International Letters of Natural Sciences
ISSN: 2300-9675, Vol. 7, pp 23-29
doi:10.18052/www.scipress.com/ILNS.7.23
© 2014 SciPress Ltd., Switzerland

Online: 2013-12-17

Identifying overlapping phylogenetic and geographic roots of HIV – 1 Evolution through computational analyses

Pankaj Kumar Singh, Rahul Banik, Hirak Jyoti Chakraborty, Sasti Gopal Das, Sayak Ganguli*, Abhijit Datta

DBT Centre for Bioinformatics, Presidency University, Kolkata, India

*E-mail address: sayakbif@yahoo.com

ABSTRACT

HIV-1 or Human Immuno Deficiency Virus-1 is the main causative agent of Acquired Immuno Deficiency Syndrome (AIDS). Human host infected with HIV - 1 extensively harbours many viral variants but very little is known about the difference in pattern[17] of evolution of phylogenetic lineages of HIV-1 non recombinant, normal inter subtype recombinant and main two specific recombinant forms of HIV-1 i.e., Circulating Recombinant Forms (CRFs) and Unique Recombinant Forms (URFs). This study is mainly concerned with study of the difference in evolutionary lineages of non-recombinant and recombinant sequences of HIV-1 genome sequences and identification of geographically rich areas which has reported high degree of HIV-1 occurrence and variety. Total 1550 HIV-1 genome sequences were obtained from HIV Los Alamos Database. The sequences were aligned using MAFFT (Multiple Alignment using Fast Fourier Transform) web server tool. Alignment was carried out using 10 different set of alignment parameter values. After alignment the aligned file was used for constructing N-J phylogenetic tree using Clustal X2 tool. Phylogenetic analysis was performed keeping in mind the category to which the sequence belongs. Upon analysis it was observed that the clade containing the probable ancestor belongs remained constant in all cases of different alignment values. Non recombinant isolates, inter subtype recombinants, CRFs, URFs all followed different patterns of evolution. Non recombinant sequences were found geographically specific and subtype specific to some extent whereas, normal recombinants were subtype specific and less geographically specific. CRFs showed variation among the pattern of their evolution. At some instances the sequences occurred as sister taxa of non-recombinant or normal inter subtype recombinant sequences, while at some instances as sister taxa of other CRFs where they were geographically specific. Three CRFs existed as completely diverged sequences. URFs were four in number; two of them were Indian isolates of while other two were Japanese isolates. URFs were found to be totally geographically specific. Geography wise high rate of variation was observed in India and Japan as these two countries had sequences belonging to all of the above categories. Cameroon and South Africa have very large number isolates and a considerable amount of genetic variation among isolates but they lack URFs.

Keywords: HIV-1; CRFs; URFs; Phylogenetic; recombinant strains; genome

1. INTRODUCTION

Genetic variability[14] has been a prime feature of HIV-1 during its course of evolution. Presently three groups[13] (M, main; O, outlier; N, neither) have so far been recognized. HIV-

1 group M viruses are responsible for more than 99 % of viral infection worldwide and are further classified into nine (A-D, FH, J and K) subtypes.

The HIV-1 has many recombinant forms which may be further classified under two categorised under two categories-

I. Special Recombinant Forms
II. Normal Inter Subtype Recombinant Forms

Special recombinant forms include Circulating Recombinant Forms or CRFs and Unique Recombinant Forms or URFs. These two forms are more complicated than other non recombinant forms if we consider their genetic variability. On the other hand the normal recombinant Forms of HIV-1 are less complex when compared with URFs and CRFs.

In this work it is a more of an attempt to deal with high rate of genetic variability in the HIV genome which a major setback in countering or halting the viral epidemic .Major forces that influence the viral evolution are – High mutation rate, genomic recombination and therapy and immune mediated pressure[17]. By the means of Phylogenetic analysis is possible to trace the source of mutation and which will definitely aid the efforts of drug and vaccine designing processes. This work is mainly concerned with obtaining a consensus sequences of provided sequences[11] and a tree is been generated to observe a particular model of evolution of HIV-1 sequences[7,12] which may provide further us an idea about Particular patterns of gene transfer and role recombination and mutation in which whether they were able to overcome influence of time and geographical factors to survive[10] and evolve. It may also help us to understand the difference in the nature of evolution[9] of normal strains and drug – resistance strains. This work is mainly concerned with obtaining a consensus sequences of provided sequences and a tree is been generated to observe a particular model of evolution of HIV-1 sequences[4] which may provide further us an idea about particular patterns of gene transfer[8] and role recombination and mutation in which whether they were able to overcome influence of time and geographical factors to survive and evolve.

2. MATERIALS AND METHODS

Total 1550 HIV-1 genome sequences were obtained from HIV Los Alamos Database. The sequences were aligned[6] using MAFFT (Multiple Alignment using Fast Fourier Transform) web server tool. Alignment was carried out using 10 different set of alignment parameter values. After alignment the aligned file was used for constructing N-J phylogenetic tree using Clustal X2[3] tool. Phylogenetic analysis was carried keeping in mind the category to which the sequence belongs. The main idea behind using ten different set of values for the process of alignment was obtaining aligned files from the possible range of gap open and extension values of MAFFT program.

MAFFT[7] is a multiple sequence alignment program used for UNIX-like operating system. It offers a range of multiple alignment methods depending upon the number of sequences to be aligned. The alignment was done here using the 10 different set of alignment parameters of MAFFT.

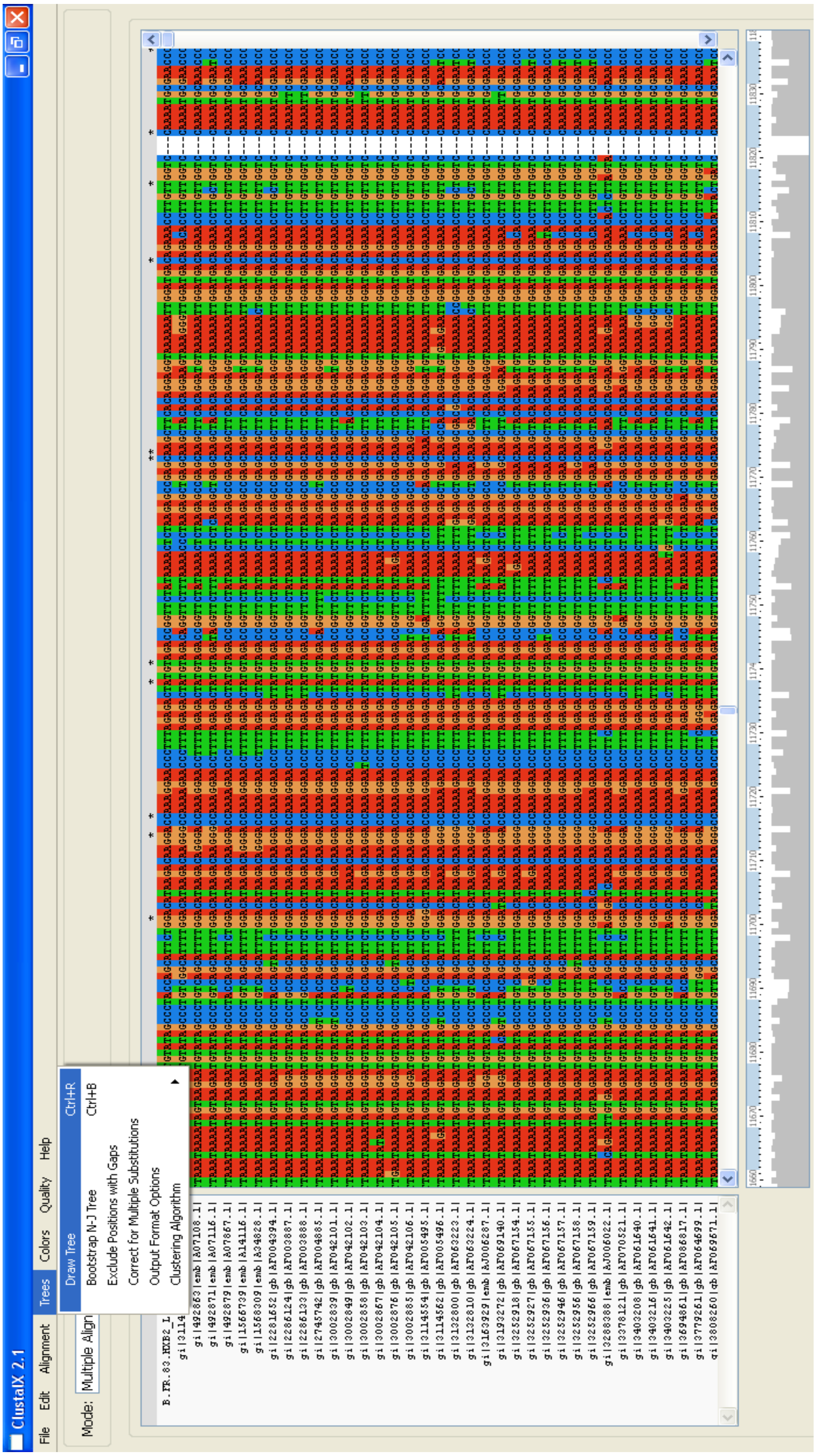

Figure 1. Tree generation using NJ method by Clustal X2.

Upon alignment the aligned sequence file was downloaded and then a Phylogenetic tree was constructed using Neighbor-Joining Algorithm using default values in Clustal X2. Tree was obtained in phylip (.ph) format. This tree was visualized in both cladogram and phylogram form using Sea View 4.0.0[16]. For cross checking this tree was converted into Newick (.nwk) format and visualized using njplot[20]. Tree generated in both .ph and.nwk format were same.

The scope of molecular Phylogenetic studies[1] for inferring short- and long-term evolutionary histories[9] of organisms and multigene families has expanded greatly beyond molecular systematic[2] due to an explosive growth in the number of sequences available in genetic databases. Phylogenetic tree[19] construction takes place after proper alignment of the concerned sequences. In this work the Phylogenetic tree[15] was constructed upon aligning the sequences and obtaining the resultant aligned (.aln) file; using CLUSTAL X2[3] software using NJ method With this growth, data sets for molecular phylogenetics have increased in terms of the number of sequences being analyzed, and the neighbor joining (NJ) method has become one of the most commonly used methods. It is computationally efficient, has desirable statistical properties, and is known to produce trees as accurate as, or better than, more computationally intensive and global searching methods NJ METHOD[18] was proposed by NAROUYA SAITOU and MASATOSHI NEI in the year 1987. N-J method is a distance based method for construction of Phylogenetic tree .It is very similar to UPGMA method of tree construction. But its main advantage is that it uses an evolutionary rate correction step before tree building. NJ method produces a unique final tree under the principle of minimum evolution. This method also does not produce minimal evolution tree, but computer stimulations has shown that it is quite efficient in obtaining the correct tree topology. It is applicable to any type of evolutionary data mainly large set of evolutionary data.

3. RESULTS AND DISCUSSIONS

The ancestor sequence among the isolated sequence is a Cameroonian isolate which is closely related to Simian Immuno Deficiency Virus of Gorillas (SIVgor). Another important feature of this sequence is that this sequence is a non recombinant one

Throughout the tree constructed number of CRFs is much higher than the number of URFs. The various patterns of evolution of the CRFs can be observed throughout the tree. In some places CRFs can be found to exist as sister taxa of non recombinant sequences, in some places these can be found in the form of sister taxa as other related CRF and apart from these, third way of occurrence of these sequences is their existence as the highly divergent sequences to their respective clades. The other important aspect is India is one of a few countries from which both recombinant form of the HIV-1 have been isolated. The URFs have formed a cluster together. Towards the end frequency of occurrence of recombinant sequences had increased significantly. An important feature is existence of highly related sequences showing 100% occurrence that is these sequences exist as extended sister groups.

The branch length of these sequence which exist as sister group is same. Another example of same observation is of sequences having Accession. No.:

[gi|238635780|gb|FJ496167.1|gi|238635790|gb|FJ496168.1|,gi|238635800|gb|FJ496169.1|,gi|238635865|gb|FJ496177.1|,gi|238635901|gb|FJ496181.1|,gi|238635910|gb|FJ496182.1|,gi|238635815|gb|FJ496171.1|,gi|238635855|gb|FJ496176.1|,gi|238635806|gb|FJ496170.1|, gi|238635891|gb|FJ496180.1|,gi|238635882|gb|FJ496179.1|gi|238635874|gb|FJ496178.1|,gi|2

38635845|gb|FJ496175.1|,gi|238635824|gb|FJ496172.1|,gi|238635928|gb|FJ496184.1|,gi|238635839|gb|FJ496174.|]

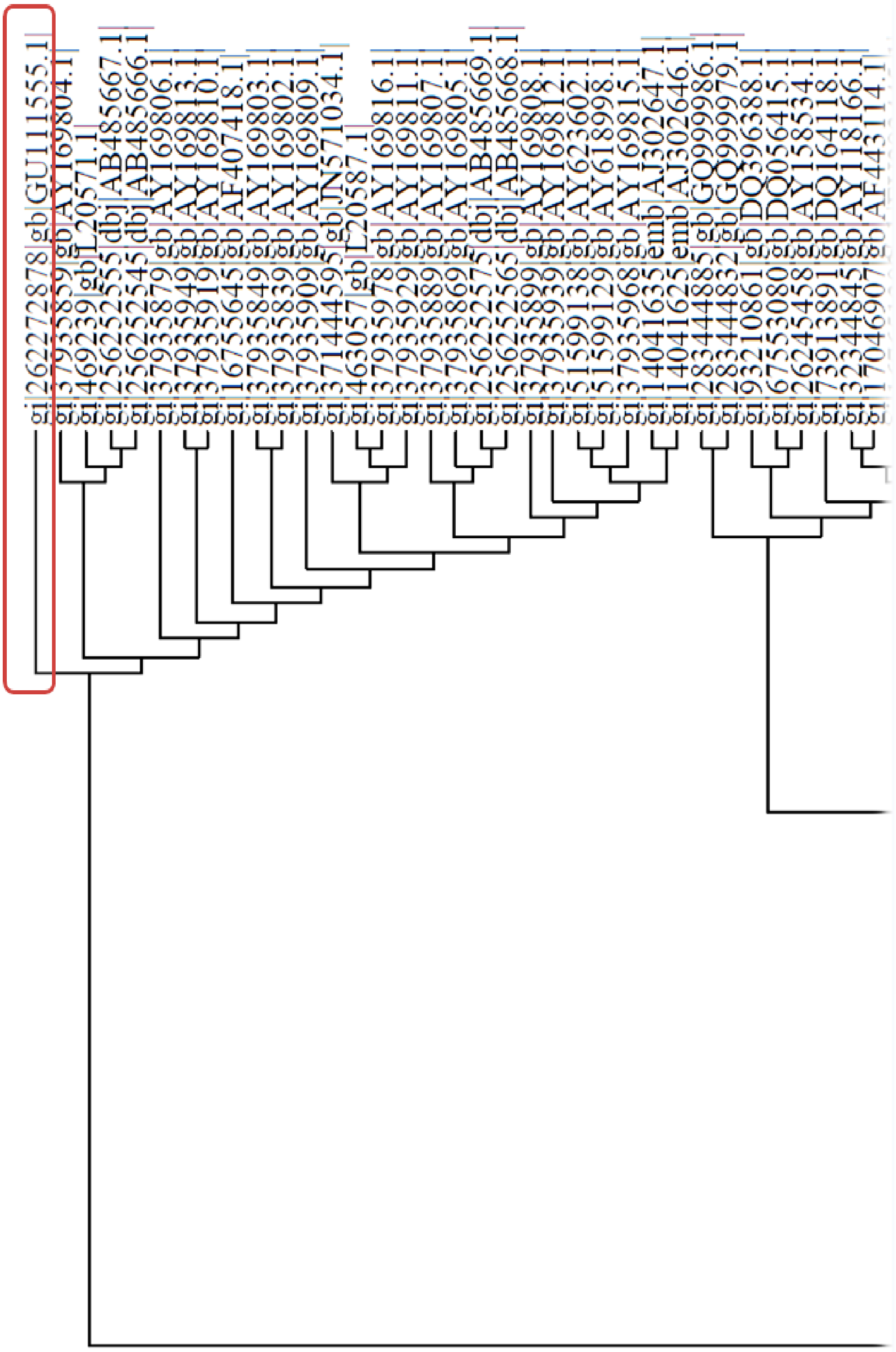
gi|262272878|gb|GU111555.1
gi|37935859|gb|AY169804.1
gi|469239|gb|L20571.1
gi|256252555|dbj|AB485667.1
gi|256252545|dbj|AB485666.1
gi|37935879|gb|AY169806.1
gi|37935949|gb|AY169813.1
gi|37935919|gb|AY169810.1
gi|16755645|gb|AF407418.1
gi|37935849|gb|AY169803.1
gi|37935839|gb|AY169802.1
gi|37935909|gb|AY169809.1
gi|371444595|gb|JN571034.1
gi|463057|gb|L20587.1
gi|37935978|gb|AY169816.1
gi|37935929|gb|AY169811.1
gi|37935889|gb|AY169807.1
gi|37935869|gb|AY169805.1
gi|256252575|dbj|AB485669.1
gi|256252565|dbj|AB485668.1
gi|37935899|gb|AY169808.1
gi|37935939|gb|AY169812.1
gi|51599138|gb|AY623602.1
gi|51599129|gb|AY618998.1
gi|37935968|gb|AY169815.1
gi|14041635|emb|AJ302647.1
gi|14041625|emb|AJ302646.1
gi|283444885|gb|GQ999986.1
gi|283444832|gb|GQ999979.1
gi|93210861|gb|DQ396388.1
gi|67553080|gb|DQ056415.1
gi|62454538|gb|AY158534.1
gi|73913891|gb|DQ164118.1
gi|32344845|gb|AY118166.1
gi|17046907|gb|AF443114.1

Figure 2. Snapshot of phylogenetic tree showing clade to which the ancestor sequence belongs to. This is a view in form cladogram.

These sequences have USA as their geographical location and were isolated in 2000. All of these sequences were submitted in PUBMED article: 19487424 this extended group has branch length of 0.041. As all of these sequences has same branch length we can consider them as quite similar from the evolutionary point of view. Even we can observe that certain sequences appearing as extended sister groups due having high similarity among their sequences. For example; we can observe an extended sister group of sequences bearing Acc.

No.: [FJ496203, FJ496207, FJ496199, FJ496203, FJ496204, FJ496207, FJ496197, FJ496195].

We can observe from Database that these sequences are isolated from ZAMBIA, around 2000-2004 A.D., and were submitted together in PUBMED article - 19487424.

On other hand the URF strains were geography specific as Indian URF strains have appeared as sister taxa of each other and Japanese URFs have shown similar way of evolution to each other and exist as sister taxa.

The normal inter subtype recombinants were found to be subtype specific and less geography specific. That means during evolution subtype have played more significant role in the evolution of these strains. Normal non recombinant sequences were more subtype and time specific instead of over harboring dependence on subtypes. As far as the distribution of HIV-1 is considered the HIV-1 is distributed throughout Africa in versatile manner. If we consider the number of sequences present in each continent this will be the order

I. Africa
II. Asia
III. Europe
IV. North America
V. South America
VI. Australia

4. CONCLUSIONS

Upon analysis Sequence having accession no: GU111555.1is considered as the probable ancestor sequence among all of these isolates. This sequence was isolated from a Cameroonian woman and this isolates is non recombinant one and is closely related to Gorilla’s Simian Immuno deficiency Virus (SIVgor) .i.e., HIV-1 has originated from primates and is from Africa. The results of this analysis supports the hypothesis of Africa been origin of HIV1 and this sequence is closely related to primates.

Throughout the analysis, most of the sequences having same Geographical location and time of isolation in a range of 4-5 years can be found to exist together in clusters without much variation. This occurrence is quit expected as Time and geographical location are two most important influencing factors having high impact on the evolution of any organism. Most of these sequences are non recombinant strains.

Sequences having high similarity can be found to be existing in form of extended sister clade. The main reason behind this occurrence is high similarity among sequences due to which they have identical branch lengths.

It can be concluded that main factor involving the evolution of HIV-1 lineages is their geographical location to which it belongs to this may be due to the variation n immunological property of individuals belonging to distant places.

ACKNOWLEDGEMENT

The authors will like to thank DBT – GOI; BTBI program for funding towards the maintenance of the facility and HIV database curators of Los Alamos HIV server for answering queries related to the use of their tool HIVAlign.

References

[1] Huson D. H., Bryant D., *Mol Biol Evol.* 23(2) (2006) 254-67.

[2] T. L. Goldberg, *Preventive Veterinary Medicine* 61 (2003) 59-70.

[3] Ling Su, et al, *Journal of Virology* (2000) 11367-11376.

[4] Larkin M. A., et al, *Bioinformatics* 23 (2007) 2947-2948.

[5] Kristen Chalmet, et al, *BMC Infectious Diseases* 10 (2010) 262.

[6] Gaschen B., Kuiken C., Korber B., Foley B., *Bioinformatics* 17(5) (2001) 415-8.

[7] Katoh Standley, *Molecular Biology and Evolution* 30 (2013) 772-780.

[8] http://www.hiv.lanl.gov/content/sequence/HIV/REVIEWS/HXB2.html

[9] Sonia Resik, et al, *Aids Research And Human Retroviruses* 23(3) (2007) 347-356.

[10] H. R. Naderi, et al, *Infectious Agents and Cancer* 1 (2006) 4.

[11] Sabri Saeed Sanabani, et al, *Virology Journal* 2009, doi:10.1186/1743-422X-6-78

[12] English, et al. *Retrovirology* 8 (2011) 54.

[13] Dong-Hun Lee1, Yeup Yoon, Chan-Hee Lee1, *The Journal of Microbiology* (2003) 232-238.

[14] N. Pierre Roque, et al, *AIDS* 18 (2004) 1371-1381.

[15] Liam J. Revell, Luke J. Harmon, David C. Collar, *Journal: Systematic Biology – SYST BIOL* 57(4) (2008) 591.

[16] Gouy M., Guindon S., Gascuel O., *Molecular Biology and Evolution* 27(2) (2010) 221-224.

[17] Philippe Lemey, et al, *PLoS Computational Biology* 3(2) (2007) e29.

[18] Saitou N., Nei M., *Mol Biol Evol*. 4(4) (1987) 406-25.

[19] W. M. Fitch, *Phil. Trans. R. Soc. Land. B* 349 (1995) 93-102.

[20] Perrière G., Gouy M., *Biochimie* 78 (1996) 364-369.

(Received 02 December 2013; accepted 07 December 2013)