ISSN: 2230-9926

Available online at http://www.journalijdr.com

IJDR

International Journal of Development Research
Vol. 09, Issue, 11, pp. 31054-31057, November, 2019

RESEARCH ARTICLE OPEN ACCESS

SED: AN EDITOR FASTA USING LABELS OF PHYLOGENY AND ASSOCIATED TRACES

***Marta Barreiros and Élcio Leal**

Institute of Biological Sciences, Federal University of Pará (UFPA), Belém-PA, ZIP/Zone 66075-110 – 479, Brazil

ARTICLE INFO

Article History:
Received 08th August, 2019
Received in revised form
11th September, 2019
Accepted 16th October, 2019
Published online 20th November, 2019

Key Words:

Alignment editor;
Phylogeny; Files; HIV-1.

Corresponding author:
Marta de Oliveira Barreiros

ABSTRACT

Phylogenies are used to treatvarious biological functions, such as the relationships between tree topology and traits of observed species. Currently, tools are available that allow viewing and editing on phylogenetic trees, as well as creating and editing alignments. However, there is a need for a simplified tool that automatically selects sequences from an alignment based on phylogenetic information. Here we created a tool called SED (data editing software), which uses phylogeny-based files to perform statistical summaries and generate subsets of alignments. SED automatically generates an analysis of previously marked data in phylogeny clades, defined by the user from sequencesets. Its input is a tree file (Nexus format), an alignment (Fasta format) and numeric data in a spreadsheet (CSV format) for statistics. Using labels marked a priori on the tree file, SED files will be generated (Fasta format) containing selected labels, it also compares the traits (CSV file) of the selected sequences with those of the entire alignment. Efficiently, SED was tested by analyzing 261 samples from the vif gene of HIV-1 subtype B with TG mutation detection (tryptophan) in eight samples (8/261 ~ 3.06). SED is a tool used to generate subset sequences from a Fasta file and generate analyzes with tagged taxa.

Citation: Marta de Oliveira Barreiros and Élcio Souza Leal. 2019. "Sed: an editor fasta using labels of phylogeny and associated traces", *International Journal of Development Research*, 09, (11), 31054-31057.

INTRODUCTION

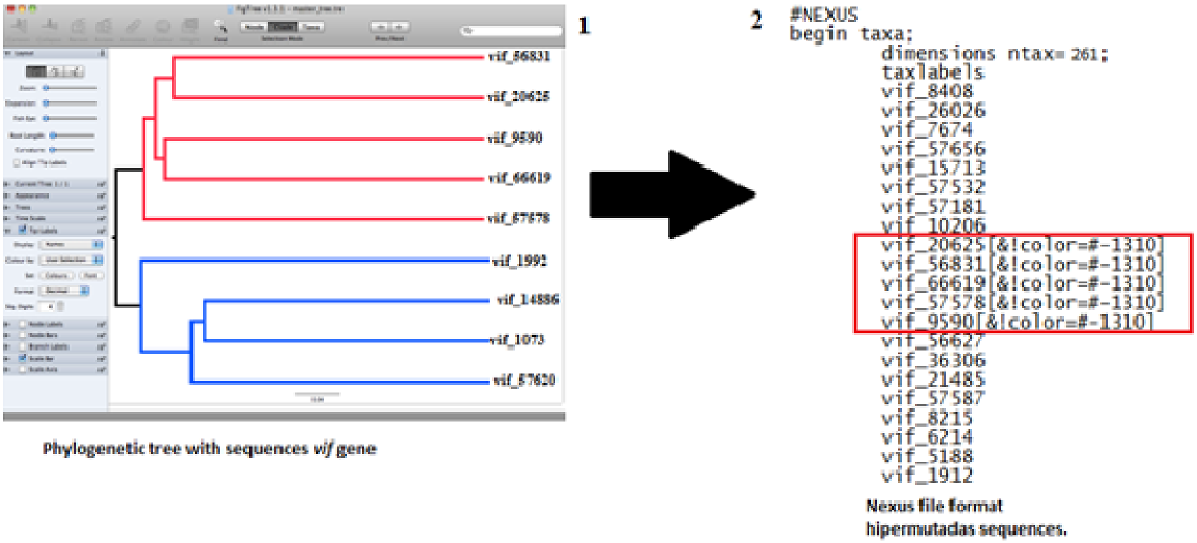
Phylogenetic reconstruction methods are used to estimate the evolutionary relationships of taxa and ancestry of organisms (Yang and Rannala, 2012). Usually, during the phylogenetic analysis some issues are necessary. For example, it is sometimes necessary to add or remove alignment rate based on the initial tree analysis. Another common practice is the correlation analysis of characteristics associated with taxa sharing a common ancestor. There are currently programs to view and edit phylogenetic trees, such as MEGA software, which is an integrated tool for performing sequence alignment, which allows inferring phylogenetic trees, estimating times of divergence, estimating rates of molecular evolution, inferring ancestral sequences, and testing evolutionary hypotheses (http://www.megasoftware.net/), PAUP software is a software package most used for inference of evolutionary trees (paup.csit.fsu.edu/), PhyML software is used to estimate phylogeny based on the principle of maximum likelihood (http://www.atgcmontpellier.fr/phyml/binaries.php)and Figtree software is a graphic viewer of phylogenetic trees, and allows the user to make edits in the Phylogeny (*i.e.*, coloring clades, viewing trees in different models), and exporting in different file formats (http://tree.bio.ed.ac.uk/software/figtree/). However, there is still a need for which automatically selects sequences from an alignment based on phylogenetic information. Here we create a tool with friendly graphical interface called SED (data editing software), written in Java that reads alignments and builds sets of sequences that were previously marked in a tree file. Our program also performs summary statistics on selected sequences if the associated characteristic information is available.

METHODOLOGY

Implementation and use: SED uses modules in Java to do the analyzes requested by the user, input commands and routines and data output were implemented using the java.io package (https://docs.oracle.com/javase/6/docs/api/java/io/package-summary.html). The package org.apache.commons.math (http://commons.apache.org/proper/commons-math/) was used to implement statistical analyzes. In addition, the tool can provide richer results by employing graphic tables, in which it was implemented using the package org.jfree.chart (http://www.jfree.org/jfreechart/). A free online CASE tool was used for graphical representation of system functionality (https://creately.com). The SED tool is available for download from the http://sed2.webnode.com/ project page as a compressed file and can be simply run by unpacking the file without any special installation procedures. For use, must install the Java platform (www.java.com) on the machine.

Sed input: The default entry for SED is an editable document generated from the taxon markings from a phylogenetic tree in Nexus format (.nxs or .trees) containing a list of genetic alignment headers. This type of file can be generated using the Figtree program (http://tree.bio.ed.ac.uk/software/figtree/), it is necessary to have a label made previously in the predefined sequences for analysis (Fig. 1). The list of input sequences, by definition, contains headers that are also related to the other files used in the same analysis, such as the Fasta (genetic alignment) file and the CSV file (comma-separated values), created by a spreadsheet, containing data for statistical analysis and identification of each sequence. It is important that the sequence identification is in the first column of the spreadsheet and the other data from the second column, thus, can save the file to CSV.

For example, the genetic sequences of a certain subtype of HIV-1 (*i.e.*, subtype B) of a data set may be used as input, and subsequently correlated with clinical data from patients (i.e., viral load and CD4 + T cells) with other subtypes (i.e., subtypes A and C) unmarked in the same data set. Fig. 1 - Illustration of the steps of creating an extended Nexus file generated from the phylogenetic tree using the FigTree software. The arrow indicates the extension of the Nexus formatted file, and the clade marked red indicates the hypermutated sequences of the *vif*gene analyzed during the search. Steps: 1) Make a phylogenetic tree from an alignment, then 2) the exported Nexus file coding sequence ahead of the labeled sequences (Underlined with red color). To validate the developed tool, a dataset of 261 HIV-1 subtype B *vif* gene samples from infected individuals was tested, previously generated in another study (Bizinoto*et al.*, 2013). These sequences were initially aligned by ClustalW software (Larkin, 2007), and a formatted Fasta file was saved. Then, we used the Hyphy software (Pond, Frost& Muse, 2005) to infer a phylogenetic tree by maximum likelihood using the neighbor-joining method.



```
#NEXUS
begin taxa;
        dimensions ntax= 261;
        taxlabels
        vif_8408
        vif_26026
        vif_7674
        vif_57656
        vif_15713
        vif_57532
        vif_57181
        vif_10206
        vif_20625[&!color=#-1310]
        vif_56831[&!color=#-1310]
        vif_66619[&!color=#-1310]
        vif_57578[&!color=#-1310]
        vif_9590[&!color=#-1310]
        vif_56627
        vif_36306
        vif_21485
        vif_57587
        vif_8215
        vif_6214
        vif_5188
        vif_1912
```

Nexus file format
hipermutadas sequences.

Fig. 1. **Illustration of the steps of creating an extended Nexus file generated from the phylogenetic tree using the FigTree software. The arrow indicates the extension of the Nexus formatted file, and the clade marked red indicates the hypermutated sequences of the *vif* gene analyzed during the search. Steps: 1) Make a phylogenetic tree from an alignment, then 2) the exported Nexus file coding sequence ahead of the labeled sequences (Underlined with red color)**

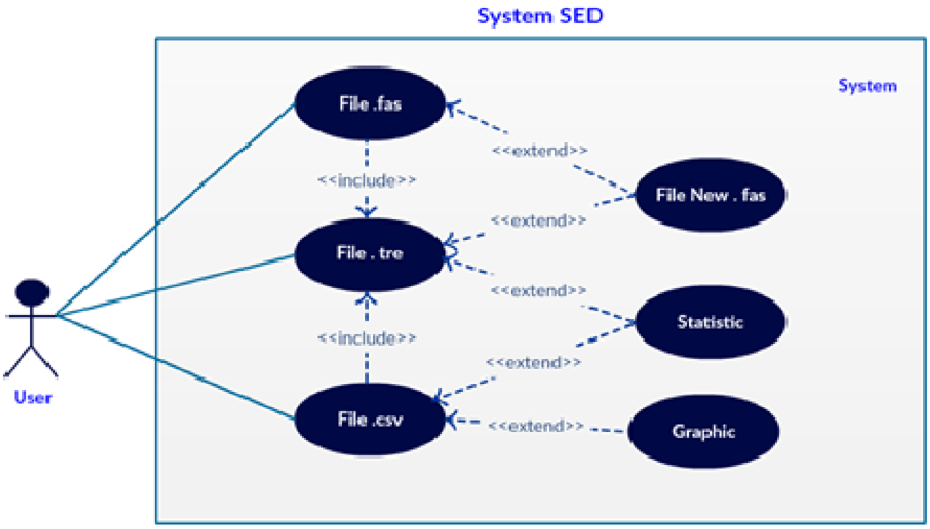
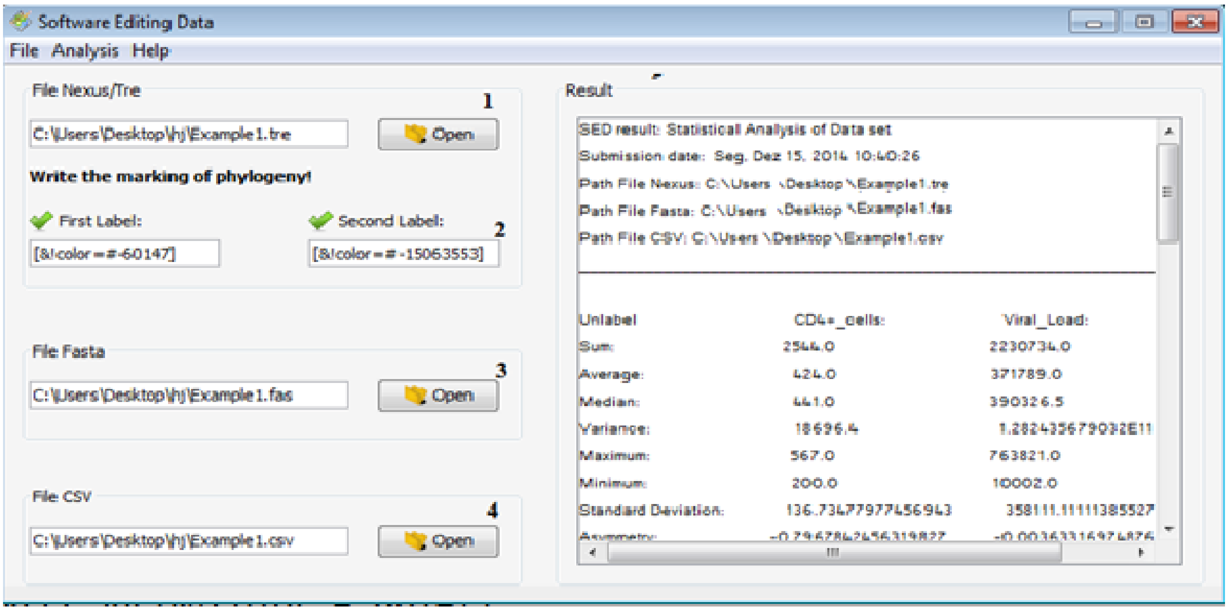


Fig. 2. UML representation of the SED functionality. To initialize the system, the user needs to enter the three types of files (Fasta, Nexus and CSV), the file in Nexus format controls the analysis, since it has the taxa marked. Then, can extract a new Fasta file or a statistical analysis



C:\Users\Desktop\hj\Example1.tre

[&!color=#-60147]

[&!color=#-15063553]

C:\Users\Desktop\hj\Example1.fas

C:\Users\Desktop\hj\Example1.csv

```
SED result: Statistical Analysis of Data set
Submission date:  Seg, Dez 15, 2014 10:40:26
Path File Nexus: C:\Users  \Desktop \Example1.tre
Path File Fasta: C:\Users  \Desktop \Example1.fas
Path File CSV: C:\Users \Desktop \Example1.csv
______________________________________________________________

Unlabel                  CD4+_cells:              Viral_Load:
Sum:                     2544.0                   2230734.0
Average:                 424.0                    371789.0
Median:                  441.0                    390326.5
Variance:                18696.4                  1.282435679032E11
Maximum:                 567.0                    763821.0
Minimum:                 200.0                    10002.0
Standard Deviation:      136.73477977456943       358111.11111385527
Asymmetry:               -0.7967842456319827      -0.00363316974876
```

Fig. 3. Illustration of the SED graphical interface. 1) Loading the file in Nexus format, 2) marking code insertion field created from phylogenetic tree inference, 3) loading the file in Fasta format, 4) loading the file in CSV format and the Result field is the output of a statistical analysis

The tree has been viewed and edited with Figtree software. By labeling (color-coding) tags in the taxa sets a Nexus file was generated. Fig. 2 shows the UML (Unified Modeling Language) representation of the main steps required to perform a SED analysis. Initially a Fasta file containing all the sequences needs to be opened, next to the Nexus file (tagged tree file) also needs to be loaded. Fig. 2 - UML representation of the SED functionality. To initialize the system, the user needs to enter the three types of files (Fasta, Nexus and CSV), the file in Nexus format controls the analysis, since it has the taxa marked. Then, can extract a new Fasta file or a statistical analysis.

SED output

The final resultof SED is a new file generated in theformat Fast a with genetic sequences that were previously marked in the edition of the phylogeny, can be chosen two options to save the data: previously marked sequences or unlabeled sequences. It isimportantto note that SED is not intended to view or edittrees, it onlyreadstree files, andthen uses labels to select sequences from an alignment file. Optionally, if some kindofnumerical data associatedwithsequences (here denoted continuous trace) isavailable, then it canbeusedtoperform a statistical analysis. In additiontodescriptivestatistics, it is also possible to perform a t-test to verify that the sequence trace continuum values (labeled sequences) differ from the original sequences in the Fasta file. Thereisalsoanoptionalgraphic output, including the parameter with title sdescribed bytheuser (Histogram, Box-Plot, Histogram Plotand Pizza). The main SED window, withall files loaded, isrepresented in Fig. 3. Fig. 3 – Illustrationofthe SED graphical interface. 1) Loadingthe file in Nexusformat, 2) marking code insertion field created from phylogenetic tree inference, 3) loadingthe file in Fastaformat, 4) loadingthe file in CSV format and the Result field is the output of a statistical analysis.

RESULT AND DISCUSSION

The methodology used for a set of samples of the HIV-1 *vif* gene to analyze the functionality of SED was described in (Bizinoto*et al*., 2013). To illustrate the functioning of SED, we used virus loads (expressed in copies of RNA / ml plasma) and the CD4 + T cell counts (cells/mm3) levels associated with each sample. Results from data set analyzes of the HIV-1 subtype B *vif* gene samples in infected individuals, showed hypermutation of the type AG in codons TGG (tryptophan). These hypermutations are induced by the human protein APOBEC3G (present in CD4 + T lymphocytes) and when they occur in the tryptophan codon induce the appearance of termination codons (TGA) in the viral genome. It has recently been shown that the substitution of tryptophan for termination codons in the HIV-1 genome is sufficient to prevent viral replication (Consol, 2014). We observed this type of substitution in eight samples (8/261 ~ 3.06%), mean CD4 + T lymphocytes in this group were 685.87 (standard deviation ± 390.53) cells / mm3 and median 735 cells / mm3, whereas in the samples without hypermutation the mean CD4 + T lymphocytes were 370.97 (standard deviation ± 287.37) cells / mm3 and median of ± 334 cells / mm3.

The mean viral load in hypermutation patients was 5,604.37 copies / mL (standard deviation ± 9,127,065) with a median of 2,900 copies / mL, but in the non-hypermutation patients the mean was 191,809.29 copies / mL (mean deviation Standard ± 368,304.85) and the median ± 80 copies / mL. The analysis of the association between hypermutations and clinical data was not statistically significant for CD4 + T lymphocyte levels (p = 0.064) nor for viral load (p = 6.02). SED is a tool that allows users to generate new files (Fasta format) and perform an analysis using any type of cloning in clades of a phylogenetic analysis and simultaneously generate graphs and statistical analysis with the numerical data.

Conclusion

SED is a computational tool used locally to generate sub sets of sequences from a Fasta file taking into account the edition of the phylogeny as signed by the Figtree program. The program includes the preparation of statistical calculations and graphs through the interaction between the files in Nexus and CSV formatsto help in better understanding the analysis addressed.

REFERENCES

Bizinoto, M.C., Yabe, S., Leal, E., Kishino, H., Martins Lde O, *et al*., 2013. Codon pairs of the HIV-1 vif gene correlate with CD4+ T cell count. *BMC Infect Dis* 13: 173.

Consol, P. *et al*., 2014. HIV infection *en route*to endogenization: two cases. *Clínical Microbiology and Infection*, v. 20, n. 12, p. 1280-1288.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, *et al*., 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23: 2947-2948.

Pond SL, Frost SD, Muse SV., 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics 21*: 676-679.

Yang Z, Rannala B., 2012. Molecular phylogenetics: principles and practice. *Nat Rev Genet* 13: 303-314.

****