IJCSC International Journal of Computer Science & Communication Vol. 1, No. 1, January-June 2010, pp. 215-219

A Tajima-NEI Method for Detecting HIV

Raninder Kaur[1], Vijay Shakti[2] & Shavinder Kaur[3]

[1]Lecturer (IT) GNDEC, Ludhiana Punjab
[2]M.Tech. CSE GNDEC, Ludhiana Punjab
[3]Lecturer (MCA) KIMT, Ludhiana Punjab
[2]rgill@gndec.ac.in

ABSTRACT

Due to large explosions of data, Knowledge discovery from these large and complex databases becomes very problematic. Various advanced techniques exist which can scale to the size of the problem and can be customized to the application of biotechnology. For the development of tree construction method, the phylogeny needs to be more explored. The phylogeny helps to understand the evolutionary relationship between given patterns of genes. The present work involves the development of Phylogenetics tree construction method with the calculations of mutation rates in order to estimate the HIV origin. The method developed for the estimation of origin of virus is based on nucleotide sequences. The criterion of computation is based on the information from the Genebank. The final results are displayed in graphical form for clearance.

Keywords: HIV, Phylogenetics, MATLAB, UPGMA, TAJIMA NEI, GAG,POLY and ENV Proteins.

1. INTRODUCTION

1.1 Role of Phylogeny in Bioinformatics

Bioinformatics is the science of managing and interpreting information from biological sequences and structures. The phylogeny is that branch of Bioinformatics which helps to understand the evolutionary relationship between given patterns of genes The Phylogenetic Trees are constructed on the basis of which we backtrack the various patterns of genes under investigation and finally reach at the ancestor node. Understanding the causes and consequences of genetic variation in human immunodeficiency virus (HIV) is one of the most important tasks facing medical and evolutionary biologists alike.

1.2 Phylogenetic Trees

A Phylogenetics tree, also called an evolutionary tree, is a tree showing the evolutionary interrelationships among various species or other entities that are believed to have a common ancestor. In a Phylogenetics tree, each node with descendants represents the most recent common ancestor of the descendants, with edge lengths sometimes corresponding to time estimates. Phylogenetic relationship between organisms is given by the degree and kind of evolutionary distance.

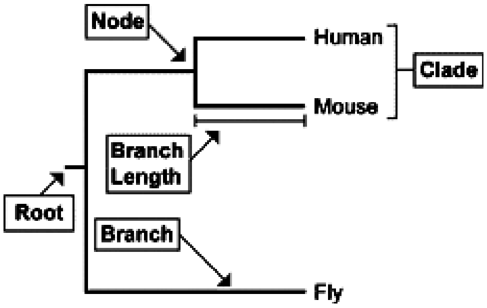


Fig 1.1: The Nodes and Branches in Phylogenetic Tree

Phylogenetics can use both molecular and morphological data in order to classify organisms.

1.3 Tree Construction Methods

There are two methods for tree costruction namely distance based and character based methods.

1. The distance-based (“phenetic”) approach is proceeded by measuring a set of distances between species, and generate the tree by a hierarchical clustering procedure.

   The Distance-based Phylogeny includes

   - Average Linkage (UPGMA) algorithm
   - Neighbour-Joining algorithm

2. The character-based (“cladistic”) approach is that which considers possible pathways of evolution,

infer the features of the ancestor at each node, and choose an optimal tree according to some model of evolutionary change (maximum parsimony, maximum likeli-hood, based on genealogy or homology).

The Character-based Phylogeny includes

- Parsimony
- Weighted parsimony (Sankoff) algorithm
- Traditional parsimony (Fitch) algorithm

1.4 UPGMA (The Method used to Build Consensus Tree)

UPGMA employs a sequential clustering algorithm, in which local topological relationships are identifeid in order of similarity, and the phylogenetic tree is build in a stepwise manner.

Steps for building UPGMA trees from a distance matrix.

1. Inspect the original matrix and find the smallest distance. Half that number is the branch length to each of those two taxa from the first Node.
2. Create a new, reduced matrix with entries for the new Node and the "unpicked" taxa. The distance from each of the unpicked will be the average of the distance from the unpicked to each of the two taxa in Node 1.
3. Inspect the reduced matrix and find the smallest distance. If these are two "unpicked taxa, then they will form a new Node with branch lengths half that distance.
4. Continue these distance calculation and matrix reduction steps until all taxa have been picked.
5. Usual convention for UPGMA (which have equal length branches from all nodes) is to use the boxy, horizontal and vertical line phylogram used in the Excel spreadsheet. This allows easy read of "time" along the X-axis.

1.5 Tajima-NEI Method

In real data, nucleotide frequencies often deviate substantially from 0.25. In this case the Tajima-Nei distance (Tajima and Nei 1984) gives a better estimate of the number of nucleotide substitutions than the Jukes-Cantor distance. This assumes an equality of substitution rates among sites and between transitional and transversional substitutions. Tajima & Nei method takes into account the base composition of the compared sequences, and is in fact the best methods as they provide accurate estimates of the number of substitutions for any C+G% of the ancestral and descendant sequences.

In the general correction of Tajima and Nei (1984), the evolutionary distance is estimated by:

$$d_{AB} = -b\ln\left(1-\frac{1}{b}f_{AB}\right)$$

where

$$b = 1-\sum_{i eN} f_i^2$$

and f_i is the frequency of the i-th type of nucleotide belonging to the set of possible nucleotide types N (= A, G, C, U or T) in the sequences being compared. This equation holds for the model of nucleotide substitutions with equal substitution rates between different nucleotides and does not take into account unequal rates of substitution among different nucleotide pairs.

2. PROBLEM FORMULATION

Despite the wide range of efforts in HIV origin prediction, the output of experimentally determined methods, typically by time-consuming and relatively expensive methods, is lagging far behind the output of gene sequence for origin prediction of the virus. A number of factors exist that make it a very difficult task to reach at the root cause of HIV. The two main problems are that the similarity of gene sequences structures is extremely large, and that the physical basis of nucleotide sequence stability is not fully understood. Mutations accumulate in the genomes of pathogens, in this case the human/simian immunodeficiency virus, during the spread of an infection. This information can be used to study the history of transmission events, and also as evidence for the origins of the different viral strains.

3. METHODOLOGY

There are two characterized strains of human AIDS viruses: type 1 (HIV-1) and type 2 (HIV-2). Both strains represent cross-species infections. The primate reservoir of HIV-2 has been clearly identified as the sooty mangabey (Cercocebus atys). The origin of HIV-1 is believed to be the common chimpanzee (Pan troglodytes).

To achieve this target following steps is followed:

1. *Retrieve Sequence Information from GenBank:* In this work, the variations in three longest coding regions from seventeen different isolated strains of the Human and Simian immunodeficiency virus are used to construct a phylogenetic tree. The sequences for these virus strains can be retrieved from GenBank using their accession numbers. The three coding regions of interest, the

gag protein, the pol polyprotein and the envelope polyprotein precursor, can then be extracted from the sequences using the CDS information in the GenBank records.

2. *Phylogenetic Tree Reconstruction:* The seqpdist and seqlinkage commands are used to construct a phylogenetic tree for the GAG coding region using the ‘Tajima-Nei’ method to measure the distance between the sequences and the unweighted pair group method using arithmetic averages, or ‘UPGMA’ method, for the hierarchical clustering. The 'Tajima-Nei' method is only defined for nucleotides, therefore nucleotide sequences are used rather than the translated amino acid sequences.

Next construct a phylogenetic tree for the POL polyproteins using the ‘Jukes-Cantor’ method to measure distance between sequences and the weighted pair group method using arithmetic averages, or ‘WPGMA’ method, for the hierarchical clustering. The ‘Jukes-Cantor’ method is defined for amino-acids sequences, which, being significantly shorter than the corresponding nucleotide sequences, means that the calculation of the pairwise distances will be significantly faster.

```
gagd = seqpdist (gag, ‘method’, ‘Tajima-Nei’,
        ‘Alphabet’, ‘NT’, ‘indel’, ‘pair’);

gagtree = seqlinkage (gagd, ’UPGMA’, data (:,1))

plot (gagtree, ’type’, ‘angular’);

title (‘Immunodeficiency virus (GAG protein)’)
```

Construct a phylogenetic tree for the ENV polyproteins using the normalized pairwise alignment scores as distances between sequences and the ‘UPGMA’, method for hierarchical clustering.

3. *Build a Consensus Tree:* The three trees are similar but there are some interesting differences. For example in the POL tree, the ‘SIVmnd5440 Mandrillus sphinx’ sequence is placed close to the HIV-1 strains, but in the ENV tree it is shown as being very distant to the HIV-1 sequences. Given that the three trees show slightly different results, a consensus tree is built using a weighted average of the three trees.

4. *Origins of the HIV Virus:* The phylogenetic tree resulting from our analysis illustrates the presence of two clusters and some other isolated strains. The most compact cluster includes all the HIV2 samples; at the top branch of this cluster we observe the sooty mangabey which has been identified as the origin of this lentivirus in humans. The cluster containing the HIV1 strain, however is not as compact as the HIV2 cluster. From the tree it appears that the Chimpanzee is the source of HIV1, however, the origin of the cross-species transmission to humans is still a matter of debate amongst HIV researchers.

SNAPSHOTS

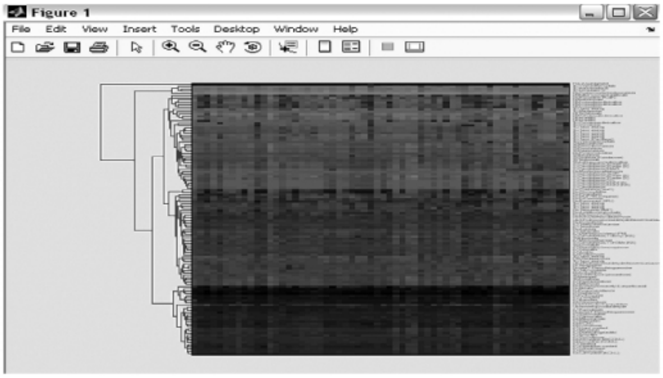


Fig 3.1: Phylogenetic Tree for the POL Polyproteins

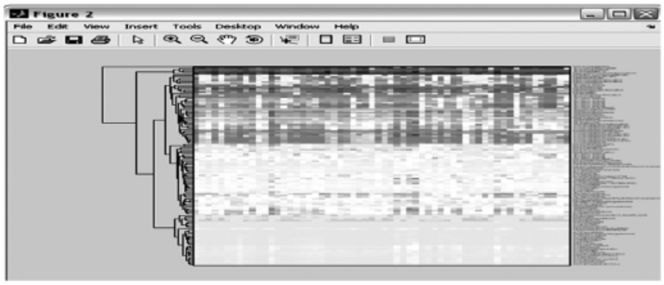


Fig 3.2: Phylogenetic Tree for the Gag Protein.

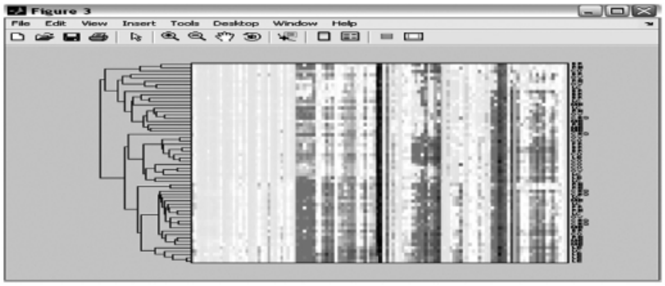


Fig 3.3: Phylogenetic Tree for the Env Protein.

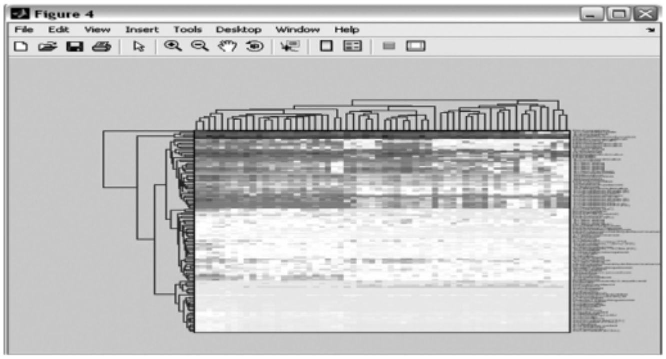


Fig 3.4: Consensus Phylogenetic Tree.

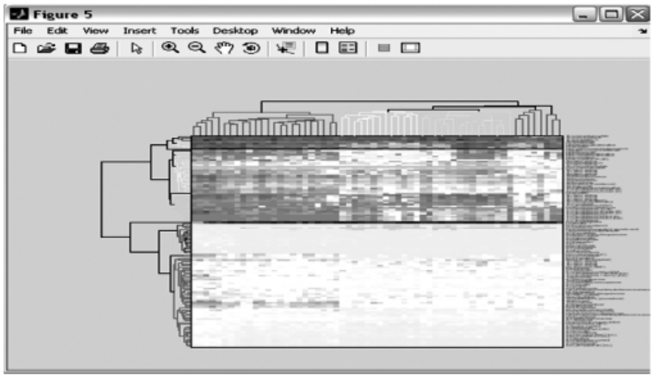


Fig 3.5: Tajima Nei Based Final Tree

3.1 Estimation of the Mutation Rates

Figure 3.4 shows the consensus tree from all three proteins, but it does not give any realistic clue about the exact location where we can suspect for the virus, so we add another approach i.e. The Estimation of Mutations in this Weighted Consensus Tree.

Mutation rates of the viruses can be defined as the chance of occurrence of mutation in each cycle of replication. In the figure 3.5, we observe fifteen various possible mutations. This clearly implies that there have been occurred fifteen different transitions. This implies High Mutation Rate is directly proportional to the transitions which in turns are the authentic cause of the HIV virus. Mutations help to increase the possibility of detection of right location for viruses.

3.2 The Plot of Interference for Three Coding Regions

The three coding regions pol, gag, env proteins are put to a window analysis and it was found that ENV shows much probability at the center where maximum interference was obtained. The plot in green line at the center of the figure 3.6 shows the peak rate of interference obtained for ENV protein. Here, the upward plots show the frequency that is increasing from bottom to top. The right hand side plot from left to right are the time span say for one year. This diagram is obtained from the plot of interference from the nodes of consensus tree .

In case of ENV Proteins, all the gene patterns are substituted to a window analysis for just to check the frecuency of occurrence of changes in all species. The one which is prone to highest frequency will adapt the very fast changes and due to transitions , more will be prone to virus.

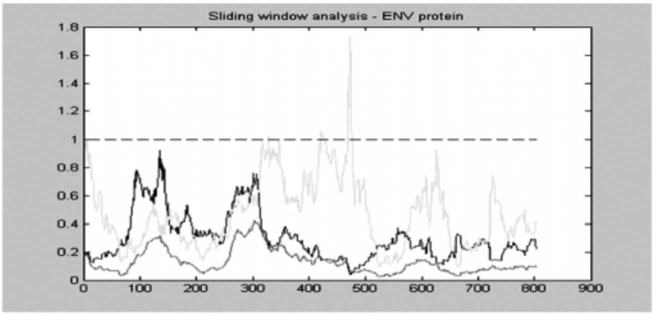


Fig 3.6: Interference in ENV Proteins Shown by the System

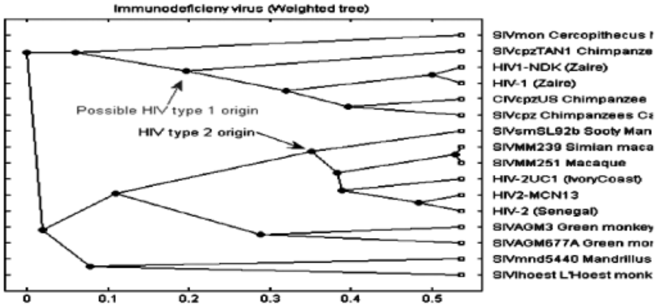
The red lines show species with negligible changes adapted. The Blue ones are slightly prone to few changes The green lines show the extent to which the transitions have occurred.

Similarly, for the second case, the GAG protein is analysed.

4. RESULTS AND DISCUSSIONS

The mutation rates and interferences are applied in existing Tajima-Nei method and is compared on various synthetic and real-world nucleotide sequences.

The phylogenetic tree resulting from mutation analysis and three proteins regions illustrates the presence of two clusters of HIV virus along with some other isolated strains and it is found that the calculation of Mutation rates and interference for POL, GAG and ENV Proteins along with an implementation of UPGMA in Tajima -Nei leads to a considerable improvement on the detection of root node of HIV Virus. In comparison with the older schemes, the new approach has been found to be more robust to give a reliable estimation to reach at the root cause for the problem. To investigate the epidemic history of HIV, we estimate the rate of mutation based on an equal sample of gag, pol, and env sequences. This approach provides a good fit to the detection of origin, as evaluated by other schemes like likelihood ratio testing as shown in fig 4.1..



Both the types of HIV are highlighted at 0.3 and 0.4, 0.1 and 0.2 time intervals respectively.

As, the unique node obtained is in between 0.1 and 0.3and no more node is closely related t6o this. So is considered as the suspected origin for the HIV origin.

5. CONCLUSION

Viruses are associated with man or animals from the time immemorial, and man is always trying to get rid off and to extinguish them. Due to their immense ability, viruses conquer the life time and again. Viruses evolve against all the physical, chemical or environmental barriers. The present model computes the mutation rates for given set of nucleotide sequences and provides the result in the form of Phylogenetics trees that can be compared for different sequences under considerations. The results are also shown graphically which further provides a detailed and complete description of the diversity during the analysis of the sequences.

REFERENCES

[1] Tajima F, Nei M. Estimation of evolutionary distance between nucleotide sequences. Mol Biol Evol. 1984; 1:269-285.

[2] Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987; 4:406-425.

[3] F. Gao, et al., "Origin of HIV-1 in the Chimpanzee Pan Troglodytes Troglodytes", Nature 397(6718), 1999, pp. 436-41.

[4] H.W. Kestler, et al., "Comparison of Simian Immunodeficiency Virus Isolates", Nature 331(6157), 1998, pp. 619-622.

[5] M. Alizon, et al., "Genetic Variability of the AIDS Virus: Nucleotide Sequence Analysis of two Isolates from African Patients", Cell 46(1), 1986, pp. 63-74.

[6] Mathematical Properties of Some Measures of Evolutionary Distance Journal of Theoretical Biology, **245**, Issue 4, 21 April 2007, Pages 790-792 Yun-Huei Tzeng, Wen-Hsiung Li and Trees-Juen Chuang.

[7] Comparison of the evolutionary distances among syngens and sibling species of Paramecium Molecular Phylogenetics and Evolution, **38,** Issue 3, March 2006, Pages 697-704 Manabu Hori, Izumi Tomikawa, Ewa Przybo? and Masahiro Fujishima.

[8] Efficient and Robust Global Amino Acid Sequence Alignment with Uncertain Evolutionary Distance Modern Information Processing, 2006, Pages 371-381 Matthias C.M. Troffaes.

[9] The Influence of Selection on the Evolutionary Distance Estimated from the Base Changes Observed between Homologous Nucleotide Sequences Journal of Theoretical Biology, **213,** Issue 2, 21 November 2001, Pages 129-144 Jinya Otsuka, Yosuke Kawai and Nobuyoshi Sugaya.

[10] Higher Selection Pressure from Antiretroviral Drugs in Vivo Results in Increased Evolutionary Distance in HIV-1 pol,: **259,** Number 1 (1999), pages 154-165.