

6. Locate the 'RAxML-style partition definitions' section to find the best model for each partition.

   a. If you are following along with the papillomavirus E6-E7 example, the output from the PartitionFinder2 analysis will contain the following text:

   ```
   RAxML-style partition definitions
   Warning: RAxML allows for [...] Note that these re-runs
   will be quick!
   JTTF, Subset1 = 1-222, 223-371
   ```

7. Open a text editor and copy and paste the optimal scheme.

   When entering the model information, ensure that RAxML compatible models are entered. RAxML requires that all models in a partitioned analysis be assigned to the same type of model: No heterogeneity (e.g., GTR); +G (e.g., GTR+G); or +I+G (e.g., GTR+I+G). It is possible that the rate heterogeneity models differ between partitions in your dataset.

In this case, it is recommended that you run separate PartitionFinder analyses for each type of rate heterogeneity and select the scheme with the lowest AICc score (Lanfear et al., 2017).

The best model for the Alphapapillomavirus E6 (positions 1 to 222 in the alignment) and E7 (positions 223 to 371) was estimated to be *"JTT+I+G+F" for both partitions*.

Therefore, the corresponding text file should contain:

```
JTTF, Subset1 = 1-222, 223-371
```

8. Save the text file. This file will be used in Basic Protocol 4.

BASIC PROTOCOL 4

MAXIMUM LIKELIHOOD PHYLOGENETIC ANALYSIS

Once a substitution model has been selected (see Alternate Protocols 1 to 3), a typical maximum likelihood analysis will propose many different hypotheses about the evolutionary history of the molecular sequences. These trees are evaluated based on the likelihood that the hypothesis (i.e., the tree) would give rise to the sequence data. The maximum likelihood tree is the hypothesis with the highest probability (Strimmer & von Haeseler, 1997; Schmidt and von Haeseler).

This protocol will use RAxML to compute phylogenetic trees. RAxML infers maximum likelihood phylogenetic trees in a comparatively short amount of time. RAxML-HPCv8 is accessible through the CIPRES Science Gateway.

Materials

CIPRES user account (Basic Protocol 1)
Amino acid or nucleotide multiple sequence alignment (Basic Protocol 2)
Text file containing the best fitting substitution model or partitioning scheme (see Alternate Protocol 1, 2, and 3,)

1. Upload the best-fitting model file to the data folder in CIPRES.
2. Select "RAxML-HPC v.8 on XSEDE - NEW Interface!".
3. Select the alignment file as data.
4. Set individual parameters under the Set Parameters tab.
   a. Set value for "Maximum Hours to Run".
      For the papillomavirus example, set this to one hour.
   b. "Set a name for output files (−n)": Set a name for output files.
   c. "Outgroup (−o) () one or more comma-separated outgroups, see comment for syntax)": Select an outgroup if desired.
      Set BPV1 as outgroup for example analysis.
   d. "Specify a random seed value for parsimony inferences (−p)": check the box.
   e. "Enter a random seed value for parsimony inferences (−p "value" gives reproducible results from random starting tree)": enter a random seed value (e.g., `12345`).
   f. Select the uploaded model file (Alternate Protocols 1 to 3) from the "Use a mixed/partitioned model? (−q)" pull down menu. If used, this option will set the evolutionary model.
   g. Depending on the model, select whether you want to "Estimate proportion of invariable sites (GTRGAMMA + I)".

Note that using a model including invariable sites (+I) is not recommended by the author of RAxML. If this option is selected, CIPRES will produce a warning. If estimation of invariable sites is important to your analysis, this warning can be ignored.

h. Click advanced parameters.
i. "Please select the Data Type": Select Protein or nucleotide from the pull-down menu.
   i. When analyzing DNA: select 'GTRGAMMA' to choose a model for the bootstrapping phase for nucleotide alignment analysis.
   ii. When analyzing protein: select 'Protein GAMMA' for protein alignment analysis.
j. Under 'Configure the Analysis':
   i. From the 'Select the Analysis Type' pull-down menu, select 'Conduct a rapid Bootstrap analysis and search for the best-scoring ML tree in one single program run. (−f a)'.
k. Under Configure Bootstrapping select:
   i. Conduct rapid bootstrapping? (−x).
   ii. Enter a random seed value for bootstrapping (e.g., `12345`).
   iii. "Specify bootstrap protocol" select "Let RAxML halt bootstrapping automatically".
   iv. "Select Bootstopping Criterion: (autoMRE is recommended)": autoMRE.

5. The phylogenetic tree with support values is stored in the `RAxML.bipartitions.result` file.

VISUALIZE PHYLOGENETIC TREE

BASIC PROTOCOL 5

The maximum likelihood phylogeny will be visualized and annotated using FigTree.

Materials

Phylogenetic analysis of multiple sequence alignment, computed by RAxML-HPCv8 (Basic Protocol 4)
Java v8 (or better), available at *https://www.java.com/en/download/win10.jsp*
FigTree, available at *https://tree.bio.ed.ac.uk/software/figtree/*

1. Open FigTree v.1.4.3.
2. Import phylogenetic tree analysis data files from Basic Protocol 4 by selecting File → Open.
   a. Upload `RAxML bipartitions.result` file from Basic Protocol 4.
3. Change `label` to `bootstrap` in the Input window.
4. Select the additional drop-down menus from the left panel to adjust treeview options as desired, for example:
   a. Under Tip Labels, Increase the font size
   b. Under Node Labels → Display Bootstrap.

      Assuming a symmetric phylogeny with equal rates of change, bootstrap proportions of ≥70% roughly correspond to a probability of ≥95% that the associated clade is real (Hillis & Bull, 1993).

   c. Under Layout: Tree orientation and layout can be adjusted by clicking the icons.
5. Select File → export to SVG.
6. SVG file can be opened and edited using Adobe Illustrator (or comparable graphics software).

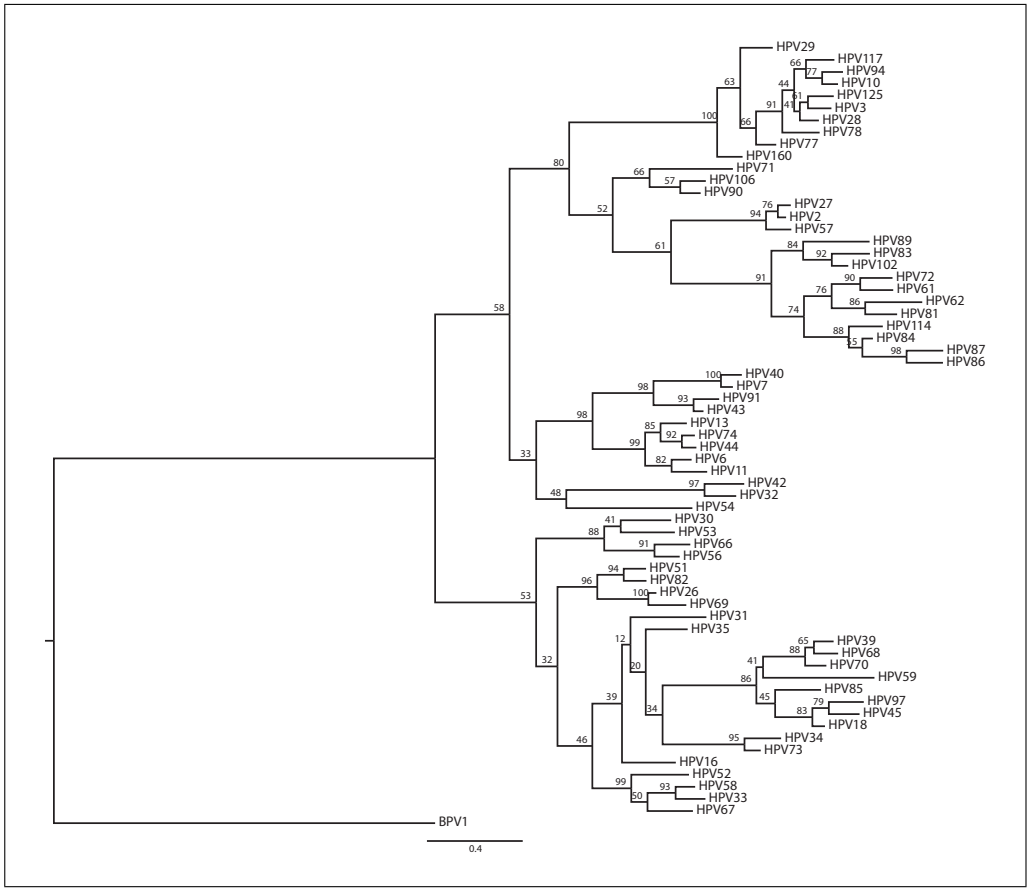

Figure 7 Maximum Likelihood phylogenetic tree representing the evolutionary history of the of the Alphapapillomavirus E6 and E7 proteins. The E6 and E7 protein sequences were aligned using MAFFT and concatenated. A Maximum Likelihood phylogenetic tree was constructed using RAxML, implementing the partitioning scheme determined by PartitionFinder2. The tree was rooted on the bovine papillomavirus 1 (BPV1). The final tree was visualized in FigTree and further edited using Adobe Illustrator. The numbers near the nodes represent bootstrap support. The scale bar represents 0.4 amino acid substitutions per site.

7. Save phylogenetic tree to a local folder.
8. Figure 7 shows a rooted phylogenetic tree showing the evolutionary relationship between the Alphapapillomavirus based on a partitioned E6 and E7 protein alignment.

COMMENTARY

Background Information

In the past couple of decades, analysis of molecular sequences has mostly replaced morphological data for assessing evolutionary relationships. Essentially, phylogenetic trees attempt to construct a hypothesis that provides the best explanation for the available (sequence) data. Over the years, these inference methods have become more accurate, but also more computationally demanding.

UPGMA (unweighted pair group method with arithmetic means) is probably the oldest method used to reconstruct a phylogenetic tree, and is based on progressively identifying the smallest value in the pairwise distance matrix. The Neighbor-Joining method builds a tree by identifying pairs of taxa that are connected by a single interior node. Given a multiple sequence alignment, maximum parsimony identifies the tree topology that requires the fewest substitutions. Maximum likelihood methods attempt to identify the tree that, given a specific model of evolution, maximizes the probability of observing the sequences in the alignment. Finally, like Maximum Parsimony and Maximum Likelihood, Bayesian methods use an optimality criterion. However, unlike the other approaches, Bayesian phylogenetics identifies a distribution of likely trees, not a single 'optimal' tree (Vandamme, 2009).

This unit uses maximum likelihood to identify the optimal tree. The use of Maximum

Likelihood methods in phylogeny was pioneered in the seminal PHYLIP package by Joe Felsenstein (Felsenstein, 1981). New implementations of the maximum likelihood search algorithms have improved their ability to handle large datasets. In addition to RAxML, many popular software packages exist. A recent article compares the performance of RAxML to other popular programs (Nguyen, Schmidt, von Haeseler, & Minh, 2015).

Critical Parameters

In this unit, we provide different protocols aimed at building a phylogenetic tree. For each protocol in the unit (Fig. 1), we use a specific software package. While these are great algorithms, for most steps, there are alternative options. It is important to remember that different algorithms make distinct assumptions and that it is, in the end, the responsibility of the researcher to identify the best approach. We have selected software that is actively being developed and updated on a regular basis. We strongly urge readers to use the most up-to-date version of all scientific software. Similarly, while we tried to minimize the use of websites for performing certain tasks, websites may not be continually maintained.

Multiple sequence alignment is the foundation on which the remaining analyses are performed. It is critical that this alignment, a homology hypothesis, be properly analyzed before proceeding. Attempts at providing objective measures to assess the quality of an alignment have been made (Ahola, Aittokallio, Vihinen, & Uusipaikka, 2006; Sadreyev & Grishin, 2008; Tomovic & Oakeley, 2007). However, users will have to employ their biological knowledge and experience to decide whether an alignment is (biologically) meaningful.

Statistical Analyses

Maximum likelihoods attempt to estimate the optimal tree, but likely will not identify the absolute best tree. Thus, it is important to know how much, if any, confidence we should have in the tree. We can use statistical approaches to assess the reliability of a certain cluster. The bootstrap, initially introduced in phylogenetics by Felsenstein (1985), provides a level of "confidence" for each individual clade of an observed tree. In practical terms, a phylogenetic tree is initially estimated based on an alignment with m sequences of length n. During each bootstrap resampling, n nucleotides are randomly chosen from each sequence with replacements, thus generating a new alignment with m rows and n columns. This resampled alignment is used to construct a new tree, which is compared to the original tree. This resampling is repeated and the percentage of times each interior branch matches the original tree is reported. While bootstrap support values are typically added to a node, they refer to splits of the tree and not to the actual nodes of the tree. Importantly, it has been proposed that the bootstrap is consistently too conservative (Hillis & Bull, 1993). There is no consensus on a minimum bootstrap value to support a split. However, even a bootstrap value of 100% only reflects the phylogenetic signal as interpreted by the method that was used. For example, it is possible to get high bootstrap support even if systematic errors were made during the tree reconstruction (e.g., sequences are not representative). Therefore, a high bootstrap indicates support for a specific cluster but makes no claims about the (biological) validity of the phylogenetic tree. A plethora of alternative support tests have been proposed (Anisimova, Gil, Dufayard, Dessimoz, & Gascuel, 2011). Among these, the nonparametric test, the SH-aLRT (Shimodaira–Hasegawa approximate likelihood ratio test; Guindon et al., 2010; Shimodaira & Hasegawa, 1999) is extremely fast, while maintaining statistical power. This test is implemented in RAxML and can be performed through CIPRES.

Understanding Results

The robust statistical framework implicit in maximum likelihood methods allows researchers to compare phylogenetic trees. However, computational requirements are a major limitation of these methods. The number of possible trees with n sequences rapidly increases according to Equation 1.

$$(2n - 3)!/\,(2n - -2\,(n - 2))!$$

Even for a relatively small alignment ($n = 10$), only a subset of trees can be explored. Various strategies to examine tree space have been described, but it is not guaranteed that the best tree is recovered (Vandamme, 2009).

Maximum likelihood trees are always unrooted. Rooting on the branch leading to the outgroup is a drawing option that does not change the fact that the tree is, mathematically speaking, still unrooted.

Time Considerations

The time needed to complete each of the procedures described in the unit depends entirely on the size of the multiple sequence

alignment (number of sequences by the length of alignment) used during the analysis. Construction of a maximum likelihood tree based on the concatenated papillomavirus E6 and E7 alignment was completed in 4 to 5 hr. However, building trees based on larger alignments can take weeks.

Acknowledgements

The authors are grateful to Drs. Varsani and Warburton for critically reading the manuscript. The authors thank members of the Van Doorslaer lab for testing the different protocols of this unit. This work was supported by a State of Arizona Improving Health TRIF and by the National Institute of Food and Agriculture, U.S. Department of Agriculture, Hatch NC229.

Literature Cited

Ahola, V., Aittokallio, T., Vihinen, M., & Uusipaikka, E. (2006). A statistical score for assessing the quality of multiple sequence alignments. *BMC Bioinformatics*, *7*, 484. doi: 10.1186/1471-2105-7-484.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*, 403–410. doi: 10.1016/S0022-2836(05)80360-2.

Anderson, D. R., & Burnham, K. P. (2002). Avoiding pitfalls when using information-theoretic methods. *The Journal of Wildlife Management*, *66*, 912–918.

Anisimova, M., Gil, M., Dufayard, J.-F., Dessimoz, C., & Gascuel, O. (2011). Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. *Systematic Biology*, *60*, 685–699. doi: 10.1093/sysbio/syr041.

Arbogast, B. S., Edwards, S. V., Wakeley, J., Beerli, P., & Slowinski, J. B. (2002). Estimating Divergence Times from Molecular Data on Phylogenetic and Population Genetic Timescales. *Annual Review of Ecology and Systematics*, *33*, 707–740. doi: 10.1146/annurev.ecolsys.33.010802.150500.

Brewer, M. J., Butler, A., & Cooksley, S. L. (2016). The relative performance of AIC, AICC and BIC in the presence of unobserved heterogeneity. *Methods in Ecology and Evolution*, *7*(6), 679–692.

Darriba, D., Taboada, G. L., Doallo, R., & Posada, D. (2012). jModelTest 2: More models, new heuristics and parallel computing. *Nature Methods*, *9*, 772. doi: 10.1038/nmeth.2109.

Darriba, D., Taboada, G. L., Doallo, R., & Posada, D. (2011). ProtTest 3: Fast selection of best-fit models of protein evolution. *Bioinformatics*, *27*, 1164–1165. doi: 10.1093/bioinformatics/btr088.

Felsenstein, J. (1985). Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution; International Journal of Organic Evolution*, *39*, 783. doi: 10.1111/j.1558-5646.1985.tb00420.x.

Felsenstein, J. (1981). Evolutionary Trees From Gene Frequencies and Quantitative Characters: Finding Maximum Likelihood Estimates. *Evolution; International Journal of Organic Evolution*, *35*, 1229. doi: 10.1111/j.1558-5646.1981.tb04991.x.

Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Research*, *39*, W29–37. doi: 10.1093/nar/gkr367.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology*, *59*, 307–321. doi: 10.1093/sysbio/syq010.

Hillis, D. M., & Bull, J. J. (1993). An Empirical Test of Bootstrapping as a Method for Assessing Confidence in Phylogenetic Analysis. *Systematic Biology*, *42*, 182–192. doi: 10.1093/sysbio/42.2.182.

Katoh, K. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*, 3059–3066. doi: 10.1093/nar/gkf436.

Katoh, K., & Toh, H. (2010). Parallelization of the MAFFT multiple sequence alignment program. *Bioinformatics*, *26*, 1899–1900. doi: 10.1093/bioinformatics/btq224.

Ladunga, I. (2017). Finding Homologs in Amino Acid Sequences Using Network BLAST Searches. *Current Protocols in Bioinformatics*, *59*, 3.4.1–3.4.24.

Lanfear, R., Calcott, B., Ho, S. Y. W., & Guindon, S. (2012). Partitionfinder: Combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular Biology and Evolution*, *29*, 1695–1701. doi: 10.1093/molbev/mss020.

Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T., & Calcott, B. (2017). PartitionFinder 2: New Methods for Selecting Partitioned Models of Evolution for Molecular and Morphological Phylogenetic Analyses. *Molecular Biology and Evolution*, *34*, 772–773.

Lemey, P., Salemi, M., & Vandamme, A.-M. (2009). *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*. Cambridge, UK: Cambridge University Press.

Luo, A., Qiao, H., Zhang, Y., Shi, W., Ho, S. Y., Xu, W., ... & Zhu, C. (2010). Performance of criteria for selecting evolutionary models in phylogenetics: A comprehensive study based on simulated datasets. *BMC Evolutionary Biology*, *10*(1), 2.

Maddison, W. P., & Wiens, J. J. (1997). Gene Trees in Species Trees. *Systematic Biology*, *46*, 523–536. doi: 10.1093/sysbio/46.3.523.

Martin, D. P., Murrell, B., Khoosal, A., & Muhire, B. (2017). Detecting and Analyzing

Genetic Recombination Using RDP4. *Methods in Molecular Biology*, *1525*, 433–460. doi: 10.1007/978-1-4939-6622-6_17.

Martin, D., & Rybicki, E. (2000). RDP: Detection of recombination amongst aligned sequences. *Bioinformatics*, *16*, 562–563. doi: 10.1093/bioinformatics/16.6.562.

Miller, M. A., Schwartz, T., Pickett, B. E., He, S., Klem, E. B., Scheuermann, R. H., ... O'Leary, M. A. (2015). A RESTful API for Access to Phylogenetic Tools via the CIPRES Science Gateway. *Evolutionary Bioinformatics Online*, *11*, 43–48.

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*, 268–274. doi: 10.1093/molbev/msu300.

Padidam, M., Sawyer, S., & Fauquet, C. M. (1999). Possible emergence of new geminiviruses by frequent recombination. *Virology*, *265*, 218–225. doi: 10.1006/viro.1999.0056.

Pearson, W. R. (2013). An introduction to sequence similarity ("homology") searching. *Current Protocols in Bioinformatics*, *42*, 3.1.1–3.1.8. doi: 10.1002/0471250953.bi0301s42.

Pearson, W. R. (2016). Finding Protein and Nucleotide Similarities with FASTA. *Current Protocols in Bioinformatics*, *53*, 3.9.1–3.9.25.

Pfeiffer, W., & Stamatakis, A. (2010). Hybrid MPI/Pthreads parallelization of the RAxML phylogenetics code. *In* 2010 IEEE International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum (IPDPSW). Available at: https://doi.org/10.1109/ipdpsw.2010.5470900.

Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T. J, Manuel, M. Wörheide, M. M., & Baurain, D. (2011). Resolving difficult phylogenetic questions: Why more sequences are not enough. PLoS *Biology*, *9*(3), e1000602. doi: 10.1371/journal.pbio.1000602.

Posada, D. (2009). Selection of Models of DNA Evolution with jModelTest. *Methods in Molecular Biology*, *537*, 93–112. doi: 10.1007/978-1-59745-251-9_5.

Posada, D., & Crandall, K. A. (2002). The effect of recombination on the accuracy of phylogeny estimation. *Journal of Molecular Evolution*, *54*, 396–402. doi: 10.1007/s00239-001-0034-9.

Sadreyev, R. I., & Grishin, N. V. (2008). Accurate statistical model of comparison between multiple sequence alignments. *Nucleic Acids Research*, *36*, 2240–2248. doi: 10.1093/nar/gkn065.

Salminen, M., & Martin, D. (2010). Detecting and characterizing individual recombination events. In P. Lemey, M. Salemi, & A. M Vandamme (Eds.) *The Phylogenetic Handbook*, 2nd ed., pp. 519–548. Cambridge, UK: Cambridge University Press.

Schierup, M. H., & Hein, J. (2000). Recombination and the Molecular Clock. *Molecular Biology and Evolution*, *17*, 1578–1579. doi: 10.1093/oxfordjournals.molbev.a026256.

Schmidt, H. A., & von Haeseler, A. Phylogenetic inference using maximum likelihood methods. In P. Lemey, M. Salemi, & A. M Vandamme (Eds.) *The Phylogenetic Handbook*, 2nd ed., pp. 181–209. Cambridge, UK: Cambridge University Press.

Shimodaira, H., & Hasegawa, M. (1999). Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. *Molecular Biology and Evolution*, *16*, 1114–1116. doi: 10.1093/oxfordjournals.molbev.a026201.

Smith, J. M. (1992). Analyzing the mosaic structure of genes. *Journal of Molecular Evolution*, *34*, 126–129. doi: 10.1007/BF00182389.

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*, 1312–1313. doi: 10.1093/bioinformatics/btu033.

Strimmer, K., & von Haeseler, A. (1997). Likelihood-mapping: A simple method to visualize phylogenetic content of a sequence alignment. *Proceedings of the National Academy of Sciences of the United States of America*, *94*, 6815–6819. doi: 10.1073/pnas.94.13.6815.

Sullivan, J., & Joyce, P. (2005). Model Selection in Phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, *36*, 445–466. doi: 10.1146/annurev.ecolsys.36.102003.152633.

Thornton, J. W., & DeSalle, R. (2000). Gene family evolution and homology: Genomics meets phylogenetics. *Annual Review of Genomics and Human Genetics*, *1*, 41–73. doi: 10.1146/annurev.genom.1.1.41.

Tomovic, A., & Oakeley, E. J. (2007). Quality estimation of multiple sequence alignments by Bayesian hypothesis testing. *Bioinformatics*, *23*, 2488–2490. doi: 10.1093/bioinformatics/btm366.

Vandamme, A.-M. (2009). Basic concepts of molecular evolution. *In* P. Lemey, M. Salemi, & A.-M. Vandamme (Eds.), *The Phylogenetic Handbook*, 2nd ed., pp. 3–30. Cambridge, UK: Cambridge University Press.

Van Doorslaer, K., Li, Z., Xirasagar, S., Maes, P., Kaminsky, D., Liou, D., ... McBride, A. A. (2017). The Papillomavirus Episteme: A major update to the papillomavirus sequence database. *Nucleic Acids Research*, *45*, D499–D506. doi: 10.1093/nar/gkw879.

Van Doorslaer, K., Tan, Q., Xirasagar, S., Bandaru, S., Gopalan, V., Mohamoud, Y., ... McBride, A. A. (2013). The Papillomavirus Episteme: A central resource for papillomavirus sequence data and analysis. *Nucleic Acids Research*, *41*, D571–8. doi: 10.1093/nar/gks984.

de Villiers, E.-M., Fauquet, C., Broker, T. R., Bernard, H.-U., & zur Hausen, H. (2004). Classification of papillomaviruses. *Virology*, *324*, 17–27. doi: 10.1016/j.virol.2004.03.033.

Wernersson, R., & Pedersen, A. G. (2003). RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. *Nu-*

cleic Acids Research, *31*, 3537–3539. doi: 10.1093/nar/gkg609.

Xia, X. (2017). DAMBE6: New Tools for Microbial Genomics, Phylogenetics, and Molecular Evolution. *The Journal of Heredity*, *108*, 431–437. doi: 10.1093/jhered/esx033.

Xia, X., & Lemey, P. (2009). Assessing substitution saturation with DAMBE. *In* P. Lemey, M. Salemi, A.-M. Vandamme, P. Lemey, M. Salemi, & A.-M. Vandamme (Eds.), *The Phylogenetic Handbook*, 2nd ed., pp. 615–630. Cambridge, UK: Cambridge University Press.

Xia, X., Xie, Z., Salemi, M., Chen, L., & Wang, Y. (2003a). An index of substitution saturation and its application. *Molecular Phylogenetics and Evolution*, *26*, 1–7. doi: 10.1016/S1055-7903(02)00326-3.

Xia, X., Xie, Z., Salemi, M., Chen, L., & Wang, Y. (2003b). An index of substitution saturation and its application. *Molecular Phylogenetics and Evolution*, *26*, 1–7. doi: 10.1016/S1055-7903(02)00326-3.

Key References

Lemey et. al. , 2009. See above.

Authoritative book on phylogenetic analyses.

Altschul et al., 1990. See above.

Initial Description of the BLAST algorithm (Basic Protocol 1).

Katoh et al., 2002. See above.

Description of the MAFFT algorithm (Basic Protocol 2).

Martin et al., 2017. See above.

Description of the RDP46 software (Basic Protocol 3).

Darriba et al., 2011. See above.

Description of the ProtTest algorithm (Alternate Protocol 1).

Darriba et al., 2012. See above.

Description of the jModelTest algorithm (Alternate Protocol 2).

Lanfear et al., 2017. See above.

Description of the PartitionFinder algorithm (Alternate Protocol 3).

Stamatakis, 2014.

Description of the RAxML algorithm (Basic Protocol 4).

Xia, 2017. See above.

Description of the DAMBE6 software (Support Protocol 4).

Internet Resources

https://www.oxfordjournals.org/nar/database/subcat/5/18.

A list of Viral genome databases maintained by the journal Nucleic Acids Research.

https://evolution.genetics.washington.edu/phylip/software.html

A list of phylogeny packages maintained by Joseph Felsenstein at the University of Washington.

https://blast.ncbi.nlm.nih.gov/

Homepage for the NCBI maintained web based BLAST portal.

https://pave.niaid.nih.gov/

Homepage of the Papillomavirus Episteme (PaVE). Maintained by NIAID. The example alignments were downloaded from PaVE.

https://www.phylo.org/

Homepage of the CIPRES Science Gateway.

https://www.cbs.dtu.dk/services/RevTrans/

Homepage to the RevTrans server. This tool is used in Support Protocol 3.

https://dambe.bio.uottawa.ca/DAMBE/dambe.aspx

Homepage for the lab of Dr. Xia at the University of Ottawa. The DAMBE6 program can be downloaded from this page. This tool is used in Support Protocol 4.

https://web.cbio.uct.ac.za/~darren/rdp.html

Homepage to the lab of Dr. Martin at the University of Cape Town. The RDP4 program can be downloaded from this page. This tool is used in Basic Protocol 3.

https://github.com/ddarriba/prottest3/releases

Github repository for ProtTest 3. This tool is used in Alternate Protocol 1.

https://users-birc.au.dk/biopv/php/fabox/alignment_joiner.php

Homepage to the FASTA alignment joiner. This tool is used in Support Protocol 6.

https://tree.bio.ed.ac.uk/software/figtree/

Homepage for the lab of Dr. Rambaut at the University of Edinburgh. The FigTree program can be downloaded from this page. This tool is used in Basic Protocol 5.