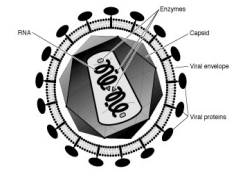

Structure of Human Immunodeficiency Virus (HIV)

Hands-On Twelve Origins of HIV

We are going to build phylogenetic trees with 12 sequences from HIV1, HIV2 and SIV. The sequences are from NCBI. The table lists, from left to right:

- **Num**: the number representing the order in which the sequences were obtained from NCBI
- **Isolate**: the isolate names of the genomes as given by NCBI
- **Accession Number**: GenBank accession numbers of the whole genomes
- **Length in base pairs**: the length of the genome
- **Posted Entry Date:** the date found in the corresponding database entry
- **Organism**: the organism to which the genome belongs.

Num	Isolate	Accession Number	Length in base pairs	Posted Entry Date	Organism
1	HIV1_ELI	K03454	9176	27-FEB-2002	Human
2	HIV1_BRU	K02013	9229	02-AUG-1993	Human
3	HIV1_MAL	X04415	9229	18-APR-2005	Human
4	HIV1_NDK	M27323	9143	02-AUG-1993	Human
5	HIV2_D205	X61240	10269	14-NOV-2006	Human
6	HIV2_ROD	M15390	9671	23-MAY-1996	Human
7	HIV2_ST	M31113	9672	29-MAY-1996	Human
8	HIV2_UCI	L07625	10271	02-AUG-1993	Human
9	SIV_Mm251	M19499	10277	29-NOV-2000	Macaque
10	SIV_CPZ	X52154	9811	18-APR-2005	Chimpanzee
11	SIV_AGM	M58410	9623	31-MAR-2006	Af. green monkey
12	SIV_SMM	X14307	10241	18-APR-2005	Sooty mangabey mk.

Three files, env_protein_sequences.txt, gag_protein_sequences.txt, and pol_protein_sequences.txt containing the amino acid sequences for the env, gag and pol proteins from the 12 isolates were created. In other words:

- env_protein_sequences.txt contains 12 env protein sequences, one from each of the 12 genomes found in the table.
- gag_protein_sequences.txt contains 12 gag protein sequences, one from each of the 12 genomes found in the table.
- pol_protein_sequences.txt contains 12 pol protein sequences, one from each of the 12 genomes found in the table.

Part A:

- Retrieve “env_protein_sequences.txt” and copy all 12 sequences.
- Go to http://www.ebi.ac.uk/Tools/msa/clustalo and paste all 12 sequences in the window.

©2016 Sami Khuri

- Click on "More options…" under STEP 2 and open the dropdown window under OUTPUT FORMAT.
- Choose "PHYLIP" from the pull-down menu. We need to have the resulting alignment in the phylip format for subsequent steps.
- Click on "Submit" of STEP 3 to align the 12 env protein sequences and wait for the result.
- The new page contains the actual alignment. The multiple sequence alignment is preceded by 2 numbers:
    - The first number represents the number of sequences (rows). In our example: 12.
    - The second represents the total number of columns in the alignment. In our example: 950.
- Use "Download Alignment File" to save the alignment under "clustalo-env.txt" on your desktop.

In a new window:

- Go http://mobyle.pasteur.fr. We are going to use the PHYLIP package to construct phylogenetic trees.
- Click on "phylogeny" under "Programs"
- Click on "distance" and then on "protdist". The protdist program from the PHYLIP package will compute the pairwise distances of the 12 sequences from the multiple sequence alignment.
- Copy the entire content of the "clustalo-env.txt" you obtained above, including the first row with the two numbers, and paste it in the Phylip protdist window.
    - Alternatively, you can choose the "upload" button under "Alignment File" by clicking on it and then clicking on "Browse" to upload "clustalo-env.txt". Then choose "clustalo-env.txt" from the drop-down window next to "select" and click on "select". The contents of "clustalo-env.txt" should appear in the display window.
- Enter your email address in the appropriate window, after clicking on "set email" at the top of the PHYLIP page, on the right-hand side and click on "OK".
- Scroll down to "Bootstrap options" and choose "Yes" in the box to the right of "Perform a bootstrap before analysis ?".
- Type "1" in the window to the right of "Random number seed (must be odd)". The random seed is used to get randomness in the bootstrap algorithm.

- Keep the default value of 100 for the number of replicates. We want the package to calculate distance matrices for the 100 bootstrap replicates (the 100 problem instances that are randomly generated).
- Scroll up to the top of the page and click on the "Run" button.
- You might be prompted to validate your submission. Do it please.
- You will get a new page, and will have to wait (sometimes up to a few minutes) until you see an output file, "protdist.outfile" under "Outfile (*PhylipDistanceMatrix)*". You will also get email messages from PHYLIP.
- The "protdist.outfile" file will be used as input file to PHYLIP's neighbor.
- Go to the row right underneath the "protdist.outfile" window (that starts with "full screen") and choose "neighbor (inflile)" from the drop-down window right before "further analysis".
- Click on "further analysis" to obtain a new page.
- From the new page, click on "advanced options"
- Scroll down to "Bootstrap options":
    - Choose "Yes" for "Analyze multiple data sets"
    - Type "100" for "How many data sets"
    - Type "1" in the "Random number seed for multiple dataset (must be odd)" window
    - Choose "Yes" for "Computer a consensus tree"
- Scroll back up and click on "Run".
- From the new page, you will have several different output files under "results". We are interested in the very first one: "consense.outfile". If you scroll down in that window, you will see the consensus tree produced by the 100 bootstrap runs.
- We would like to save "consensus.outfile" (in a readable format).
- Click on "full screen" right underneath the "Consense output file"
- Save the resulting page under "env consensus"
- Open the file you have just saved: env_consensus, and scroll down to see the consensus tree with the bootstrap values for each branch.

Part B:

1) Repeat the procedure described in part A with gag_hiv_siv_sequences.txt and save the resulting consensus trees under gag_consense_outfile.txt.

2) Repeat the procedure described in part A with pol_hiv_siv_sequences.txt and save the resulting consensus trees under pol_consense_outfile.txt.

Part C:

Study carefully the three consensus trees and answer the following questions:

1) What do the trees show with regards to the HIV and SIV relationships?
2) Why do SIV's cluster with both HIV-1 and HIV-2?
3) Which HIV type, HIV-1 or HIV-2, is more closely related to the SIV from the sooty mangabey? Which type is more closely related to the SIV from the chimpanzee? What does this tell you about the origin of HIV-1 and HIV-2?

This project was adapted from "The Origin and Evolution of HIV" from Siv Andersson's laboratory at Uppsala University, Sweden.
http://artedi.ebc.uu.se/course/ugsbr/hiv/.

More about the HIV Genome and Virus

1) http://www.hiv.lanl.gov/
The HIV databases contain data on HIV genetic sequences, immunological epitopes, drug resistance-associated mutations, and vaccine trials. The website also gives access to a large number of tools that can be used to analyze these data. This project is funded by the Division of AIDS of the National Institute of Allergy and Infectious Diseases (NIAID), a part of the National Institutes of Health (NIH).

2) http://www.mcld.co.uk/hiv/
This hypertext looks at HIV from a molecular point of view, using an indexed set of entries.
Important remark from that site:
"The gag and pol genes are right next to each other in the HIV genome - in fact, they overlap a little - and when the proviral genome is being transcribed from DNA into fresh RNA, sometimes the cell machinery makes a "mistake". Instead of finishing copying out the gag gene, it hops onto the pol gene and also copies that code. So the RNA which comes out is the gag and pol code run together into one longer code.
From the point of view of HIV this isn't a mistake. The combined gag-pol transcript encodes some of its most important proteins, such as reverse transcriptase and integrase. In fact HIV encourages this particular "mistake", with strategically-placed bumps in its DNA to encourage the jump to occur. As a result, gag-pol gets produced around 1 in every 20 occasions. This means that its resultant proteins are found at around one twentieth of the concentration of the "normal" products."
