

Comparative computational Analysis of HIV-1 sequence variants in Africa and Europe

Abstract:

The human immunodeficiency virus (HIV) is a retrovirus has history of harmful effects lead to million cases of deaths as it causes AIDs disease when attacking the immune system. HIV-1 as other RNA viruses get mutations over time. Therefore, in this study we compare the consensus sequence of 10 HIV complete genome sequences from Africa with the consensus sequence of 10 complete genome sequences from Europe. Using our python script, we will extract the data from NCBI and do MSA alignment using ClustalW. Additionally, obtaining the consensus sequence using our already made consensus function. Then run the alignment again to get the dissimilar region and drawing the phylogenetic tree to show the relation between the sequences. After observed different functional products/interpretations using Pfam between the two sequences and results showed that dissimilarity between sequences may affect the functional products as well.

Keywords: HIV/AIDS, HIV complete genome, dissimilar region detection, and multiple genome sequencing, phylogenetic tree, CG content.

Introduction:

Human immunodeficiency virus type-1 (HIV-1), the causative agent of AIDS, is blamed for over 34 million deaths and is poised to claim over two million lives a year in the absence of efficient therapeutic intervention.^[1] Once people get HIV, they have it for life. Untreated HIV replication causes progressive CD4+ T cell loss and a wide range of immunological abnormalities, leading to an increased risk of infectious and oncological complications. But with proper medical care, HIV can be controlled. People with HIV who get effective HIV treatment can live long, healthy lives and protect their partners. ^[2]

Each viral particle of HIV-1 contains two copies of the full-length viral genomic RNA. Encapsidating two copies of genomic RNA is one of the characteristics of the retrovirus family. The two RNA molecules are both positive-sense and often identical; furthermore, each RNA encodes the full complement of genetic information required for viral replication. The two strands of RNA are intricately entwined within the core of the mature infectious virus as a ribonuclear complex with the viral proteins, including nucleocapsid.^[3] HIV-1 as any RNA virus it forms mutations while transmitting. These mutations may affect its functional products. Therefore, in this study we compare HIV-1 sequences from two different continents, Africa and Europe, to highlight the probable effect that may result using full automated method by python script.

Material and methods:

All computational steps are done using python 3^[4]

Data collection:

We downloaded twenty whole genome sequences for HIV-1 from Africa and Europe, ten sequences for each continent, by the help of Entrez function from NCBI virus database. Then, we created 2 new files to save our sequences separately.

Sequence alignment:

Multiple alignments were performed using Clustalw Command Line. As, we aligned all sequences from Africa to get only one consensus sequence and did the same for Europe. We aligned these two consensus sequences for detecting the dissimilar regions. The last alignment was for the consensus sequence of Africa and all Europe sequences which will be used in the upcoming steps

Dissimilar regions extraction:

We created a function to extract the index of dissimilar regions and also return its original sequence and the dissimilar one

CG content calculations:

We created a function to get CG content and represented the results in a scatter plot

Drawing Phylogenetic tree:

By using phylo packages we got phylogenetic tree for the aligned file of the consensus sequence of Africa and all sequences of Europe.

After that, where we have imported os packages to remove unnecessary files since it is no longer needed to make the scrip produce only the needed files.

Functional products/interpretations:

After obtaining the dissimilar region data we imported our 2 consensus sequences to NCBI ORF finder tool to get all open reading frames in our 2 sequences and their corresponding functional products using pfam website.

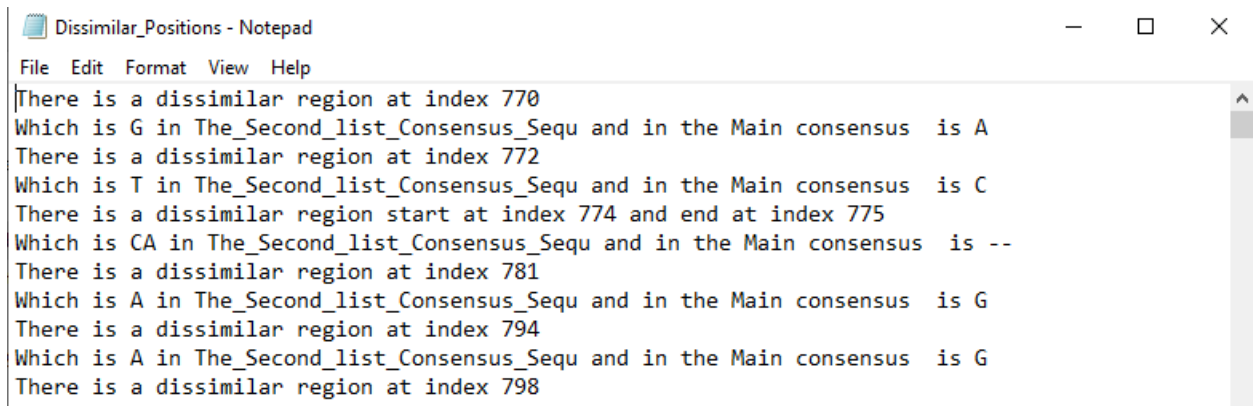
Results and Discussion:

Alignment: All performed alignments was saved in separate files

Dissimilar regions:

All indexes of dissimilar regions and also its original sequence and the dissimilar one is saved in a Dissimilar_positions file which automatically get downloaded after running the code as shown in figure1.

The results shows that although the sequences represent the same virus but as there are from different continents the sequences represent a number of dissimilar regions along the whole genome



```
File Edit Format View Help
There is a dissimilar region at index 770
Which is G in The_Second_list_Consensus_Sequ and in the Main consensus is A
There is a dissimilar region at index 772
Which is T in The_Second_list_Consensus_Sequ and in the Main consensus is C
There is a dissimilar region start at index 774 and end at index 775
Which is CA in The_Second_list_Consensus_Sequ and in the Main consensus is --
There is a dissimilar region at index 781
Which is A in The_Second_list_Consensus_Sequ and in the Main consensus is G
There is a dissimilar region at index 794
Which is A in The_Second_list_Consensus_Sequ and in the Main consensus is G
There is a dissimilar region at index 798
```

Figure 1. extracting the dissimilar region in separate file

CG content:

The below scatter plot represents the CG content of the consensus sequence of Africa and all sequences of Europe shows that sequence with accession no. MZ327295.1 has the highest CG content and sequence with accession no. MZ32729.1 has the lowest and each sequence has different value.

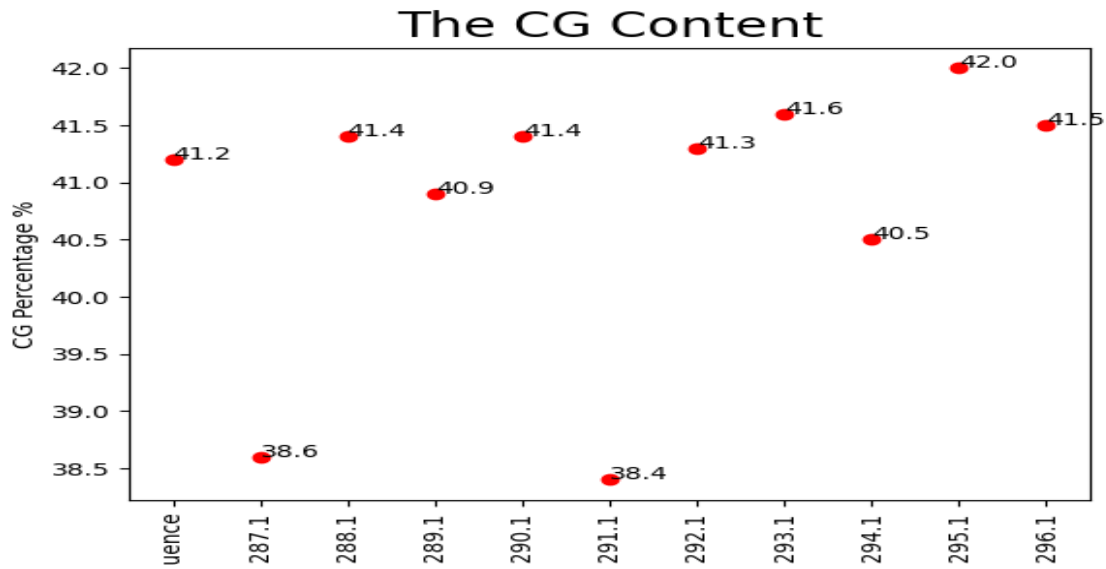


Figure 2 The scatter plot represents the CG content of the consensus sequence of Africa and all sequences of Europe

The phylogenetic tree:

The below figure shows the phylogenetic tree of the consensus sequence of Africa and all sequences of Europe and represent the relationships of these sequences. Also, as shows that the most related sequences of the consensus sequence of Africa are MZ327292.1 and MZ327290.1. while in the opposite way MZ327287.1 and MZ327294.1 take place.

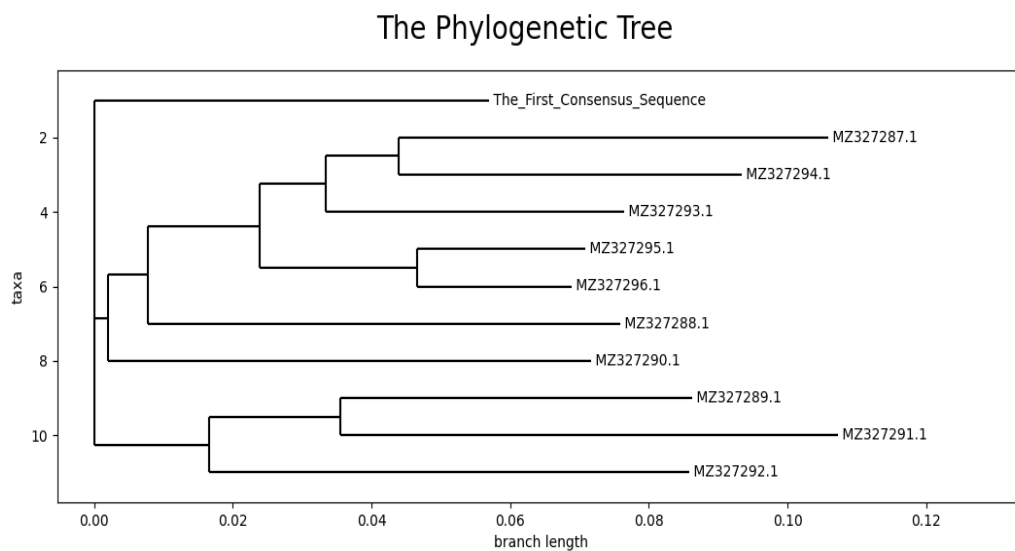


Figure 3. shows the phylogenetic tree of the consensus sequence of Africa and all sequences of Europe

Functional products/interpretations:

In this study, we compared four predicted ORFs of the resulted two consensus sequences, the first from Africa and the second from Europe.

NCBI orf finder website results show that for the first consensus sequence ORF1 correspond ORF15 in the second sequence as shown in figures 5 and 6.

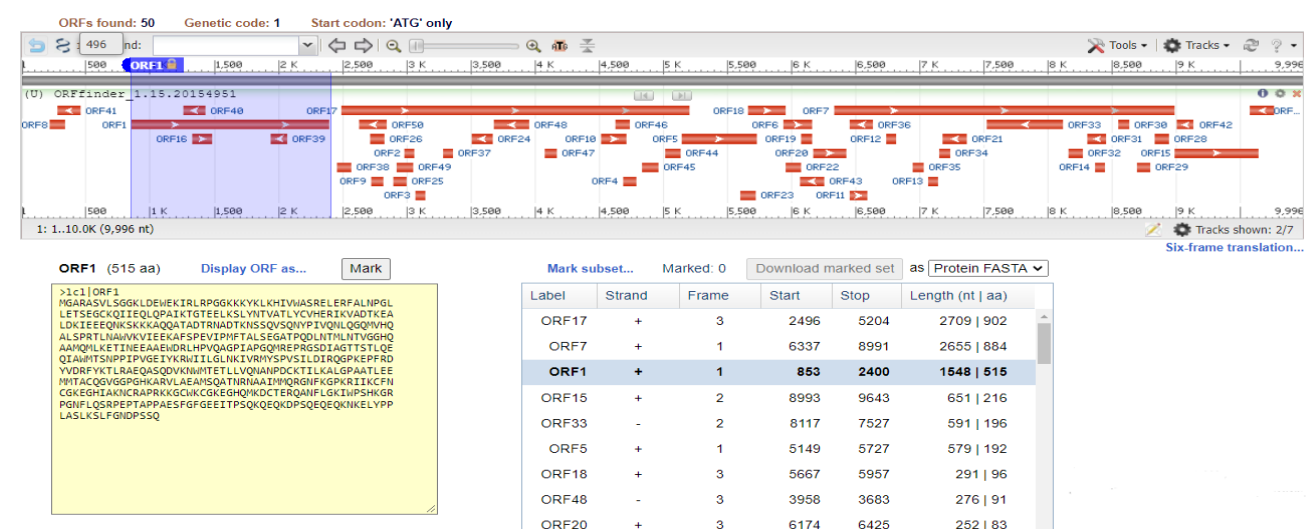


Figure 4. ORF1 in the consensus sequence of Africa

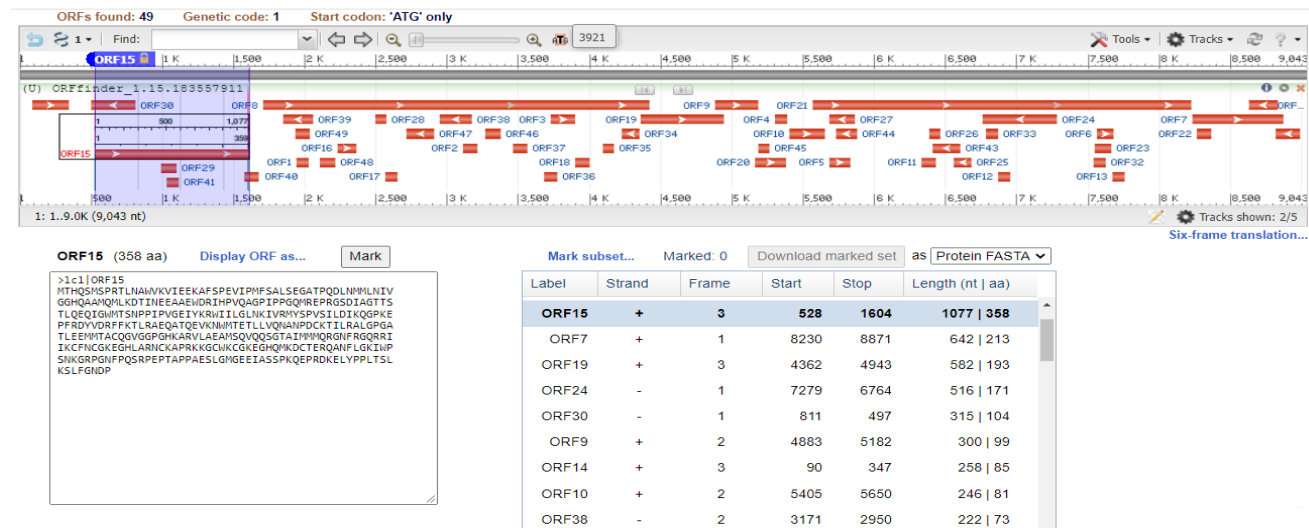


Figure 5. ORF15 in consensus sequence 2 of Europe which is the corresponding of ORF1 in the 1st sequence

By analyzing the protein sequences by Pfam website, the results show that the second sequence from Europe lacks Gag_p17 and Gag_p24 N-terminal domain as shown in figure 7. These sequences contain 129 dissimilar regions which may lead to variation in the functional products.

These Gag proteins contain all the viral elements required for virus assembly. In particular, The HIV-1 matrix protein p17 is a structural protein critically involved in most stages of the life cycle

of the retrovirus. It participates in the early stages of virus replication as well as in RNA targeting to the plasma membrane, incorporation of the envelope into virions and particle assembly.^[5]

The presented Mutations may lead to inefficient Gag targeting to the plasma membrane, resulting in dramatically reduced virus production^[1]

The first consensus sequence (Africa)

The second consensus sequence (Europ)

| Family | Description | Family | Description |
|---------------------------|---------------------------------------|---------------------------|-----------------------------------|
| Gag_p17 | gag gene protein p17 (matrix protein) | Gag_p24_C | Gag protein p24 C-terminal domain |
| Gag_p24_C | Gag protein p24 C-terminal domain | Gag_p6 | Gag protein p6 |
| Gag_p6 | Gag protein p6 | zf-CCHC | Zinc knuckle |
| zf-CCHC | Zinc knuckle | zf-CCHC | Zinc knuckle |
| zf-CCHC | Zinc knuckle | | |
| Gag_p24 | gag protein p24 N-terminal domain | | |




Figure 6. it shows the functional proteins in the 2 consensus sequences in ORF1 for the first sequence presented in the right-handed side and ORF15 for the second sequence represented in the left handed side

Gag proteins contain all the viral elements required for virus assembly. In particular, The HIV-1 matrix protein p17 is a structural protein critically involved in most stages of the life cycle of the retrovirus. It participates in the early stages of virus replication as well as in RNA targeting to the plasma membrane, incorporation of the envelope into virions and particle assembly.^[5]

The presented Mutations may lead to inefficient Gag targeting to the plasma membrane, resulting in dramatically reduced virus production^[1]

Furthermore, three more ORFs, mentioned in the below table, were compared but although there are dissimilar regions in their sequences but this have no effect on their functional products

| | | | |
|--------------------------------------|-------|-------|-------|
| The 1st Consensus sequence of Africa | ORF17 | ORF5 | ORF7 |
| The 2nd Consensus sequence of Europe | ORF8 | ORF19 | ORF21 |
| No. of dissimilar regions | 177 | 41 | 207 |

Conclusion:

This study was performed on the same virus, HIV-1 but it shows that the environment has an effect in the virus development as by changing the continent the whole genome of the virus got different nucleotides which may represent different functional products at the end as shown in result. The

more critical this functional products to the virus, the more the notable effect on its behavior and transmit.

References:

- [1] Ghanam, Ruba H.; Samal, Alexandra B.; Fernandez, Timothy F.; Saad, Jamil S. (2012). Role of the HIV-1 Matrix Protein in Gag Intracellular Trafficking and Targeting to the Plasma Membrane for Virus Assembly. *Frontiers in Microbiology*, 3(), –. doi:10.3389/fmicb.2012.00055
- [2] Deeks, Steven G.; Overbaugh, Julie; Phillips, Andrew; Buchbinder, Susan (2015). HIV infection. *Nature Reviews Disease Primers*, (), 15035–. doi:10.1038/nrdp.2015.35
- [3] Moore, M. D., & Hu, W. S. (2009). HIV-1 RNA dimerization: It takes two to tango. *AIDS reviews*, 11(2), 91–102.)
- [4] Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.
- [5] Fiorentini, S., Marini, E., Caracciolo, S., & Caruso, A. (2006). Functions of the HIV-1 matrix protein p17. *The new microbiologica*, 29(1), 1–10.

Member contribution:

All members have equally contributed to present the work

Members:

Ahmed Nabil Elghamry

Mohamed Elsayed Elmanzalawi

Sherouk Mahmoud AbdelNaby