Student: Yu Cheng (Jade)

ICS 675

Analysis Project:The Florida Dentist Case Revisited

October 05, 2009

Download the Dataset

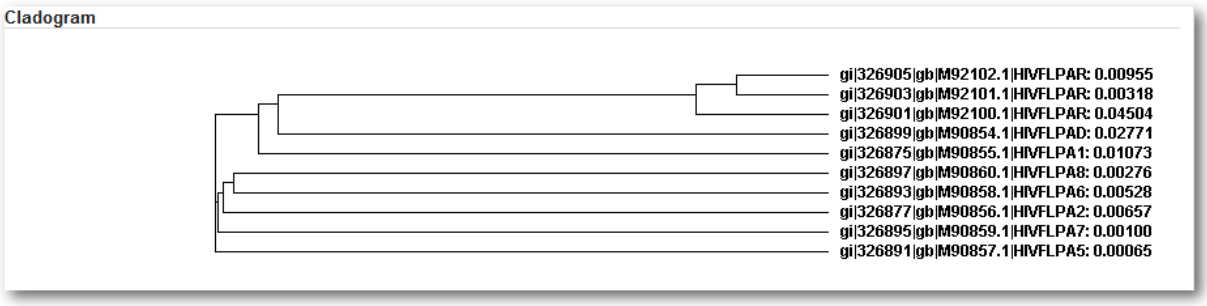
In NCBI's Genbank, I searched for "HIV-1 v3 gene" for Florida dentist case related sequences and control group sequences. I downloaded several sequences for each individual related to the Florida dentist case, and several for each control group. Namely, I obtained 10 sequences for patient A, 18 sequences from patient B, 6 sequences from patient C, 6 sequences from patient D, 7 sequences from patient E, 7 sequences from patient F, 6 sequences from patient G, 6 sequences from patient H, 7 sequences from the dentist, 10 sequences from China, 10 sequences from Europe, and 11 sequences from the Liberty city Florida. The Genbank IDs for all these sequences are shown in the Cladogram trees in the following section.

Clean the Dataset

Strategy: The image in each section was the phylogenetic tree for each group obtained using Clustalw program. Base on these guide trees, I selected a couple of sample sequences to represent each group. The sequences in each experimental group were sequenced from different clones from the same individual, so the goal is to select the sequences that can represent this individual. If the particular clone's DNA sequence is farther away from other clones, it might've experienced uncommon mutations. Therefore I selected the representatives from the ones that are topologically close with each other.

On the other hand, the sequences in the control groups were not obtained from the same person. They were far apart from each other to begin with. So I just randomly selected a couple to represent each control group.

Followed by each image are the DNA nucleotide sequences and protein amino acid sequences that were selected to represent this group. As I mentioned the selection was based on the Clustalw output.



Florida Patient A (FLPA 6, 8)

FLPA6:

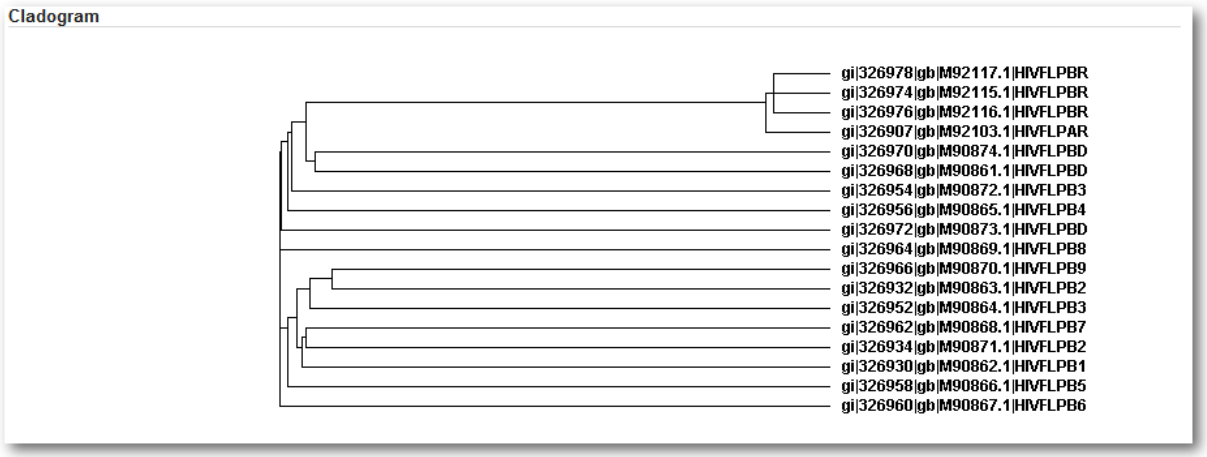
ctagcagaag aagaggtagt aattagatct gccaatttca cagacaatgc taaaatcata atagtacaac tgaatgcatc tgtaaaaatt aaatgtacaa gacccaacaa caatacaaga aaaggtatac agataggacc aggaagggca ttttatgcaa caggagaaat aataggagat ataagacaag cacattgtaa cattagtaga gaaaaatgga ataatacttt aaagcaggta gttacaaaat taagagaaca atttgagaat aaaacaataa tctttaatca ctcctcagga ggggacccag aaattgtaatg cacagttttaa ttgtggaggg gaatttttc

LAEEEVVIRSANFTDNAKIIIVQLNASVKIKCTRPNNNTRKGIQIGPGRAFYATGEIIGDIRQAHCNISREKWNNTLKQV VTKLREQFENKTIIFNHSSGGDPEIVMHSFNCGGEFF

FLPA8:

ctagcagaag aagaggtagt aattagatct gccaatttca cagacaatgc taaaatcata atagtacaac tgaatgcatc tgtagaaatt aattgtacaa gacccaacaa caatacaaga aaaggtatac agataggacc aggaagggca ttttatgcaa caggagaaat aataggagat ataagacaag cacattgtaa cattagtaga gaaaaatgga ataatacttt aaagcaggta gttacaaaat taagagaaca atttgagaat aaaacaataa tctttaatca ctcctcagga ggggaccca gaaattgtaat gcacagtttta cttgtggaggg gaatttttc

LAEEEVVIRSANFTDNAKIIIVQLNASVEINCTRPNNNTRKGIQIGPGRAFYATGEIIGDIRQAHCNISREKWNNTLKQV VTKLREQFENKTIIFNHSSGGDPEIVMHSFTCGGEFF



Florida Patient B (FLPB 2, 3)

FLPB2:

ctagcagaag aagaagtagt aattagatct gccaatttca cagacaatgc taaaatcata atagtacagc tgaatgcatc tgtagaaatt aattgtacaa gacccaacaa caatacaaga aaaggtatac atataggacc agggagggca ttttatgcaa caggagaaat aataggagat ataagacaag cacattgtaa cattagtaga gaaaaatgga ataatacttt agaacaggta aaaacaaaat taagagaaca atttgagaat aaaacaataa tctttaatca ctcctcagga ggggacccag aaattgtaacg cacagtttta attgtggaggg g

LAEEEVVIRSANFTDNAKIIIVQLNASVEINCTRPNNNTRKGIHIGPGRAFYATGEIIGDIRQAHCNISREKWNNTLEQV KTKLREQFENKTIIFNHSSGGDPEIVTHSFNCGG

FLPB3:

ctagcagaag aagaggtagt aattagatct gccaatttca cagacaatgc taaaatcata atagtacagc tgaatgcatc tgtagaaatt aattgtacaa gacccaacaa caatacaaga aaaggtatac atataggacc agggagggca ttttatgcaa caggagaaat aataggagat ataagacaag cacattgtaa cattagtaga gcaaaatgga ataatacttt aaaacaggta gaaacaaaat taaaagaaca atttaataaa acaataatct ttaagcactc ctcaggaggg gacccagaaa ttgtaatgcac agttttaattg tggaggg

LAEEEVVIRSANFTDNAKIIIVQLNASVEINCTRPNNNTRKGIHIGPGRAFYATGEIIGDIRQAHCNISRAKWNNTLKQV ETKLKEQFNKTIIFKHSSGGDPEIVMHSFNCGG

Florida Patient C (FLPC 1, 6)

FLPC6:

ctagcagaag aagaggtagt aattagatct gccaatttca cagacaatgc taaaatcata atagtacagc tgaatgcatc tgtagaaatt aattgtacaa gacccaacaa caatacaaga aaaggtatac atataggacc agggagagca gtttatgcaa cagacagaat aataggagat ataagacaag cacattgtaa cattagtaga gaaaaatgga ataatacttt aaaacaggta gttacaagat taagagaaca atttgtgaat aaaacaataa tctttactca cccctcagga ggggacccag aaattgtaatg cacagtgttaa ttgtggaggg gaatttt

LAEEEVVIRSANFTDNAKIIIVQLNASVEINCTRPNNNTRKGIHIGPGRAVYATDRIIGDIRQAHCNISREKWNNTLKQV VTRLREQFVNKTIIFTHPSGGDPEIVMHSVNCGGEF

FLPC12:

ctagcagaag aagaggtagt aattagatct gccaatttca cagacaatgc taaaatcata atagtacagc tgaatgcatc tgtagaaatt aattgtacaa gacccaacaa caatacaaga aaaggtatac atataggacc agggagagca gtttatgcaa cagacagaat aataggagat ataagacaag cacattgtaa cattagtaga gaaaaatgga ataatacttt aaaacaggta gttacaaaat taagagaaca atttgtgaat aaaccaataa tctttactca cccctcagga ggggacccag aaattt

LAEEEVVIRSANFTDNAKIIIVQLNASVEINCTRPNNNTRKGIHIGPGRAVYATDRIIGDIRQAHCNISREKWNNTLKQV
VTKLREQFVNKPIIFTHPSGGDPEI

Florida Patient D (FLPD 1, 9)

FLPD9:

ctagcagaag aagaggtagt aattagatct gcaaatttct cggacaatgc taaaaccata atagtacagc tgaataaatc tgtaaaaatt ccttgtataa gacccagcaa taatacaaga caaagtatac ctataggacc agggaaagca gtttatgcaa caggacagat aataggagat ataagaaagg cacatcgtaa ccttagtgaa gcaatatgga ataacacgtt aaaacagata gttaaaaaat taaaagaaca atttaagaat aaaacaatag tcttcaatca atcctcagga ggggacccag aaattgtaatg cacagttttaa ttgtg

LAEEEVVIRSANFSDNAKTIIVQLNKSVKIPCIRPSNNTRQSIPIGPGKAVYATGQIIGDIRKAHRNLSEAIWNNTLKQIV
KKLKEQFKNKTIVFNQSSGGDPEIVMHSFNC

FLPD12:

ctagcagaag aagaggtagt aattagatct gcaaatttct cggacaatgc taaaaccata atagtacagc tgaaagaacc tgtaaaaatt aagtgtataa gacccagcaa taatacaaga caaagtatac ctataggacc agggaaagca gtttatgcaa caggacagat aataggagat ataagaaaag cacattgtaa ccttagtgaa gcaagatgga ataacacgtt agaacagata gttaaaaaat taaaagaaca atttaagaat aaaacaataa tcctcaatca atcctcagga ggggacccag aaattgtaatg cacagttttaa ttgtg

LAEEEVVIRSANFSDNAKTIIVQLKEPVKIKCIRPSNNTRQSIPIGPGKAVYATGQIIGDIRKAHCNLSEARWNNTLEQIV
KKLKEQFKNKTIILNQSSGGDPEIVMHSFNC

Florida Patient E (FLP1E 6, 7)

FLPE6:

gaagagatag tgattagacc tgccaatttc acagacaatg ctaaagtcat aatagtacag ctgaatgcat ctgtagaaat taattgtaca agacccaaca acaatacaag aaaaggtata catataggac cagggagggc attctatgca acaggagaaa taataggaga tataagacaa gcacattgta acattagtgg agaaaaatgg aataatactt taaaacaggt agttacaaaa ttaagagaac aatttgggaa taaaacaata atctttaatc actcctcagg aggggaccca gaaattgt

EEIVIRPANFTDNAKVIIVQLNASVEINCTRPNNNTRKGIHIGPGRAFYATGEIIGDIRQAHCNISGEKWNNTLKQVVTK
LREQFGNKTIIFNHSSGGDPEIV

FLPE7:

gaagagatag taattagatc tgccaatttc acagacaatg ctaaagtcat aatagtacag ctggatgcat ctgtagaaat taattgtaca agacccaaca acaatacaag aaaaggtata catataggac cagggagggc attttatgca acaggagaaa taataggaga tataagacaa gcacattgta acattagtgg agaaaaatgg aataatactt taaaacaggt agttacaaaa ttaagagaac agtttgggaa taaaacaata atctttaatc actcctcagg aggggaccca gaaattgt

EEIVIRSANFTDNAKVIIVQLDASVEINCTRPNNNTRKGIHIGPGRAFYATGEIIGDIRQAHCNISGEKWNNTLKQVVTK
LREQFGNKTIIFNHSSGGDPEIV

Florida Patient F (FLPF 1, 6)

FLPF1:

gaagaggtag taattagatc tgaaaatttc atggacaatg ttaaaaccat aatagtgcag ctgaatgaat ctgtacaaat taattgtaca agacccaaca acaatacaag aaaaagtata catatagcac cggggagagc attttatgca acaggagaaa taataggaga tataagacaa gcacattgta accttagtag cataaaatgg aatgacactt taagacagat agctaaaaaa ttaaaagaac aatttggaaa taaaacaata atctttaatc aatcctcagg aggggaccca gaaatt

EEVVIRSENFMDNVKTIIVQLNESVQINCTRPNNNTRKSIHIAPGRAFYATGEIIGDIRQAHCNLSSIKWNDTLRQIAKK
LKEQFGNKTIIFNQSSGGDPEI

FLPF6:

gaagaggtag taattagatc tgaaaatttc aaggacaatg ttaaaaccat aatagtgcag ctgaatgaat ctgtgcaaat taattgtaca agacccaaca acaatacaag aaaaagtata catatagcac cggggagagc attttatgca acaggagaaa taataggaga tataagacag gcacattgta accttagtag cacaaaatgg aatgacactt taagacagac agctaaaaga ttaaaagaac aaattggaaa taaaacaata atctttaatc aatcctcagg aggggaccca gaaatt

NFKDNVKTIIVQLNESVQINCTRPNNNTRKSIHIAPGRAFYATGEIIGDIRQAHCNLSSTKWNDTLRQTAKRLKEQIGN
KTIIFNQSSGGDPEI

Florida Patient G (FLPG 1, 6)

FLPG1:

gaagaggtag taattagatc tgccaatttc acagacaatg ctaaaatcat aatagtacag ctgaatgcac ctgtagaaat taattgtaca agacccaaca acaatacaag aggaggtata catataggac cagggagagc attttatgca acagatagaa tagtaggaga tataagagaa gcatattgta acattagtag agaaaaatgg aataatactt taaaactggt agttacaaaa ttaagagaac aatttgtgaa taaaacaata atctttaatc actcctcagg aggggaccca gaaattgtaa tgcacagtgtt aattgtggagg ggaatttttct act

EEVVIRSANFTDNAKIIIVQLNAPVEINCTRPNNNTRGGIHIGPGRAFYATDRIVGDIREAYCNISREKWNNTLKLVVTK LREQFVNKTIIFNHSSGGDPEIVMHSVNCGGEFFY

FLPG6:

gaagaggtag taattagatc tgccaatttc acagacaatg ctaaaatcat aatagtacag ctgaatgcac ctgtagaaat taattgtaca agacccaaca acaatacaag aaaaggtata agtataggac cagggagagc attttatgca acagatagaa tagtaggaga tataagaaaa gcatattgta acattagtag agaaaaatgg aataatactt taaaactggt agttacaaaa ttaagagaac aatttgtgaa taaaacaata atctttaatc actcctcagg aggggaccca gaaattgtaa tgcacagtgt taattgtgga ggggaatttt tctact

EEVVIRSANFTDNAKIIIVQLNAPVEINCTRPNNNTRKGISIGPGRAFYATDRIVGDIRKAYCNISREKWNNTLKLVVTK LREQFVNKTIIFNHSSGGDPEIVMHSVNCGGEFFY

Florida Patient H (FLPH 2, 4)

FLPH2:

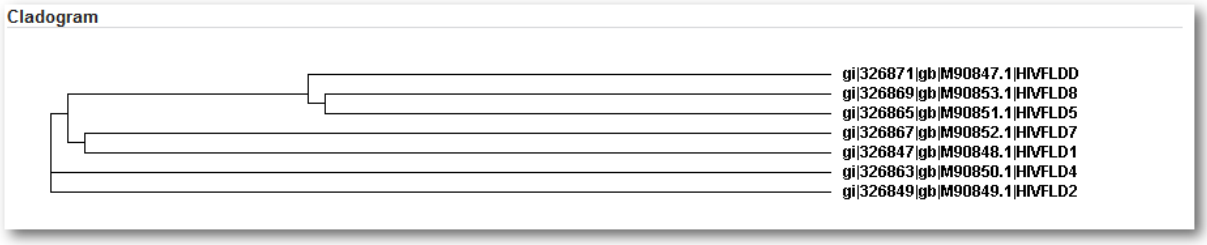
ctagcagaag gagaggtaat aattagatct gaaaatttca cggataatgc taagaccata atagtacagc tgaatgcaac tataaatatt acttgtgaaa gaccccacaa caatacaaga aaaagtatac atataggacc agggagggca ttttttgcaa caggagacat aacaggagat ataagacaag cacattgtaa ccttagtaaa ggagattggg ataacgcttt aaaacagata gttacaaaat taggagaaca atttgggagg aataaaacaa tagtctttaa gcaatcctca ggaggggacc cagaaattat aatgcacagt tttaattgtg caggggaatt ttcctactgt aat

LAEGEVIIRSENFTDNAKTIIVQLNATINITCERPHNNTRKSIHIGPGRAFFATGDITGDIRQAHCNLSKGDWDNALKQI VTKLGEQFGRNKTIVFKQSSGGDPEIIMHSFNCAGEFSYCN

FLPH4:

ctagcagaag gagaggtaat aattagatct gaaaatttca cggataatgc taaaaccata atagtacagc tgaatgcaac tataaacatt acttgtgaaa gaccccacaa caatacaaga agaagtatac atataggacc agggagagca ttttttgcaa caggagacat aacaggagat ataagacaag cacattgtaa ccttagtaga ggaggttggg ataacacttt aaaacagata gttacaaaat taagagaaca atttgggnnn aataaaacaa tagtctttaa tcaatcctca ggaggggacc cagaaattat aatgcacagt tttaattgtg caggggaatt tttctactgt aat

LAEGEVIIRSENFTDNAKTIIVQLNATINITCERPHNNTRRSIHIGPGRAFFATGDITGDIRQAHCNLSRGGWDNTLKQI
VTKLREQFGXNKTIVFNQSSGGDPEIIMHSFNCAGEFFYCN



Florida Dentist (FLD 1, 7)

FLD1:

ctagcagaag aagaggtagt aattagatct gccaatttca cagacaatgc taaaatcata atagtacagc tgaatgcatc tgtagaaatt aattgtacaa ggcccaacaa caatacaaga aaaggtatac atataggacc agggagagca ttttatgcaa caggagaaat aataggagat ataagacaag cacattgtaa cattagtaga gaaaaatgga ataatacttt aaaccaggta gttacagaat taagggaaca atttgggaat aaaacaataa cctttaatca ctcctcagga ggggacccag aaattgtaat gcacagtttt aattgtggag gggaattttt ctattgtaat

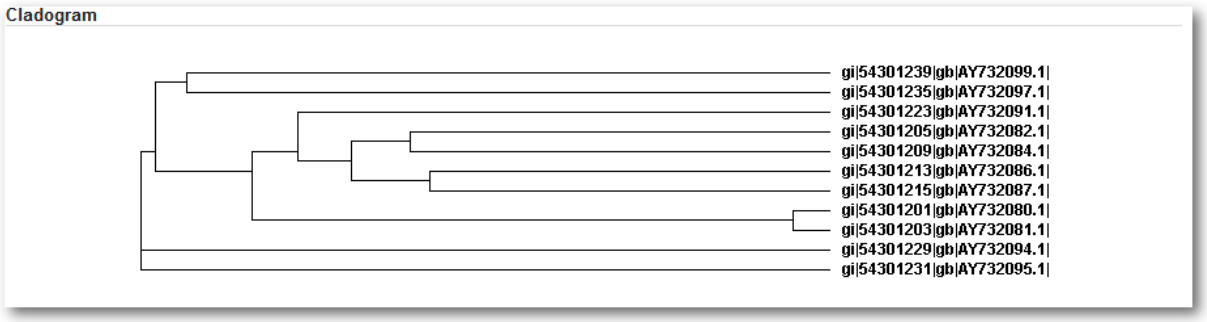
LAEEEVVIRSANFTDNAKIIIVQLNASVEINCTRPNNNTRKGIHIGPGRAFYATGEIIGDIRQAHCNISREKWNNTLNQV
VTELREQFGNKTITFNHSSGGDPEIVMHSFNCGGEFFYCN

FLD7:

ctagcagaag aagaggtagt aattagatct gccaatttca cagacaatgc taaaatcata atagtacagc tgaatgcatc tgtagaaatt aattgtacaa ggcccaacaa caatacaaga aaaggtatac atataggacc agggagagca ttttatgcaa caggagaaat aataggagat ataagacaag cacattgtaa cattagtaga gaaaaatgga ataatacttt aagacaggta gttacaaaat taagagaaca atttgggaat aaaacaataa tctttaatca ctcctcagga ggggacccag aaattgtaat gcacagtttt aattgtggag gggaattttt ctactgtaat

LAEEEVVIRSANFTDNAKIIIVQLNASVEINCTRPNNNTRKGIHIGPGRAFYATGEIIGDIRQAHCNISREKWNNTLRQV
VTKLREQFGNKTIIFNHSSGGDPEIVMHSFNCGGEFFYCN

Notes: From this group on are the control groups. The following phylogenetic trees ***do not*** mean much, as far as the topology goes. They merely show the sequence IDs that I've downloaded from NCBI. This is because the sequences were obtained from different individuals. For example, note that sample AY732080 is very close to AY732081. They were sequences accidentally chosen from two clones of the same person.



China (AY 732095, 732099)

AY 732095:

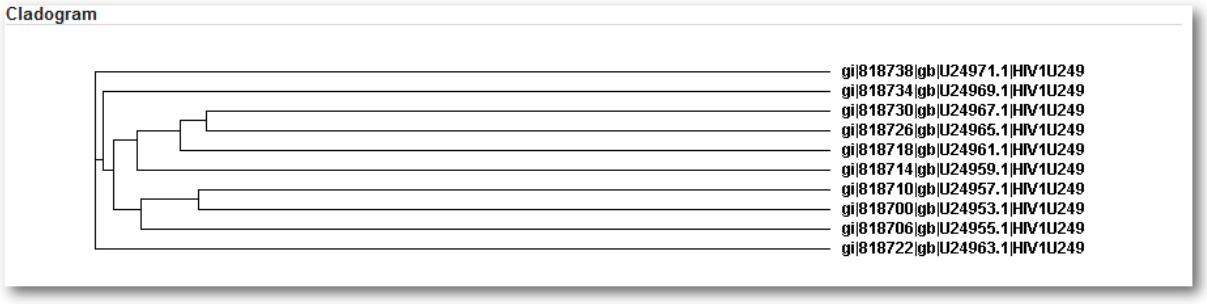
ttctcggaca atgctaaagt cataatagta cagctgaata aatctgtaga aattaattgt acaagaccta acaacaatac aagaaaaagt atacatctag gacaagggaa agcatggtat acaacagaaa taataggaga tataagacaa gcacattgta cattagtatg gaataacact ttaaaacaga taactgaaaa attaagaga

FSDNAKVIIVQLNKSVEINCTRPNNNTRKSIHLGQGKAWYTTEIIGDIRQAHCTLVWNNTLKQITEKLR

AY732099:

ttctcggaca atgctaaagt cataatagta cagctgaatg aatctgtaga aattaattgt acaagaccta acaacaatac aagaaaaagt atacatctag gacaagggaa agcatggtat acaacagaaa taataggaga tataagacaa gcacattgta cattagtatg gaataacact ttaaaacaga taactgaaaa actaagaga

FSDNAKVIIVQLNESVEINCTRPNNNTRKSIHLGQGKAWYTTEIIGDIRQAHCTLVWNNTLKQITEKLR



Europe (U 24967, 24953)

U24967:

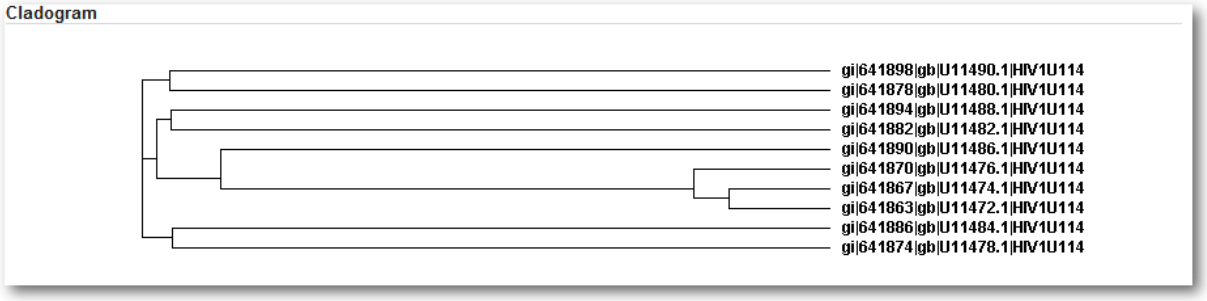
tagcagaaga agaggtagta attaggtctg aaaatttcac gaacaatgct aaaaccataa tagtacagct gaaaaaacct gtagaaatta attgcataag acccaacaac aatacaagaa aaggtataca tataggacca gggagagcat tttatacaac aggagaaata ataggaaata taagacaagc acattgtaac cttagtagag cagaatggaa tgacacctta aaacagatag ttgtcaaatt aggagaacaa tttaagaata caa

AEEEVVIRSENFTNNAKTIIVQLKKPVEINCIRPNNNTRKGIHIGPGRAFYTTGEIIGNIRQAHCNLSRAEWNDTLKQIV VKLGEQFKNT

U24953:

ggcagtctag cagaagaaga ggtagtaatt agatctgaaa atttcacaaa caatgctaaa agcataatag tacagctgaa tgaaactgta gaaattaatt gtacaagacc caacaacaat acaagaaaag gtatacatat aggaccaggc aaagcatttt atgcaacagg agatataata ggagatataa gacaagctca ttgtaacatc agtagagcaa aatggaatga cactttaaga cagatagcta tcaaattaag agaacaattt aagaataaaa caatagtctt taatcaatcc tcaggagggg acccagaaa

GSLAEEEVVIRSENFTNNAKSIIVQLNETVEINCTRPNNNTRKGIHIGPGKAFYATGDIIGDIRQAHCNISRAKWNDTLR QIAIKLREQFKNKTIVFNQSSGGDPE



Liberty City (U 11490, 11472, 11478, 11488)

U11490:

aatctcacgg acaatgctaa aaccataata gtacatttaa ataaatctgt agtgattaat tgtacaagac ccaacaacaa tacaataaaa agtatacgca taggaccagg gcgagcatgg tatacaacag gagaaataac aggagatata agacaagcac attgtaacct tagtagagca gactggaata acactttaag acaggtagtt atgaaactaa gagaacactt taaaaataaa acaatagtct ttaatcaatc ctcaggaggg gacccagaaa ttgtaatgca cagttttaat tgtggagggg aatttttc

NLTDNAKTIIVHLNKSVVINCTRPNNNTIKSIRIGPGRAWYTTGEITGDIRQAHCNLSRADWNNTLRQVVMKLREHFK NKTIVFNQSSGGDPEIVMHSFNCGGEFF

U11472:

gaagaggtag taattagatc tgccaatttc acagataata ctaaaatcat aatagtacag ctgaaggaat ctgtagaaat taattgtaca aggcccaaca acaatacaag aagaagtata aatataggac cagggagagc attttatgca acaggagata taataggaaa tataaggcaa gcacactgca acattagtag agcaaaatgg tttgatgctt taaaacaggt agctggaaaa ttaagagaac aatttgataa taaaacaata gcctttaatc aatcctcagg aggggaccta

EEVVIRSANFTDNTKIIIVQLKESVEINCTRPNNNTRRSINIGPGRAFYATGDIIGNIRQAHCNISRAKWFDALKQVAGK LREQFDNKTIAFNQSSGGDL

U11478:

aatttcacaa acaatgctaa aaccataata gtacagctga atgaaactgt agaaattaat tgtacaagac ccaacaacaa tacaagaaaa agcatacata taggaccagg cagagcattt tttacaacag gagatataat aggagacata agacaagcac attgtaacat tagtagagca agatggaatg aaactttaaa cagaatagtt acaaaattaa gagaacaatt tgggaataat aaaacaatag tctttaatca ctcctaccca ggaggggacc cagaagttgt aacacacagt tttaattgtg gaggggaatt tttctactgt aattcaaca

NFTNNAKTIIVQLNETVEINCTRPNNNTRKSIHIGPGRAFFTTGDIIGDIRQAHCNISRARWNETLNRIVTKLREQFGN
NKTIVFNHSYPGGDPEVVTHSFNCGGEFFYCNST

U11488:

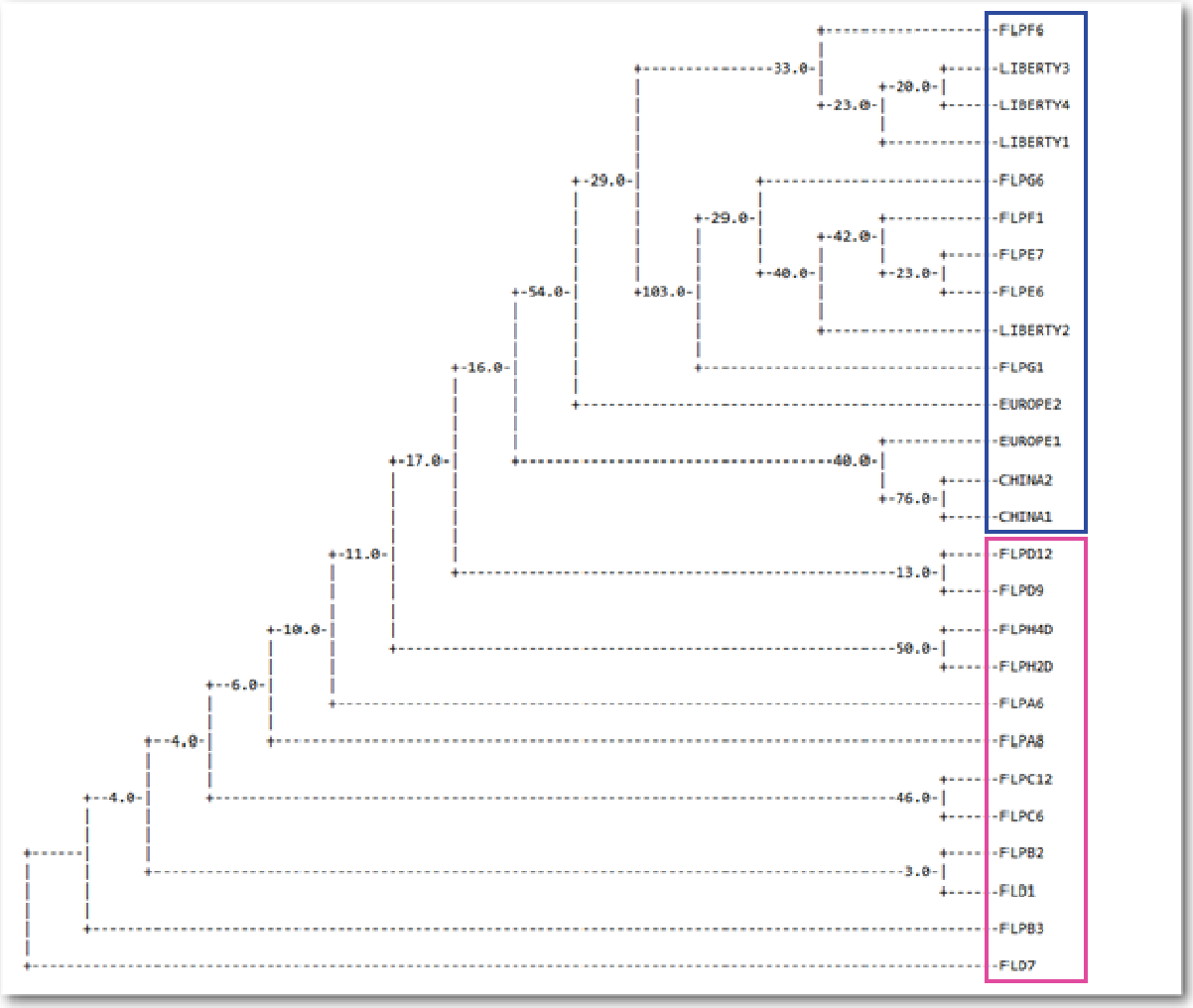
aatttcacaa acaatgctaa aaccataata gtacagctga atgaagctgt agtaattaat tgtacaagac ccaacaacaa tacaagaaaa ggtatacata taggaccagg gagagcattc tatgcaacag gagacataat aggagatata agacaagcac attgtaacct tagtaaagtg gcatggaatg aaactttaaa aaaggtagtt gaaaaattaa gagaacaatt taagaagaaa ataatagtct ttaattcatc ctcaggaggg gacccagaaa ttgtaactca cagttttaat tgtggagggg aatttttcta ctgtaataca

NFTNNAKTIIVQLNEAVVINCTRPNNNTRKGIHIGPGRAFYATGDIIGDIRQAHCNLSKVAWNETLKKVVEKLREQFK
KKIIVFNSSSGGDPEIVTHSFNCGGEFFYCNT

Distance Matrix method

Description: The following consensus tree was obtained using the Neighbor program, neighbor-joining option, with bootstrap value 120, and seed value 13. The analysis was done on the distance matrix of the protein sequences of the selected representatives from each group.

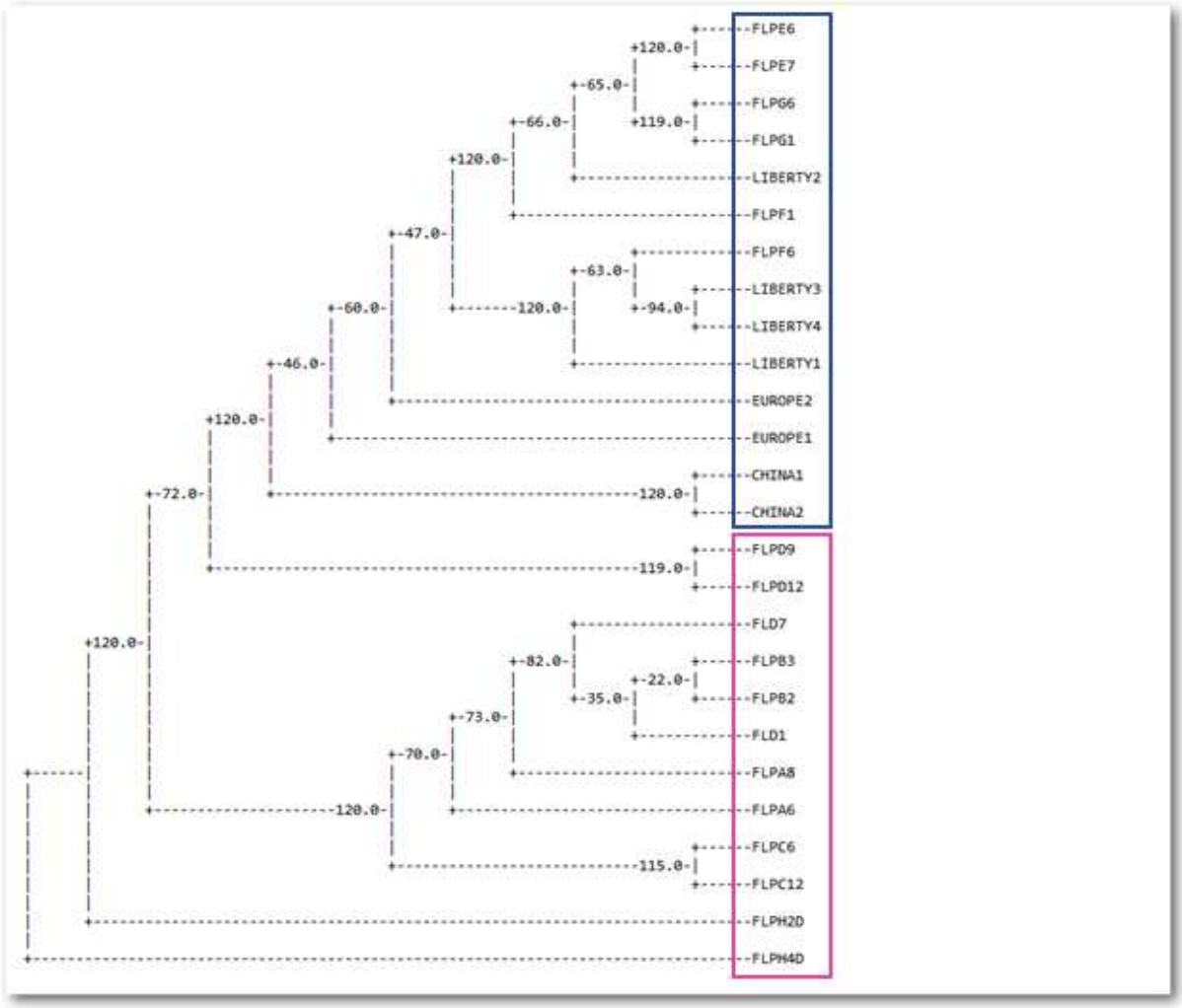
Since bootstrap value 120 was used to generate the result. The following consensus tree is the final representation of 120 distance matrices produced by the neighbor-joining method using the protein sequences.



Neighbor NJ method (protein sequences)

Description: The following consensus tree was obtained using the Neighbor program, UPGMA option, with bootstrap value 120, and seed value 13. The analysis was done on the distance matrix of the protein sequences of the selected representatives from each group.

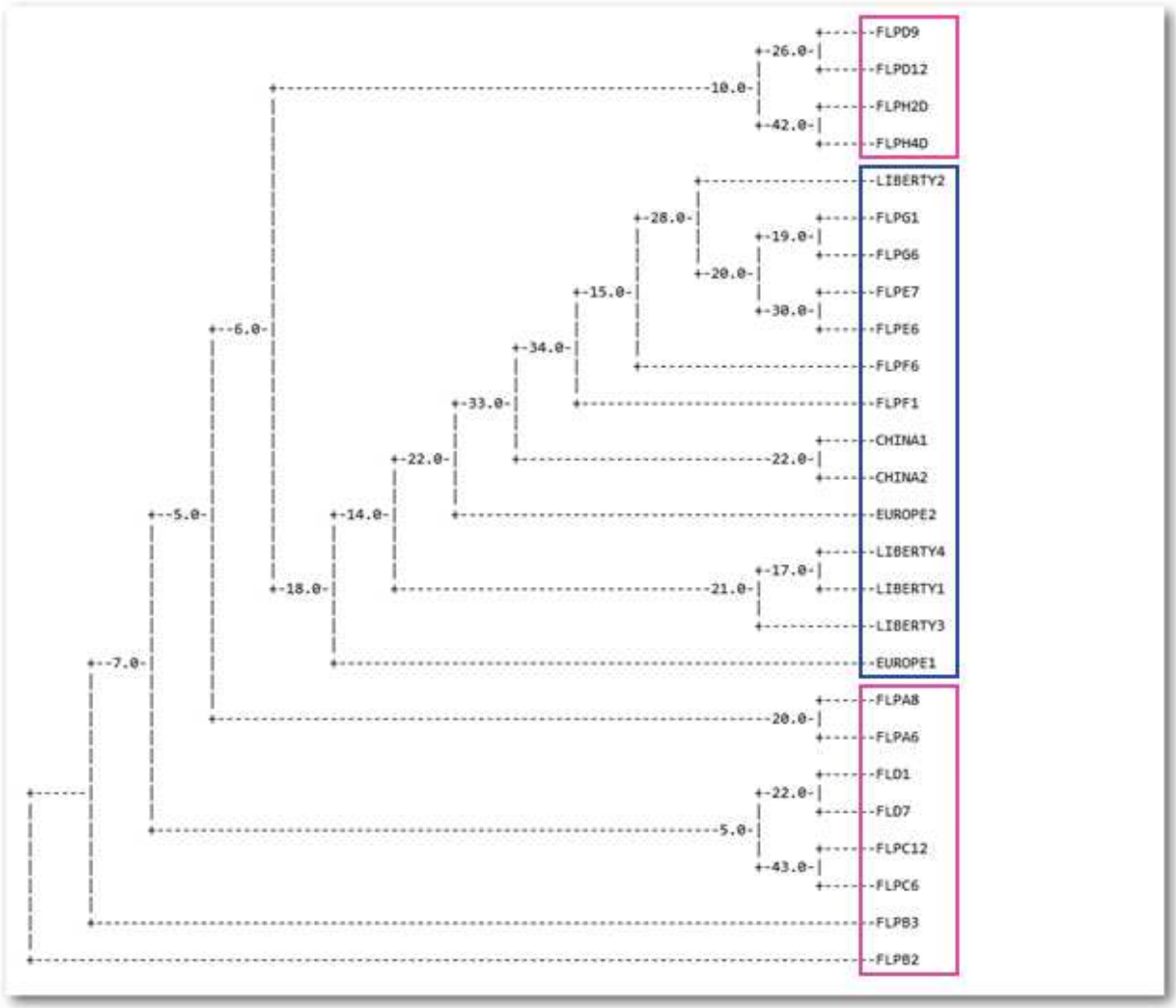
Since bootstrap value 120 was used to generate the result. The following consensus tree is the final representation of 120 distance matrices produced by UPGMA using the protein sequences.



Neighbor UPGMA method (protein sequences)

Description: The following consensus tree was obtained using the Neighbor program, neighbor-joining option, with bootstrap value 120, and seed value 13. The analysis was done on the distance matrix of the DNA sequences of the selected representatives from each group.

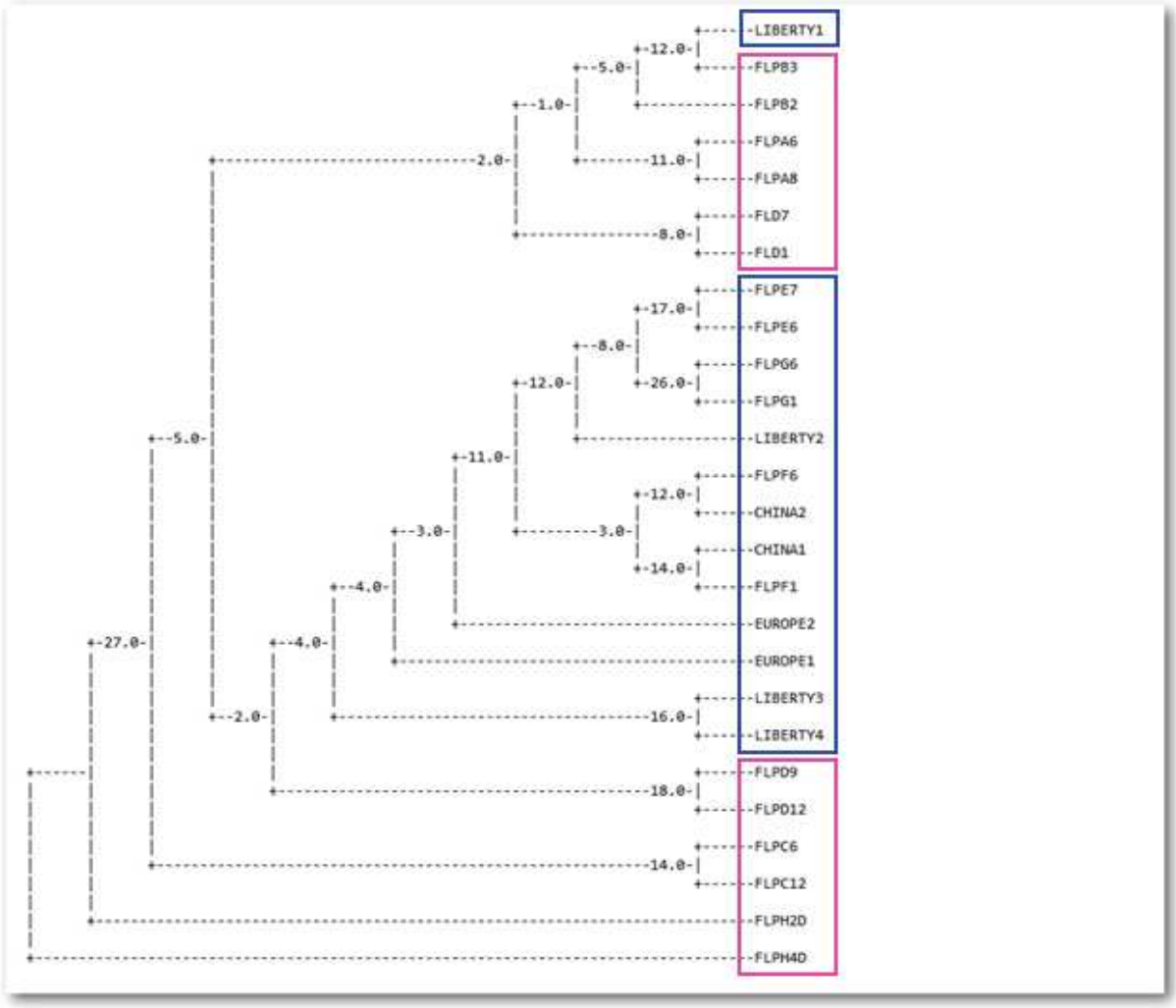
Since bootstrap value 120 was used to generate the result. The following consensus tree is the final representation of 120 distance matrices produced by the neighbor-joining using the DNA sequences.



Neighbor NJ method (DNA sequences)

Description: The following consensus tree was obtained using the Neighbor program, UPGMA option, with bootstrap value 120, and seed value 13. The analysis was done on the distance matrix of the DNA sequences of the selected representatives from each group.

Since bootstrap value 120 was used to generate the result. The following consensus tree is the final representation of 120 distance matrices produced by UPGMA using the DNA sequences.

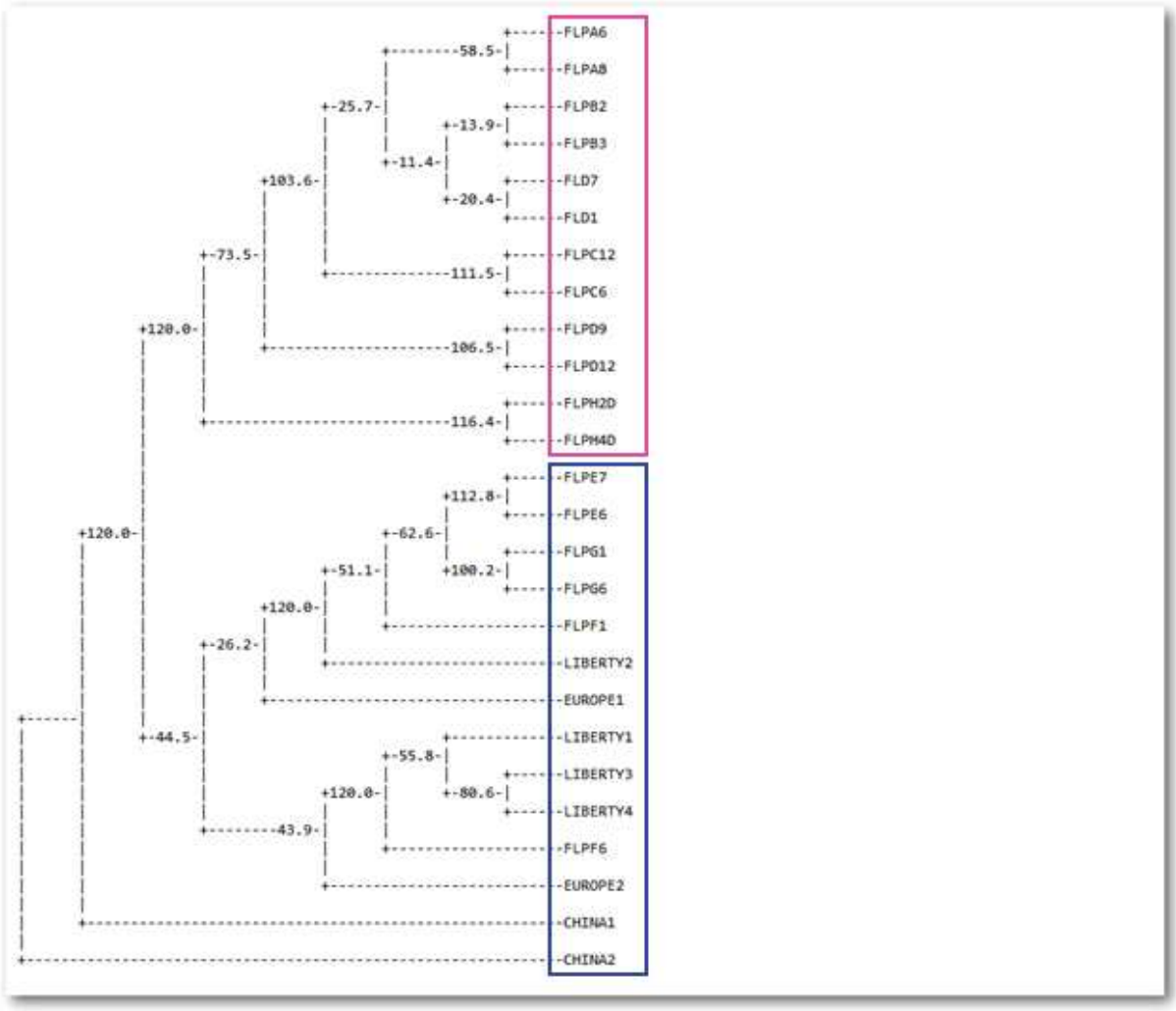


Neighbor UPGMA method (DNA sequences)

Parsimony method

Description: The following consensus tree was obtained using the Protpars program, a character based parsimony method, with bootstrap value 120, and seed value 13. The analysis was done on the distance matrix of the protein sequences of the selected representatives from each group.

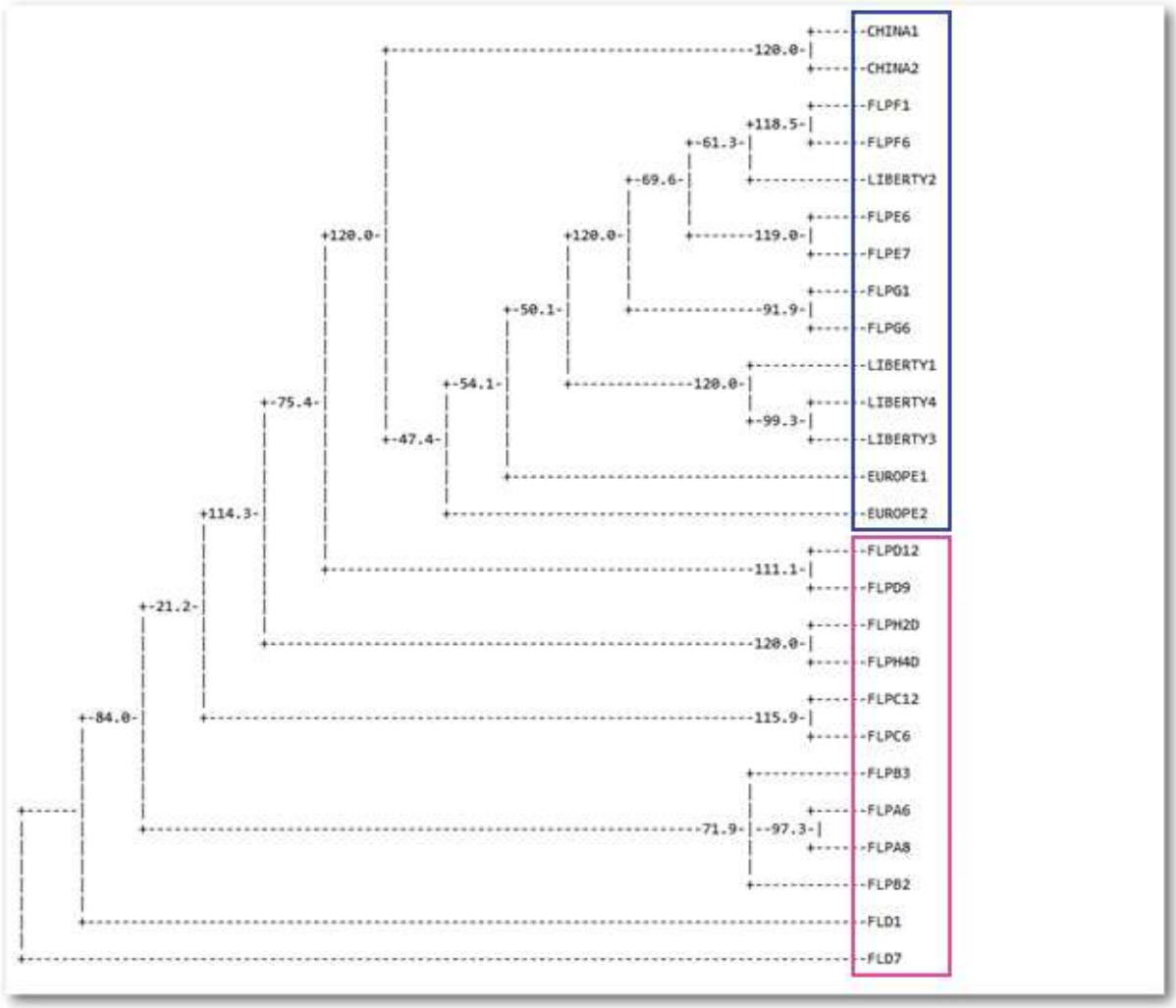
Since bootstrap value 120 was used to generate the result. The following consensus tree is the final representation of 120 phylogenic trees produced by character based Parsimony method using the protein sequences.



Protpars (protein sequences)

Description: The following consensus tree was obtained using the DNAtpars program, a character based parsimony method, with bootstrap value 120, and seed value 13. The analysis was done on the distance matrix of the DNA sequences of the selected representatives from each group.

Since bootstrap value 120 was used to generate the result. The following consensus tree is the final representation of 120 phylogenic trees produced by character based Parsimony method using the DNA sequences.



DNApars (DNA sequences)

Conclusions

Case Review

Human Immunodeficiency Virus (HIV) is a virus with a single-stranded RNA genome. Because RNA replication is highly error prone when compared to DNA replication, the HIV virus is constantly mutating. Many of these nucleotide changes result in non-functional viruses, but some produce viable viruses with altered cell surface antigens. This represents a significant challenge to producing an effective HIV vaccine.

In the late 1980's, eight patients of an HIV-positive dentist in Florida were diagnosed as being HIV-positive. Though many of the patients had had invasive dental procedures performed, an investigation did not uncover systematic hygienic lapses that might account for infection of the patients. Additionally, there were no obvious ways in which the dentist might have deliberately infected his patients.

In an attempt to determine whether the eight HIV-positive patients were infected by the dentist, researchers isolated viral RNA from blood samples from the dentist, the infected patients, and HIV-positive individuals in the area who had had no contact with the dentist. The investigators then amplified DNA copies of the genomic RNA sequences via the polymerase chain reaction (P.C.R.) and determined the nucleotide sequence of pieces of the HIV gp120 (V3 region) gene. This data was then used to determine how closely related the dentist's HIV virus strain was to that of his patients, and the HIV-positive individuals who had not had contact with the dentist.

Sequences

In my study, I first downloaded a various numbers of sequences of different colons from the eight patients and the dentist. For the control group, I also downloaded sequences of the same gene, HIV-1 V3 region, of samples from irrelevant individuals in Liberty city Florida as well as samples from irrelevant areas, China and Europe.

To clean up the dataset, I used Clustalw to find the best representatives for each of the experimental group. I selected two sequences from each experimental group. For control groups, since they are not from the same person, the sequences are not really relevant with each other. I picked two from China group, Europe group, and four from Liberty city group.

After deciding the DNA sequences to use for my analysis, I obtained the corresponding protein sequences from NCBI's Gene bank. These information is shown in the first section of this report.

Clustalw

Clustalw is used for multiple sequence alignment to find the best representatives of sequences from each experimental group. Program takes the input DNA sequences in the FASTA format and output the phylogenetic trees. For the phylogenetic trees, I was able to exclude the sequences that are not so close to the other sequences in the same group. Since the selected sequences would represent the particular experimental group, I would select from the ones that are close with each other.

While for the control groups, the sequences were not from the same individual, they are far apart from each other to start with. So I randomly selected a couple of them.

Dnadist

Dnadist is used to generate distance matrices for further analysis of DNA similarities. This program used nucleotide sequences to compute a distance matrix. The distance for each pair of species estimates the total branch length between the two species, and was used in the distance matrix programs Neighbor later in the analysis.

The program reads in nucleotide sequences and writes an output file containing the distance matrix. Although the program correctly takes into account a variety of nucleotide sequence ambiguities, I cleaned up the input sequences to be the same length of nucleotides.

Input:

```
  26      209
FLPA6
           CTAGCAGAAGAAGAGGTAGTAATTAGATCTGCCAATTTCACAGACAATGCTAAAATCATAATAGTACAACTGAATGCA
           TCTGTAAAAATTAAATGTACAAGACCCAACAACAATACAAGAAAAGGTATACAGATAGGACCAGGAAGGGCATTTTAT
           GCAACAGGAGAAATAATAGGAGATATAAGACAAGCACATTGTAACATTAGTAG
FLPA8
           CTAGCAGAAGAAGAGGTAGTAATTAGATCTGCCAATTTCACAGACAATGCTAAAATCATAATAGTACAACTGAATGCA
           TCTGTAGAAATTAATTGTACAAGACCCAACAACAATACAAGAAAAGGTATACAGATAGGACCAGGAAGGGCATTTTAT
           GCAACAGGAGAAATAATAGGAGATATAAGACAAGCACATTGTAACATTAGTAG
:          :
:          :
```

Output:

```
  26
FLPA6
            0.000000  0.004798  0.024279  0.024279  0.070499  0.070499  0.178675
            0.155193  3.503351  3.459452  3.103315  3.277940  3.045990  3.010856
            0.101572  0.096083  0.039137  0.039137  2.320965  2.337942 -1.000000
           -1.000000 -1.000000  2.995141  3.607538  4.268605
FLPA8
            0.004798  0.000000  0.019314  0.019314  0.065077  0.065077  0.172135
            0.155732  3.619171  3.576790  3.199394  3.384838  3.138913  3.097450
            0.095919  0.090488  0.034051  0.034051  2.396207  2.412907  5.352246
           -1.000000 -1.000000  3.073460  3.364575  3.910573
:           :
:           :
```

Protdist

Protdist is used to generate distance matrices for further analysis of protein similarities. This program used protein sequences to compute a distance matrix. The distance for each pair of species estimates the total branch length between the two species, and was used in the distance matrix program Neighbor later in the analysis.

The program reads in protein sequences (amino acid sequences) and writes an output file containing the distance matrix. Although the program correctly takes into account a variety of amino acid sequence ambiguities, I cleaned up the input sequences to be the same length of amino acids.

Input:

```
   26      69
FLPA6       LAEEEVVIRSANFTDNAKIIIVQLNASVKIKCTRPNNNTRKGIQIGPGRAFYATGEIIGDIRQAHCNIS
FLPA8       LAEEEVVIRSANFTDNAKIIIVQLNASVEINCTRPNNNTRKGIQIGPGRAFYATGEIIGDIRQAHCNIS
:           :
:           :
```

Output:

```
  26
FLPA6
           0.000000  0.026478  0.026478  0.026478  0.039906  0.039906   0.244764
           0.238927 38.766636 38.921718 35.849373  4.987060 39.306367  39.291311
           0.202290  0.202290  0.026478  0.026478  3.059512  3.089455  36.938763
           8.837461  6.216181 38.206270  6.510981  5.416667
FLPA8
           0.026478  0.000000  0.000010  0.000010  0.013122  0.013122   0.266059
           0.271453 38.893663 39.043006 36.192015  5.269274 39.414798  39.400308
           0.206315  0.206315  0.000010  0.000010  3.192338  3.223065  36.952001
           9.146617  6.074596 38.360603  6.939067  5.702482
:          :
:          :
```

Neighbor

Neighbor is used to generate the guide trees. Program takes the distance matrix, which is computed based on the similarity of DNA sequences or protein sequences, as the input and constructs a visual friendly tree structure to represent the phylogenetic relationships among the taxa. Taxa here are DNA nucleotide and protein amino acid sequences.

In this analysis, two variants of algorithms were used, UPGMA and Neighbor-joining. They are both iterative algorithms. Basically, program based on the current distance matrix calculates a heuristic value, finds the pair of taxa with the lowest value, and creates a node on the tree than joins these two taxa. Then program calculates the distances versus this new node, and do it until the tree is completely constructed. Neighbor-joining is a more evolved method, while UPGMA is the simplest of the distance method.

Input:

```
  26
FLPA6
          0.000000  0.026478  0.026478  0.026478  0.039906  0.039906   0.244764
          0.238927 38.766636 38.921718 35.849373  4.987060 39.306367  39.291311
          0.202290  0.202290  0.026478  0.026478  3.059512  3.089455  36.938763
          8.837461  6.216181 38.206270  6.510981  5.416667
FLPA8
          0.026478  0.000000  0.000010  0.000010  0.013122  0.013122   0.266059
          0.271453 38.893663 39.043006 36.192015  5.269274 39.414798  39.400308
          0.206315  0.206315  0.000010  0.000010  3.192338  3.223065  36.952001
          9.146617  6.074596 38.360603  6.939067  5.702482
:         :
:         :
```

Output: There were several ways to represent the output phylogenetic tree. The following output data is in the Newick standard form, which is the input format of tree structures to many programs.

```
((((((((((FLPF6:120.0,((LIBERTY3:120.0,LIBERTY4:120.0):20.0,LIBERTY1:120.0):23.0):33.0,
((FLPG6:120.0,((FLPF1:120.0,(FLPE7:120.0,FLPE6:120.0):23.0):42.0,LIBERTY2:120.0):40.0):29.
0,FLPG1:120.0):103.0):29.0,EUROPE2:120.0):54.0,(EUROPE1:120.0,(CHINA2:120.0,CHINA1:120.0):
76.0):40.0):16.0,(FLPD12:120.0,FLPD9:120.0):13.0):17.0,(FLPH4D:120.0,FLPH2D:120.0):50.0):1
1.0,FLPA6:120.0):10.0,FLPA8:120.0):6.0,(FLPC12:120.0,FLPC6:120.0):46.0):4.0,(FLPB2:120.0,
FLD1:120.0):3.0):4.0,FLPB3:120.0):120.0,FLD7:120.0);
```

Parsimony Protpars and DNApars were used to build the "evolution" trees. As character based guide tree generating programs, they infer an unrooted phylogeny from protein amino acid sequences or DNA nucleotide sequences. They allow any amino acid or nucleotide to change to any other, and counts the number of such changes needed to evolve the protein sequences or DNA sequences on each given phylogeny Neighbor is used to generate the guide trees.

The input protein alignment of my analysis contained sequences that were quite different from each other, considering the control groups. In addition, I did a 120 bootstrap replicates execution, the programs took long time to finish, and the protpars.outfile itself was 38 Mb large. Character based heuristic methods are more suitable for taxa with strong similarity.

Input:

```
  26      69
FLPA6       LAEEEVVIRSANFTDNAKIIIVQLNASVKIKCTRPNNNTRKGIQIGPGRAFYATGEIIGDIRQAHCNIS
FLPA8       LAEEEVVIRSANFTDNAKIIIVQLNASVEINCTRPNNNTRKGIQIGPGRAFYATGEIIGDIRQAHCNIS
:           :
:           :
```

OR

```
   26      209
FLPA6
           CTAGCAGAAGAAGAGGTAGTAATTAGATCTGCCAATTTCACAGACAATGCTAAAATCATAATAGTACAACTGAATGCA
           TCTGTAAAAATTAAATGTACAAGACCCAACAACAATACAAGAAAAGGTATACAGATAGGACCAGGAAGGGCATTTTAT
           GCAACAGGAGAAATAATAGGAGATATAAGACAAGCACATTGTAACATTAGTAG
FLPA8
           CTAGCAGAAGAAGAGGTAGTAATTAGATCTGCCAATTTCACAGACAATGCTAAAATCATAATAGTACAACTGAATGCA
           TCTGTAGAAATTAATTGTACAAGACCCAACAACAATACAAGAAAAGGTATACAGATAGGACCAGGAAGGGCATTTTAT
           GCAACAGGAGAAATAATAGGAGATATAAGACAAGCACATTGTAACATTAGTAG
:          :
:          :
```

Output: There were again different ways to represent the output phylogenetic tree. The following output data is in the Newick standard form. The consensus tree was shown in previous sections.

```
(((((((FLPA6:120.0,FLPA8:120.0):58.5,((FLPB2:120.0,FLPB3:120.0):13.9,(FLD7:120.0,FLD1:120
.0):20.4):11.4):25.7,(FLPC12:120.0,FLPC6:120.0):111.5):103.6,(FLPD9:120.0,FLPD12:120.0):10
6.5):73.5,(FLPH2D:120.0,FLPH4D:120.0):116.4):120.0,((((((FLPE7:120.0,FLPE6:120.0):112.8,(F
LPG1:120.0,FLPG6:120.0):100.2):62.6,FLPF1:120.0):51.1,LIBERTY2:120.0):120.0,EUROPE1:120.0)
:26.2,(((LIBERTY1:120.0,(LIBERTY3:120.0,LIBERTY4:120.0):80.6):55.8,FLPF6:120.0):120.0,EURO
PE2:120.0):43.9):44.5):120.0,CHINA1:120.0):120.0,CHINA2:120.0);
```

Bootstraping: Bootstrapping is a bias-reducing procedure in which the phylogenic analysis programs build an alignment of pseudo-sequences by picking residue positions at random and stringing the residues at those positions together until the sequence is the same length as the original alignment. From this pseudo-sequence alignment, it determines the relative number of sequence difference among the sample sequences, as determined from a random sampling of their sequences. This process was repeated, 120 times in each case of my analysis, to make 120 outputs. The tree that was ultimately produced represents a consensus of the 120 outputs.

Analysis: It should be clear from these phylogenetic trees that the dentist's strains are close to patients A, B, C, D, and H; they are as similar as different mutations in the dentist. The dentist's strains are not so close to E, F, and G, or the controls, of course. The distances of DNA and protein sequences between patients E, F, G and the dentist are just as far as the each of them versus any one from the control groups. This relationship is marked out in the phylogenetic trees in the previous sections. The red box contains the sequences from the dentist and patients A, B, C, D, and H. The blue box contains the sequences from the control groups and patients E, F, and G.