Alignment and 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://pbil.univ-lyon1.fr/software/seaview.html). 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 perform phylogenetic reconstruction using different data sets. As an example of nucleotide phylogenetic analysis, we will examine a case of HIV transmission based on two different gene data sets. Using a second data set consisting of amino acid sequences of various primate immunodeficiency viruses, we will investigate interspecies transmissions resulting in different HIV lineages.

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.

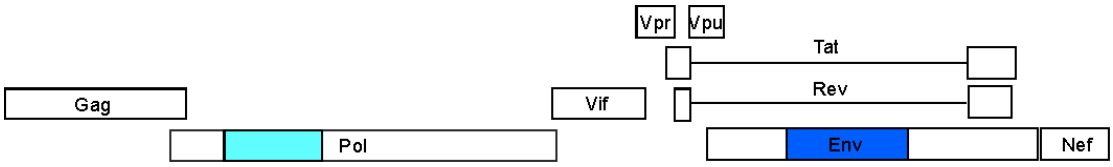


Figure 1. Organization of the coding genome of HIV-1. The analyses will be based on part of the reverse transcriptase in the *pol* gene and a fragment of the *env* gene.

To investigate the evolutionary history of more divergent primate immunodeficiency viruses, and to identify the simian viruses most closely related to HIV, we will analyze partial reverse transcriptase protein sequences. Shortly after the discovery of HIV, related lentiviruses viruses were found in different non-human primates. Simian immunodeficiency viruses (SIVs) have now been identified in no fewer than 36 different nonhuman primate species in sub-Saharan Africa. The general rule seems to be that each infected primate species has its own specific SIV and, intriguingly, naturally infected primates do not go on to be seriously sickened by the virus (but see Keele et al., 2009). The data we analyze here has been compiled in the **PIV.phy** file.

Reconstructing HIV transmission.

Pol data set.

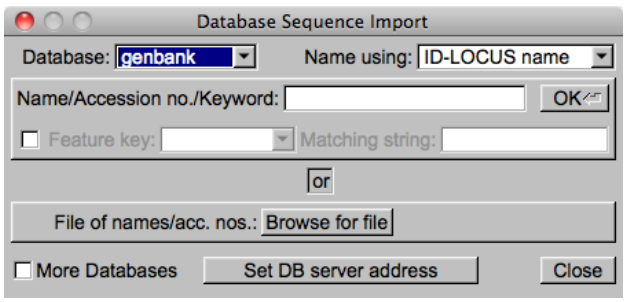
In this first exercise, we will download the RT sequence data and perform multiple alignment using the Muscle algorithm (in Seaview). To explore the sequence data, browse to the Pubmed record of the relevant paper: http://www.ncbi.nlm.nih.gov/pubmed/12388776 (this is also the first hit when entering "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 number	Strain/name	Accession number	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
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

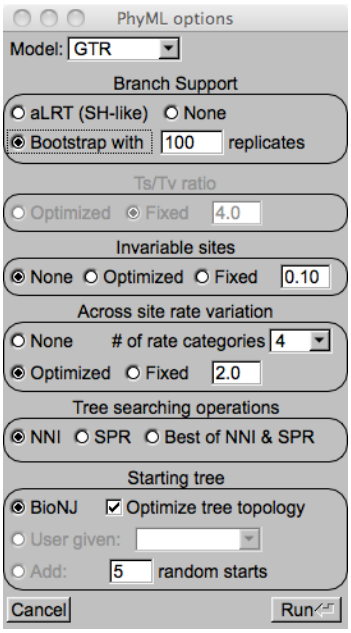
Table 1. Accession numbers and strain names of the HIV-1 RT sequences.

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.



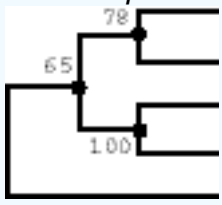
To reconstruct a phylogenetic tree, we will use the “PhyML” program under “Trees” menu. The original PhyML algorithm employs by a mixed heuristic strategy. The program uses a fast distance-based (Neighbor-Joining) method, BioNJ (Gascuel, 1997), to quickly compute a full initial tree. Then fastNNI operations 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?

Reconstructing the primate immunodeficiency evolutionary history

As an example of an amino acid phylogenetic analysis, we analyze partial *pol* amino acid sequences from different primate immunodeficiency viruses (PIV). Because of their high sequence divergence, analysis of nucleotide sequences may suffer from substitution saturation. There are two types of HIV, type 1 and type 2, that differ in their natural history, infectivity, and pathogenicity, as well as in details of their genomic structure. HIV-2 has remained largely restricted to West Africa. HIV-1 is classified in 3 different groups (M, N and O) that arose from separate cross species transmissions. While group N and O infections are largely restricted to Central Africa (essentially Cameroon), the worldwide pandemic is caused only by HIV-1 group M. Group M is further classified in 9 different pure subtypes (A, B, C, D, F, G, H, J and K), and several circulating recombinant forms (CRFs). All HIV sequences in the data set analyzed here are listed in Table 2 for easy interpretation of the trees. All other PIV sequences were sampled from non-human primates and the first part of their sequence name represents the host species (Table 3).

The PIV data is provided as aligned amino acid sequences in the **PIV.phy** file. Explore the evolutionary relationships using distance-based methods with bootstrapping (e.g., “100” replicates), and using PhyML (without bootstrapping). Note the different substitution models for amino acid data. What conclusion can be drawn in terms of interspecies transmissions: how many, and which simian viruses appear to be the progenitors of the HIV lineages?

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.

Keele BF, Jones JH, Terio KA, Estes JD, Rudicell RS, Wilson ML, Li Y, Learn GH, Beasley TM, Schumacher-Stankey J, Wroblewski E, Mosser A, Raphael J, Kamenya S, Lonsdorf EV, Travis DA, Mlengeya T, Kinsel MJ, Else JG, Silvestri G, Goodall J, Sharp PM, Shaw GM, Pusey AE, Hahn BH. Increased mortality and AIDS-like immunopathology in wild chimpanzees infected with SIVcpz. Nature. 2009 Jul 23;460(7254):515-9.

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.

Type	Group	Sequence	subtype
1	M	A1_UG_85_U455_M62320	A1
		B_US_90_WEAU160_U21135	B
		C_ET_86_ETH2220_U46016	C
		D_CD_84_84ZR085_U88822	D
		F1_BE_93_VI850_AF077336	F1
		G_SE_93_SE6165_AF061642	G
		H_CF_90_056_AF005496	H
		J_SE_93_SE7887_AF082394	J
		K_CM_96_MP535_AJ249239	K
		01_AE_TH_90_CM240_U54771	CRF01_AE
		02_AG_NG_x_IBNG_L39106	CRF02_AG
		03_AB_RU_97_KAL153_2_AF193276	CRF03_AB
		04_cpx_CY_94_CY032_AF049337	CRF04_cpx
	N	N_CM_95_YBF30_AJ006022	N/A
	O	O_BE_87_ANT70_L20587	
		O_BE_87_ANT70_L20587	
2	A	A_GW_x_ALI_AF082339	
		A_DE_x_BEN_M30502	
		A_SN_x_ST_M31113	
	B	B_BF_x_BF121_AY936769	
		B_GH_86_D205_X61240	
		B_CI_x_EHO_U27200	
	G	G_CI_92_ABT96_AF208027	
	Unknown	U_FR_96_12034_AY530889	

Table 2. Human immunodeficiency viruses included in the PIV data set.

name prefix	host species
CPZ	Chimpanzee
GOR	Gorilla
SMM	Sooty mangabey
MNE	Macaque
DRL	Drill
MND	Mandrill
ERY	Red-eared Guenon
BKM	Black mangabey
TAL	Talapoin monkey
TAN	Tantalus monkey
VER,SAB	African green monkey
RCM	Red-capped mangabey
SUN	Sun-tailed macaque
MON	Mona monkey
MUS	Moustached monkey
DEB	De Brazza's monkey
GSN	Greater spot-nosed monkey
DEN	Dent's mona monkey
LST	L'hoest monkey
SYK	Sykes' monkey
COL	Colobus monkey

Table 3. Non-human primate virus naming and their host species