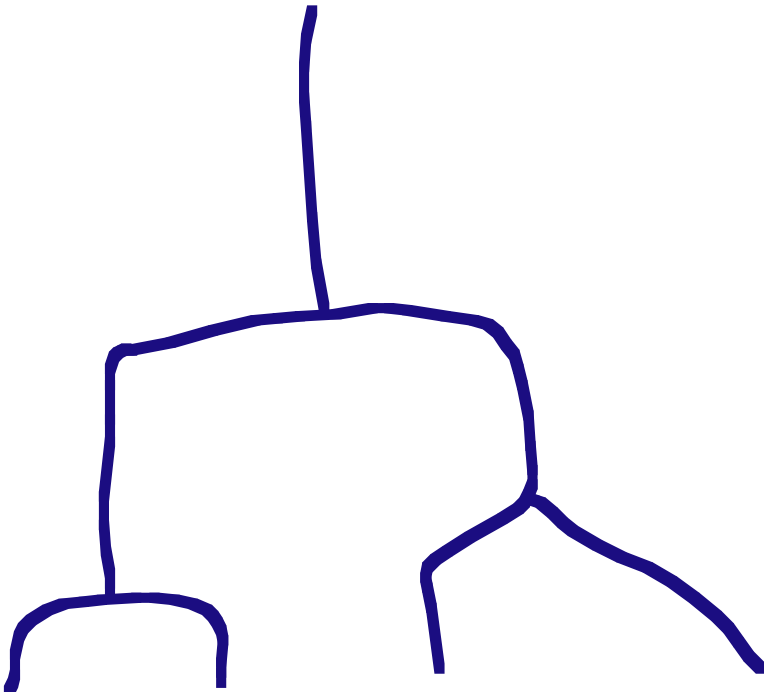


Molecular phylogenetics practical

Introduction

This session is designed to introduce you to a web service dedicated to phylogenetic inference, which can be found at:

http://www.phylogeny.fr/

Phylogeny.fr runs and connects various bioinformatics programs to reconstruct a robust phylogenetic tree from a set of sequences. It includes multiple alignment procedures, alignment curation utilities and tree visualization.

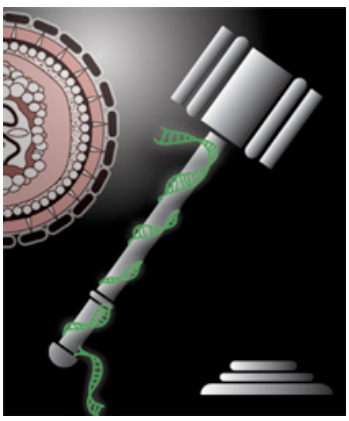
Citation:
Dereeper A., Guignon V., Blanc G., Audic S., Buffet S., Chevenet F., Dufayard J.-F., Guindon S., Lefort V., Lescot M., Claverie J.-M., Gascuel O. Phylogeny.fr: robust phylogenetic analysis for the non-specialist Nucleic Acids Research. 2008 Jul 1; 36 (Web Server Issue):W465-9. Epub 2008 Apr 19.

Note: a curated list of most other available software packages related to phylogenetic analysis can be found on the web site:

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

This web page is maintained by Joe Felsenstein, who wrote the PHYLIP package.

Background:



In this exercise, we will perform a molecular investigation of HIV transmission for which the results have 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 evaluate transmission hypotheses 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 *a priori* hypothesis of HIV transmission. Metzker *et al.* amplified and sequenced part of the reverse transcriptase (polymerase, *pol*) and part of the envelope gene (See Fig. 1). In this practical, we will use the *pol* data to test the transmission hypothesis.

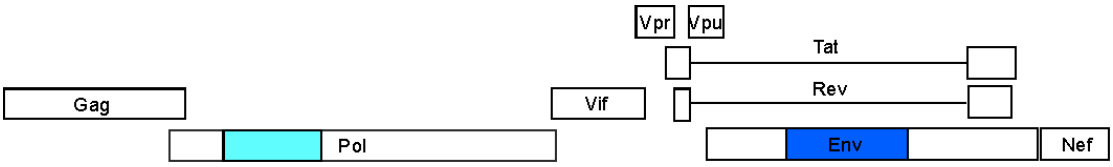


Figure 1. Organization of the coding genome of HIV-1. The original analyses were based on part of the reverse transcriptase in the *pol* gene and a fragment of the *env* gene.

In the first part of this document, an introduction to PAUP* is provided that can be used as a reference. The actual exercise starts on page 9 (2. Distance methods).

Phylogenetic inference using Phylogeny.fr

Pol analysis

In the first exercise, we will analyze the HIV *pol* population sequences, including viruses sampled from the victim, the patient and local controls. Go to http://www.phylogeny.fr and scroll down the page to the "Phylogeny analysis" section. Click on "Advanced" to manually set parameters for the various steps in the procedure. This will bring you to the "Workflow settings". Since the *pol* sequences are already aligned, there is no need to include alignment or curation, so deselect "Muscle" and "Gblocks" (Keep "PhymL" and "TreeDyn"). Next, click on "Create workflow"; this will bring you to the input data window. "Choose File" and browse to and select the "HIVpol.fasta" file. Scroll down the page to "Phylogeny: PhyML" and change the "Substitution model" to "GTR (DNA/RNA)". Before submitting at the bottom of the page, you can enter your email, but this is not necessary when following the analysis online.

Having submitted the analysis, phylogeny.fr proceeds with PhyML analysis, which should take less than a minute, and arrives at the "Tree Rendering" page.

Consider the phylogenetic relationships of the HIV strains in the victim (V1_BCM), the gastroenterologist's patient (P1_BCM) and local control sequences (LA). Concerning the hypothesis of deliberate transmission, there are two alternative tree topologies for the clustering of the victim's HIV sequence:

((V1_BCM, *LA*), P1_BCM)

((V1_CM, P1_BCM), *LA*)

The brackets define the phylogenetic hierarchy of the clustering. *Tick the one that represents your analysis.*

The red numbers at the nodes represent approximate likelihood ratio test values, which are similar to bootstrap support values. *Is the clustering you obtained well supported*?

The online tree-rendering tool (TreeDyn) allows you to change the way the phylogeny is visualized (by selecting "actions"); explore some of these options. *What is the difference between a phylogram and a cladogram?* The tree can be saved in graphics format (png, pdf, svg) or standard tree formats like "Newick".

Env analysis

If there is sufficient time left, we can proceed with the analysis of the larger *env* data set. This unaligned data set also contains cloned sequence variants of the patient and the victim. Go to the “Advanced Mode” in Phylogeny.fr and create a workflow using the default processing steps (we now have to perform alignment and curation). “Choose File” and browse to and select the “HIVenv.fasta” file. Select again the “GTR (DNA/RNA)” as “Substitution model” for this analysis and submit the analysis. Because the *env* data contains more sequences than the *pol* data set (128 vs. 27), and we require multiple sequence alignment, the analysis will take considerably longer, in particular the tree reconstruction (5-10 minutes depending on then server load).

What are the relationships of the victim, the patient and the control sequences in the env gene and what is the support for this clustering?

Have a look at the alignment curation and what alignment blocks Gblocks has decided to keep for the phylogenetic analysis. *What percentage is that of the Muscle alignment?*