1. Alignment, tree reconstruction using Seaview.

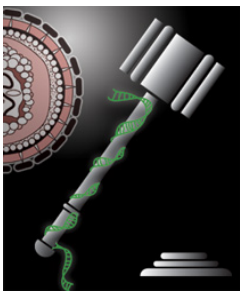
A hands-on practical

Introduction

SeaView is a multiplatform, graphical user interface for multiple sequence alignment and molecular phylogeny (available at http://doua.prabi.fr/software/seaview). SeaView drives the alignment programs Muscle or ClustalW for multiple sequence alignment and computes phylogenetic trees using parsimony, distance-based algorithms and maximum likelihood (ML, using the PhyML program). PhyML provides a wide range of options to perform phylogenetic analyses of nucleotide and amino acid sequences. Early PhyML versions used a fast algorithm to perform Nearest Neighbor Interchanges (NNIs), in order to improve a reasonable starting tree topology (Guindon and Gascuel, 2003). The most recent version 3.0 also includes an efficient algorithm to search the tree space using Subtree Pruning and Regrafting (SPR) topological moves (Hordijk and Gascuel, 2005). PhyML was designed as a heuristic to process moderate to large data sets, and to evaluate branch supports in a sound statistical framework for moderate size data sets. In this practical, we will examine a case of HIV transmission based on two different gene data sets.

Data sets

The molecular investigation of HIV transmission has previously been used as evidence in court (Metzker et al. 2002). Because of the rapid rate of HIV-1 evolution, phylogenetic analysis of HIV-1 DNA sequences is a powerful tool for the identification of closely related viral strains, which may be used to investigate putative transmission between individuals. In Lafayette, Louisiana, a gastroenterologist was accused of trying to kill his former lover by injecting her with HIV-infected blood from one of his patients. The former lover said that on the night of 4 August 1994, the gastroenterologist, who had been giving her vitamin shots, came to her house and gave her another injection against her wishes. In December, after the victim began having suspicious symptoms, her obstetrician tested her for HIV. The victim found out she carried the virus in January 1995, and in May of that year, she accused the gastroenterologist of deliberately infecting her. The gastroenterologist has pleaded not guilty, and his lawyers say he was at home with his wife on the night in question.



As part of their investigation, the police obtained samples of blood from the victim and from the gastroenterologist's only HIV-positive patient. They arranged to have Michael Metzger, then a graduate student in the lab of molecular biologist Richard Gibbs at Baylor College of Medicine in Houston, compare the genetic material from those two HIV strains to each other. They were also compared to viral sequences from 30 randomly chosen HIV patients in the Lafayette area and to hundreds of HIV sequences in the national database.

In this exercise, we will perform a phylogenetic analysis based on the data of this investigation and test the hypothesis of HIV transmission. Metzker *et al.* amplified and sequenced part of the reverse transcriptase (RT, *pol*) and part of the envelope gene (*env*, see Fig. 1). We will download and align the RT sequence data in Seaview; the unaligned *env* sequence is provided in the **HIVenv.fasta** file.

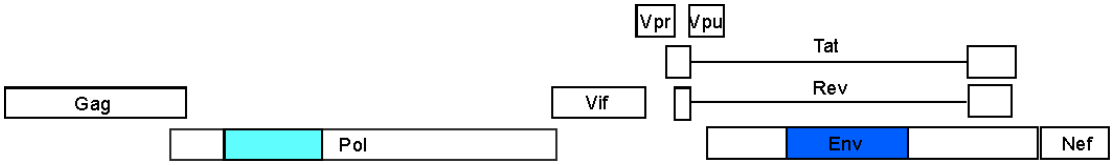
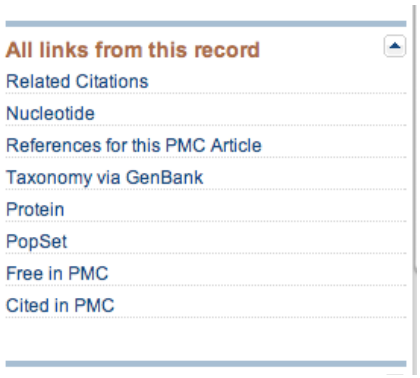


Figure 1. Organization of the coding genome of HIV-1. The analyses will be based on part of the reverse transcriptase in the *pol* gene and a fragment of the *env* gene.

Reconstructing HIV transmission.

Pol data set.

In this first exercise, we will download the RT sequence data and perform multiple alignment using the Muscle algorithm (in Seaview). To explore the sequence data, browse to the Pubmed record of the relevant paper: http://www.ncbi.nlm.nih.gov/pubmed/12388776 (this is also the first hit when entering "Metzker HIV" as search terms in Pubmed). In addition to the abstract of the paper, PubMed also provides several links from this record. Click on the PopSet link and subsequently on the second PopSet record (GI:24209939). One way to download the sequence data would be to change the Display option to "FASTA" and change "Send to" to "File". However, we will make use of SeaView's ability to retrieve sequence data from online databases directly based on a file with the accession numbers. The accession numbers are listed in the Table below and saved to the accessionNumbers.txt file. Note that different clones for the patient (P) and the victim (V) are included, whereas local controls from the Lafayette area (LA) are population sequences. The "BMC" and "MIC" sequences result from replications in two different labs.

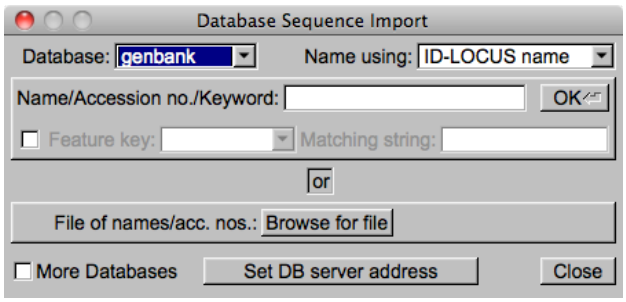


Accession	Strain/name	Accession	Strain/name
AY156734	P1.BCM.RT	AY156783	LA18.RT
AY156735	P2.BCM.RT	AY156784	LA21.RT
AY156736	P3.BCM.RT	AY156785	LA22.RT
AY156737	P4.BCM.RT	AY156786	LA23.RT
AY156738	P5.BCM.RT	AY156787	LA24.RT
AY156739	P6.BCM.RT	AY156788	LA25.RT
AY156740	P7.BCM.RT	AY156789	LA26.RT
AY156741	V1.BCM.RT	AY156790	LA27.RT
AY156742	V2.BCM.RT	AY156791	LA28.RT
AY156771	LA02.RT	AY156792	LA29.RT

Accession	Strain/name	Accession	Strain/name
AY156772	LA04.RT	AY156793	LA30.RT
AY156773	LA05.RT	AY156794	LA31.RT
AY156774	LA06.RT	AY156795	LA32.RT
AY156775	LA07.RT	AY156797	P1.MIC.RT
AY156776	LA08.RT	AY156799	P2.MIC.RT
AY156777	LA10.RT	AY156800	P3.MIC.RT
AY156778	LA12.RT	AY156801	P4.MIC.RT
AY156779	LA13.RT	AY156802	P5.MIC.RT
AY156780	LA14.RT	AY156803	P6.MIC.RT
AY156781	LA16.RT	AY156806	V1.MIC.RT
AY156782	LA17.RT	AY156807	V2.MIC.RT

Table 1. Accession numbers and strain names of the HIV-1 RT sequences.

SeaView

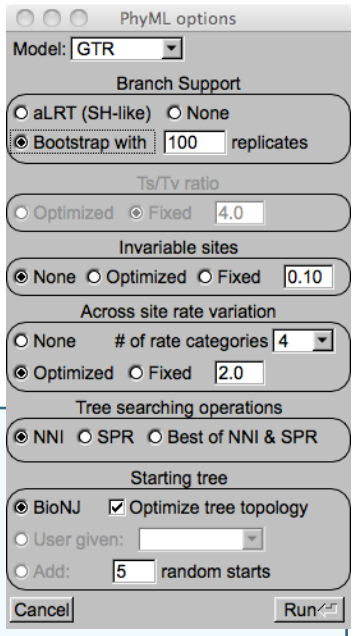


Start the SeaView application. To import the sequence data, select “File” menu, “import from DBs”. In the “Database Sequence Import” window, set “Database” to “genbank”, click “Browse for file” and browse to/select the accessionNumbers.txt file. When the program asks for a “Sequence alignment name’, enter a name (e.g., HIVpol) and click “Ok”. SeaView will import 42 sequences that need to be aligned for further phylogenetic analysis. In the “Align” menu, select “Align all”. As can followed in an “Alignment” console window, Muscle will align the sequences in a few seconds. Click “Ok” to go back to the sequence window. The aligned sequences can be viewed as proteins using the “View as proteins” option under the “Props” menu.

To reconstruct a phylogenetic tree, we will use the “PhyML” program under “Trees” menu. The original PhyML algorithm employs by a mixed heuristic strategy. The program uses a fast distance-based (Neighbor-Joining) method, BioNJ (Gascuel, 1997), to quickly compute a full initial tree. Then fastNNI operations are applied to optimize that tree. During fastNNI all possible NNI trees are evaluated (optimizing only the branch crossed by the NNI) and ranked according to their ML value. Those NNIs which increase the ML value most, but do not interfere with each other, and are simultaneously applied to the current tree. Simultaneously applying different NNIs saves time and makes it possible to walk quickly though tree space. On the new current tree fastNNI is repeated until no ML improvement is possible.

Selecting the PhyML program opens a “PhyML options” window. Select the “GTR” model of substitution (when viewing as nucleotides), bootstrapping using “50“ replicates (see Box 1 for a short summary of the bootstrapping procedure.), across site rate variation modeled using a discretized gamma distribution with “4” categories and a shape

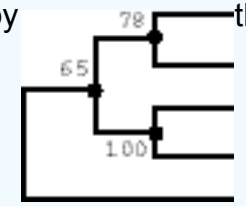
parameter that will be “Optimized” (but no invariant sites), “NNI” branch swapping and default “Starting tree” options. After clicking on “Run”, a “tree-building” console window will appear that reports the progress in the reconstruction procedure. Although a single tree may be reconstructed relatively fast, the bootstrapping procedure may take some time (about 10 minutes depending on the speed of your computer). Note that 50 replicates will probably not lead to a precise evaluation of the robustness of our inference and a higher number, e.g. 1000, is generally preferred.



Box 1. Bootstrapping

In order to assess the support for various alternative phylogenetic tree topologies it is possible to use a bootstrap procedure. This consists of sampling with replacement from the aligned sequence sites and repeating the process of phylogenetic tree reconstruction. Each time a given phylogenetic partition or clade (or group) is supported by the resampled data, its *bootstrap value* is incremented. A *consensus tree* congruent with those clades, which have the highest bootstrap values, can then be defined. Bootstrap values are associated with a given *node* in the consensus tree, and give some indication of the support for the clade defined by that node.

Bootstrap tree:



When the tree-building is completed, click “Ok” and a tree window appears. Use Table 1 to interpret the clustering of the *patient* and *victim* sequences. Bootstrap percentages can be shown at each node in the tree. Take a minute to explore other tree visualization options. Reconstruct a new tree using the “Best of NNI and SPR” branch swapping option in PhyML, but without a bootstrap analysis, and evaluate whether this changes your conclusions on the *patient-victim* clustering.

Env data set.

The *env* data is provided as unaligned sequences in the **HIVenv.fasta** file. Open this file using “File”, “Open” and align the sequences as before. Note the different degree of divergence and alignment ambiguity compared to the previous RT alignment. Explore the evolutionary relationships using distance-based methods with bootstrapping (e.g., “100” replicates), and using PhyML (without bootstrapping). Do we arrive at similar conclusions as compared to RT sequence analysis?

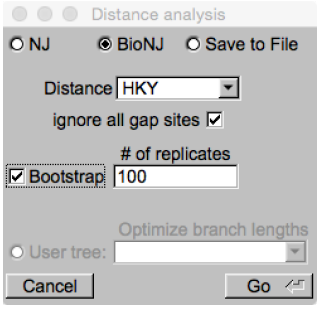
2. Investigating rooting and temporal signal in Ebolavirus.

Introduction

One of the first papers reporting on the genetic makeup of the virus responsible for the 2014 Ebolavirus outbreak suggested that a divergent variant of the Zaire ebola (EBOV) lineage was the cause of the outbreak (Baize et al., 2014). The EBOV strain has previously caused ebola outbreaks in the Democratic Republic of Congo (DRC), the Republic of Congo (RC) and Gabon. The authors published three complete genome sequences from the Guinea outbreak and performed a phylogenetic analysis using 24 sequences of the Zaire and other representative lineages. They show that the 2014 sequences fall as a divergent lineage outside the Zaire lineage suggesting that this may be a pre-existing endemic virus in West Africa rather than the result of spread of the EBOV lineage from the Central African countries that have had previous human outbreaks. These findings have been challenged by Dudas and Rambaut (2014) upon closer examination of the rooting of the EBOV lineage. Here, we will revisit this and assess the correlation between sampling time and genetic divergence using TempEst (http://tree.bio.ed.ac.uk/software/tempest/).
TempEst is a tool for investigating the temporal signal and 'clocklikeness' of molecular phylogenies. It can read and analyse contemporaneous trees (where all sequences have been collected at the same time) and dated-tip trees (where sequences have been collected at different dates). It is designed for analysing trees that have not been inferred under a molecular-clock assumption to see how valid this assumption may be. It can also root the tree at the position that is likely to be the most compatible with the assumption of the molecular clock.

We will use an alignment (Ebolavirus_cds.nex) based on the collation of all protein coding regions in the genome from strains representative of the genus Ebolavirus (which includes Bundibugyo BDBV, Reston RESTV, Sudan SUDV, Tai Forest TAFV and Zaire ebolavirus EBOV species), and a similar alignment that only contains the coding regions from the EBOV lineage (EBOV_cds.nex).

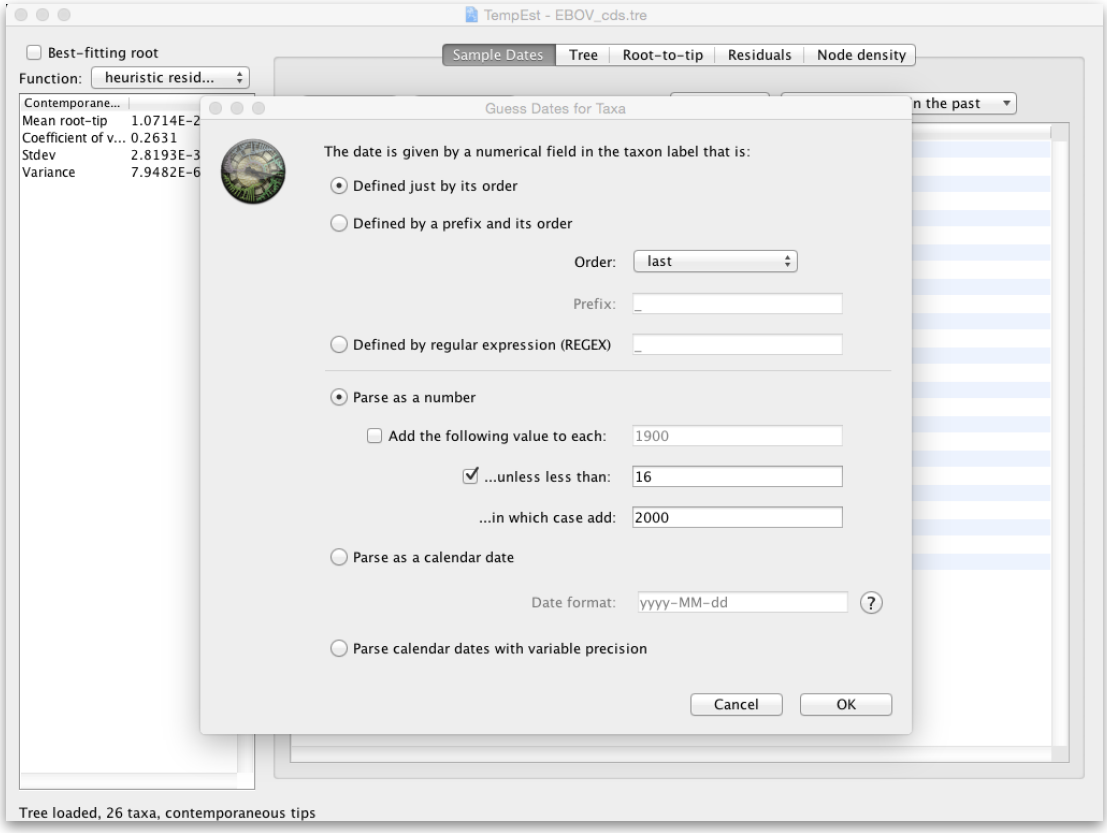
Phylogenetic reconstruction



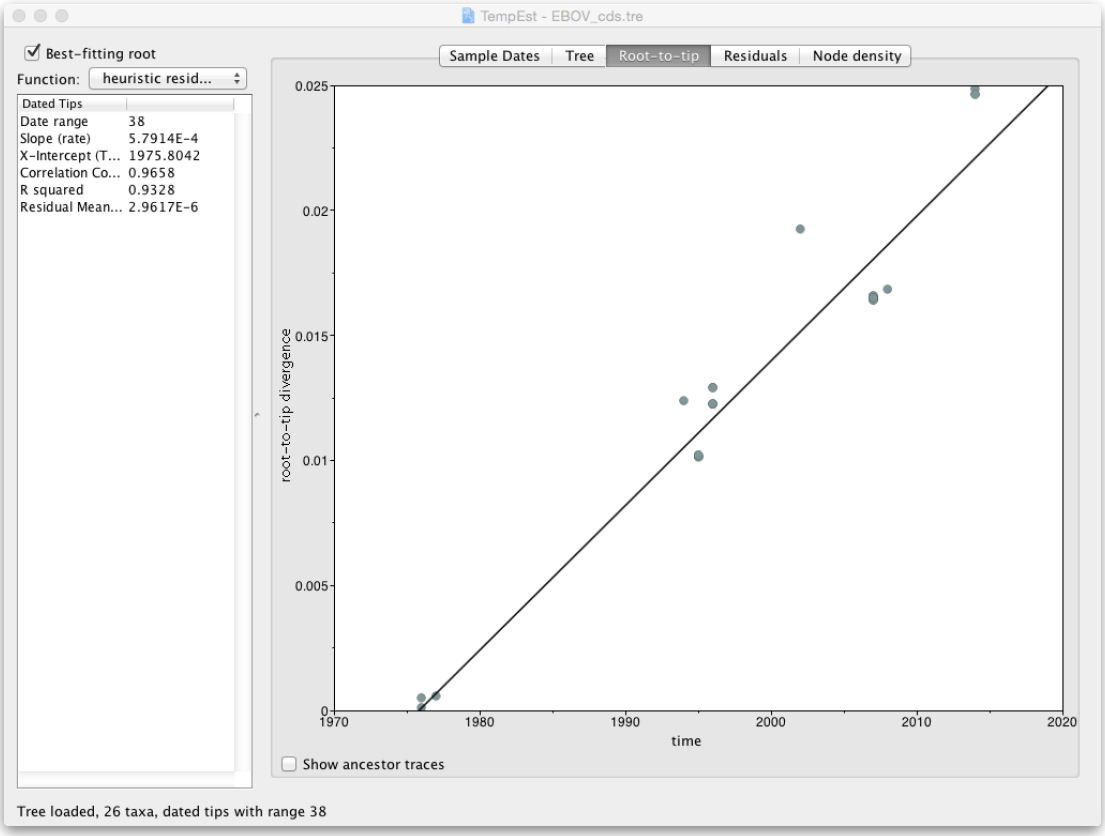
Start SeaView and open the Ebolavirus_cds.nex file (File->Open, and browse to the file). To quickly investigate the phylogenetic position of the three 2014 ebolavirus genomes, reconstruct a distance-based tree (BioNJ) using HKY model of evolution. Explore the robustness of the clustering by including a bootstrap analysis based on 100 replicates (to keep the computation time restricted). Examine the clustering of the 2014 lineage in the resulting phylogeny. Although this is in line with the conclusions by Baize *et al*. (2014), the branch leading to the Guinea outbreak is long, maybe not because it is a divergent lineage but because it is the most recently sampled and therefore had more time to evolve. Combined with a very divergent outgroup this leads to a situation where the root position of the EBOV clade can be unreliably estimated (Dudas and Rambaut, 2014).

To further examine this, load the alignment that only includes the EBOV lineage (EBOV_cds.nex) in SeaView and reconstruct a maximum likelihood tree using the PhyML implementation, using default settings, and save the (unrooted) tree (e.g. as EBOV.tre).

Start TempEst and open the tree file you have saved. In the Sample Dates panel, we need to inform TempEst about the year of sampling for every sequence. This is included at the end of the sequence names (e.g. for EBOV|AF272001|Mayinga|KinshasaDRC|1976). By clicking on the 'Guess Dates', a new window appears:



Keep the default 'Defined just by its order', but set Order to 'last' and click 'OK'. Check that the sampling years have been set correctly before proceeding to the 'Analysis' panel. The 'Root-to-tip' regression is based on the current arbitrary rooting of the tree. For a more appropriate root, select 'Best-fitting' root, which finds the root that maximises the residual mean squared, the correlation or the R-squared for the root-to-tip vs sampling time regression (the exact option is not important in this case). This root results in a clear accumulation of divergence in between outbreaks as a function of the sampling time of the viruses in these outbreaks. Check the 'Best-fitting' root in the 'Tree' window. How does this differ from the outgroup rooting in the full data set?


TempEst - EBOV_cds.tre
Best-fitting root
Function:
heuristic resid...
Dated Tips
Date range 38
Slope (rate) 5.7914E-4
X-Intercept (T... 1975.8042
Correlation Co... 0.9658
R squared 0.9328
Residual Mean... 2.9617E-6
Sample Dates
Tree
Root-to-tip
Residuals
Node density
root-to-tip divergence
0.025
0.02
0.015
0.01
0.005
0
1970
1980
1990
2000
2010
2020
time
Show ancestor traces
Tree loaded, 26 taxa, dated tips with range 38

References

Metzker ML, Mindell DP, Liu XM, Ptak RG, Gibbs RA, Hillis DM. Molecular evidence of HIV-1 transmission in a criminal case. PNAS, 2002 29;99(22):14292-7.

Gascuel O. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Molecular Biology and Evolution, 1997, 14:685-695.

Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood." Systematic Biology, 2003, 52(5):696-704.

Hordijk W., Gascuel O. Improving the efficiency of SPR moves in phylogenetic tree search methods based on maximum likelihood. Bioinformatics, 2005, 21(24), pp. 4338-4347.

Gouy M., Guindon S. & Gascuel O. (2010) SeaView version 4 : a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Molecular Biology and Evolution 27(2):221-224.

Baize S, Pannetier D, Oestereich L, Rieger T, Koivogui L, Magassouba N, Soropogui B, Sow MS, Keïta S, De Clerck H, Tiffany A, Dominguez G, Loua M, Traoré A, Kolié M, Malano ER, Heleze E, Bocquin A, Mély S, Raoul H, Caro V, Cadar D, Gabriel M, Pahlmann M, Tappe D, Schmidt-Chanasit J, Impouma B, Diallo AK, Formenty P, Van Herp M, Günther S. Emergence of Zaire Ebola Virus Disease in Guinea - Preliminary Report. N Engl J Med. 2014 Apr 16. PubMed PMID: 24738640.

Rambaut, A., T.T. Lam, L. de Carvalho, and O.G. Pybus. 2016. Exploring the temporal structure of heterochronous sequences using TempEst. Virus Evolution 2: vew007 DOI: http://dx.doi.org/10.1093/ve/vew007.

Dudas G, Rambaut A. Phylogenetic Analysis of Guinea 2014 EBOV Ebolavirus Outbreak. PLOS Currents Outbreaks. 2014 May 2. Edition 1. doi: 10.1371/currents.outbreaks.84eefe5ce43ec9dc0bf0670f7b8b417d.