COMPARATIVE ANALYSIS OF BIOINFORMATICS TOOLS USED IN HIV-1 STUDIES

Daniel Kareski
Navayo technologies
Skopje, Macedonia

Nevena Ackovska
Institute of Intelligent Systems, Faculty of Computer Science and Engineering
Skopje, Macedonia

ABSTRACT

The HIV-1 virus is one of most studied viruses because of the incredible ability to morph its genetic structure. Sequencing the virus and generating phylogenetic trees gives us the ability to recognise the evolutionary changes of the virus and possibly help in finding a cure. Bioinformatics offers great choice of tools that are used in the process of generating phylogenetic trees. Different factors contribute in the usage of one tool over another and in order to compare phylogenetic quality of a tree we use the bootstrapping method. In this paper HIV-1 bioinformatics tools used in studies conducted in the Balkan Peninsula are presented, and comparative analysis of those tools is elaborated. The tools utilised in these studies are used in processing random data. The results of their usage is also presented.

KEYWORDS:
HIV-1, phylogenetic trees, sequence, alignment, MEGA, pol gene.

I. INTRODUCTION

The HIV-1 is a retrovirus that causes *acquired immunodeficiency syndrome* (AIDS) [1], a condition that weakens the immune system and makes it more vulnerable to all kinds of attacks. This virus has a fast replication cycle [2] which makes it able to modify its structure because of the errors made while self-replicating. The HIV-1 genome [3] is quite compact and efficiently stored due evolutions way of always using the solution that works the best, so it contains coding regions that overlap. Among the coding regions which enable the virus to replicate and attack organisms, the GAG-POL coding region is essential for the reproduction function of the virus. The *pol* coding region (Fig. 1) encodes three enzymes that are most important for the replication of HIV-1: protease, reverse transcriptase and integrase. It has been shown that the *pol* gene contains sufficient genetic information to perform experiments and give conclusions on all kinds of details regarding the virus like the subtype of the virus, its closeness to others subtypes of HIV and many others. Furthermore, the *pol* gene is generally the most conserved gene throughout the virus's evolution [4].

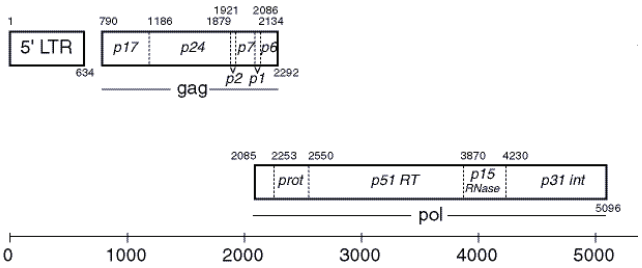

Figure 1: Part of the HIV-1 genome showing the coding regions for *gag* and *pol genes*. [3]

The efficiency of the HIV-1 virus and inability to cure the disease makes it one of the most studied viruses. Worldwide, studies have shown a tendency in which the HIV-1 virus present in a certain area develops similar types of mutations. These characteristic landmarks are used when we try to subtype the sampled HIV-1 viruses. Further studies are conducted to discover why some regions contain less infected organisms than others. Due to the massive subtyping of HIV-1, the virus's particular subtypes are divided in 4 major groups (M, N, O, P) and the most common subtypes, which are part of the group M, are divided into more groups (A, B, C, D, E, F, G, H, I, J, K).

When conducting an experiment on the HIV-1, after retrieval of the blood samples containing the virus, it is required for it be sequenced so its gene structure would be known. This is done by DNA sequence analysers. The sequenced *pol* gene retrieved in the form of a string consisting of the 4 base pairs A, C, T, G is first aligned [5] to other known sequences of the same gene. Alignment is performed by maximising the similarity of two or more strings consisted of the 4 base pairs. While comparing the strings, the only allowed operations are adding a gap, deleting a base or using the current base. The sequences used in alignment are incorporated in the alignment tools available on many web sites and databases (one example is [6]) and when the alignment is done we use the aligned string to continue with the analysis.

This study focuses on three things: performing alignment, determining the subtype of a virus and building phylogenetic trees. The work presented here is divided in five sections. In the following section some of the studies conducted on the Balkan Peninsula are represented. Further, the tools used in the studies are elaborated. In the fourth section the specific tools used in this research are explained. At the end, some future work and the conclusions are laid out.

II. Studies conducted on the Balkan Peninsula

There are several studies conducted on the Balkan Peninsula regarding HIV-1 virus. Generally the aim of the studies is to show the subtypes currently present on this territory, different statistics regarding the spreading of the virus, searching for common characteristics of different subtypes, and the way HIV entered some particular region. The studies conducted in Albania [7] and Bulgaria [8] aim to generate some insights in to the subtypes of HIV-1 virus present in both of the countries respectively using the coalescent theory. In both of the studies there were sufficient number of *pol* gene samples which are firstly sequenced in order to determine the HIV-1 subtypes, then aligned using different alignment tools and databases, and the result of the alignment is used in generating different phylogenetic trees.

In the Albanian study the results showed that most common subtypes are HIV-1A and HIV-1B. Furthermore, it was shown that there is a major resemblance with HIV-1A viruses present in Greece which implies that single major introduction of HIV-1A has occurred. The results of the study confirmed the initial hypothesis given by the coalescent theory and it was determined that the most resent common ancestor (MRCA) for HIV-1A was dated in 1980s with 757 active infections, while MRCA for HIV-1B was dated in mid-1970s with 243 active infections. The phylogenetic tree also showed that the infection spread from the city of Tirana to the outer parts of Albania.

In the Bulgarian study the result showed that 50% of the *pol* gene sequences were classified as subtype B, 22% as subtype A1, and the rest of the percentage was divided between subtypes: C, F, H, G. The phylogenetic analysis showed that subtype B viruses were intermixed with many foreign sequences from both East and West Europe, while subtype A1 viruses were closely related to sequences from West Europe (France and Spain in particular).

III. Tools used in the conducted studies

All the tools used in the studies can be divided in three categories: tools for subtyping genes (*pol* genes in our case), tools for alignment and tools for generating phylogenetic tree. Every tool has a huge impact on the eventual outcome especially considering the fact that we almost never work with 100% accuracy in bioinformatics. In order to test the quality of the built trees we use the bootstrap method. This method is quite handy because it shows how well the new phylogeny tree can handle the data that was used to create it. Next, we utilize the same combination of tools used in the studies mentioned above on some random HIV-1 *pol* genes in order to determine the bootstrap level of the generated trees over the same data sample.

A. Popular tools used in studies all over the world

The BioAfrica online tools [9] and Los Alamos HIV databases [10] are the main sources of tools used in HIV-1 studies worldwide. They generally contain all the latest bioinformatics techniques which allow the researches an easy and mathematically proven method of generating all kinds of statistics, generating phylogenetic trees, subtyping using HIV-1 genes, calculating drug resistance etc. It would shorten the time to make assumptions and prove or decline those assumptions. It also allows a unified and error-safe method of transferring data from one laboratory to other. Generally all the tools are based on some fundamentally same algorithms, so the only reason why one would choose one tool over another is just familiarity with the tools interface and way of functioning.

B. Albanian study

In the Albanian study the *pol* gene subtypes were determined using the RIP tools [11] and sequences were aligned using clustalw tool available at [12] followed by manual editing. For the creation of the phylogenetic tree PAUP* 4.0 [13] was used.

C. Bulgarian study

In the Bulgarian study Rega subtyping tool [14] was used in order to perform subtyping. Alignment was also performed using the clustalw tools and manual editing using BioEdit software [15]. The Bayesian phylogenetic tree was created with MrBayes tool [16].

IV. Specifics of tools in our research

One of the main reasons why this study is conducted is because there is a hypothesis that the number of HIV-1 infected people in Macedonia is drastically low compared to other Balkan countries. There might be many reasons for this, such as: people are generally not tested for infection, the knowledge of the people about HIV is low, or something else. However, the patients that are confirmed as infected, are under HAART [17] treatment. The treatment drastically reduces the efficiency of the replication mechanism governed by the HIV-1 virus.

This study is statistics-oriented so we could have a better view on the numeric details regarding HIV-1 virus, the time of its origin on this territory, subtypes, dynamics of spreading, etc. Just like in the other studies, we also work with the *pol* gene because it is a reliable way to determine the details that are focus of this research. The tools presented here are chosen over other bioinformatics tools mostly because of good reviews, simple user interface and reliable output.

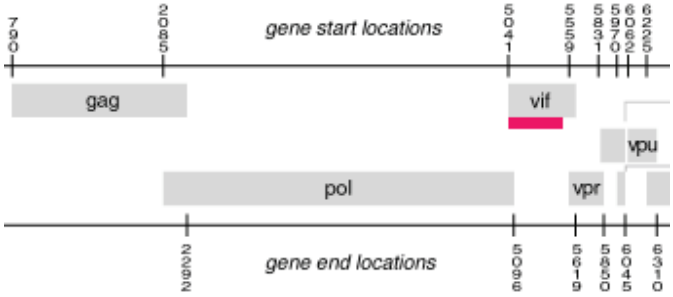

Figure 2: Result of the alignment tool [6] showing the input sequence (marked red) compared to part of the HIV-1 genome

We have chosen a random data set of few *pol* genes sequences from European countries, which are aligned with an alignment tool available at [6], in order to spot the difference in the phylogenetic trees built by different tools. This data is to be compared to the samples that we have obtained from the samples in Macedonia. The alignment tool [6] gives an extremely detailed output regarding the position of the current sequence compared to the overall HIV-1 genome, which HIV-1 gene it is aligned to and a graphical representation the result. Also, a detailed text output is shown representing every alignment between the current input and all the similar sequences available in the database. Fig. 2 shows an example of aligning process of a subsequence of the *vif* gene (in red colour) and its position compared to the HIV-1 genome.

Furthermore, four software packages are presented, used in generating one of the most common HIV-1 analysis, inferring phylogenetic trees. The most common phylogenetic trees used in bioinformatics are: maximum likelihood tree (ML), maximum parsimony and neighbor-joining tree (NJ).

The maximum likelihood and maximum parsimony methods are character-based methods because they use the individual substitutions among the sequences to determine the most likely ancestral relationships. Maximum parsimony method utilises the tree generated at a particular step. We calculate the number of changes that need to be performed on a set of characters (part of the observed data) so they could fit in the tree's structure. The best hypothesis is the tree requiring the fewest changes. Different subtypes of the maximum parsimony method imply different penalties when changes occur. Maximum likelihood uses an explicit model of how the character state changes occur. The models describe the relative rates of different changes i.e. they differ in terms of the parameters used to describe the rates at which one nucleotide replaces another during evolution. In this research with the test samples Jukes-Cantor model is used, due to its simplicity (it assumes equal base frequencies). The NJ tree is part of the distance-based methods which first calculate the overall distance between all the pairs of sequences, and then calculate a tree based on those distances. The NJ method is generally much faster compared to the other two methods previously mentioned, but much more error prone when working with biased data (for example HIV data).

A. Mega software package

The Molecular Evolutionary Genetics Analysis (MEGA) software package is an integrated tool for conducting sequence alignment, generating phylogenetic trees, mining web-based databases, constructing evolutionary trees and testing evolutionary hypotheses. Fig. 3 shows the phylogenetic tree created with the Mega software package [18]. In this research 11 samples of randomly chosen *pol* gene sequences from [10] are used. The sequences were first aligned using the alignment tool part of the MEGA package. This package allows the user to choose from 5 different tree models: maximum likelihood tree (ML), minimum evolution tree, maximum parsimony, UPGMA trees and neighbor-joining tree. We choose the ML tree with Jukes-Cantor model and a bootstrap test of 100 iterations. The numbers shown in Fig. 3 represent the bootstrap index. Bootstrap index of 100 means that every recombination of nucleotides, which are part of the pol sequences of the current edges, are supported by the tree structure.

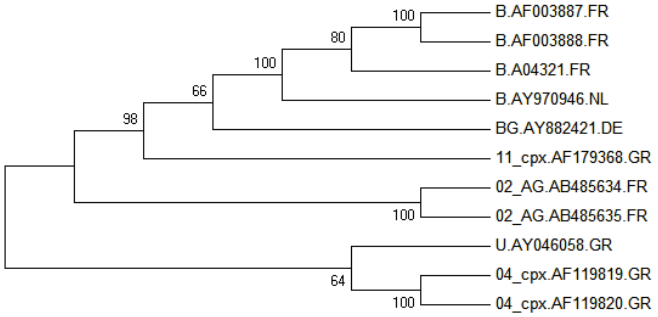

Figure 3: Phylogenetic tree generated by Mega.

B. SeaView software package

SeaView is a multiplatform program designed to facilitate multiple alignment and phylogenetic tree building from molecular sequence data through the use of a graphical user interface. SeaView reads and writes various file formats (NEXUS, MSF, CLUSTAL, FASTA, PHYLIP, MASE, Newick) of DNA and protein sequences and of phylogenetic trees. SeaView drives programs muscle [19] or Clustal Omega [20] for multiple sequence alignment, and also allows to use any external alignment algorithm able to read and write FASTA-formatted files. Phylogenetic trees are computed by parsimony, using PHYLIP's dnapars/protpars [21] algorithm, distance, with NJ or BioNJ [22] algorithms on a variety of evolutionary distances and maximum likelihood. Fig. 4 represents the phylogenetic tree created with the SeaView software package [23]. We use the same input of 11 *pol* gene samples, the Jukes-Cantor model and the ML tree. The numbers represent the bootstrap index.

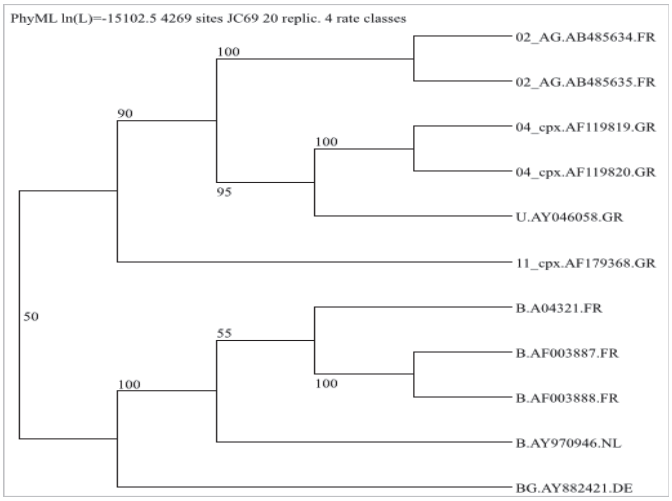

Figure 4: Phylogenetic tree generated by SeaView.

C. *MrBayes software tool*

MrBayes is a program for Bayesian inference and model choice across a wide range of phylogenetic and evolutionary models. MrBayes uses Markov chain Monte Carlo (MCMC) methods to estimate the posterior distribution of model parameters. This software uses an input in the NEXUS data format which was generated using [24]. Just like in the previous tools we use the same input sequences and ML tree with Jukes-Cantor model and the consensus tree is shown in Fig. 5. The consensus tree has the posteriors, i.e. the support values for the different clades (different *pol* gene sequences) or in other words the probability that these clades are correct. These support values, correspond to the number of trees that recovered that clade during the building process.

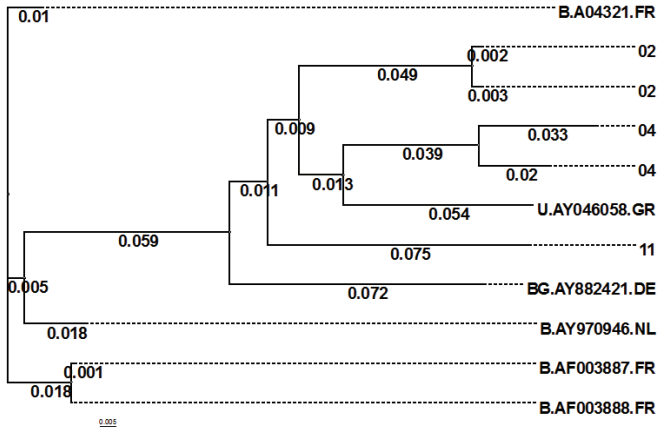

Figure 5: Phylogenetic tree generated by MrBayes.

As mentioned above, the selected software packages are among the most used packages which provide good user experience and detailed output. When it comes to comparing the three tools, the Mega software package provided the best user interface. It also provided excellent control in manipulating with the customization of the tree building procedure and was definitely faster than the SeaView tool. We should keep in mind that ML trees generate probably the best output, but are not much effective when the input size is big. In the first two cases we used bootstrap method to evaluate the quality of the trees (bootstrap index of 100 is the best case scenario) and the result produced by the Mega package has slightly higher overall index. The bootstrap method is not available for the MrBayes tool in which case we use the support values.

V. FUTURE WORK

Future work regarding this study is focused in retrieving and sequencing blood samples from population in Macedonia. This process is already in progress. After its completion, we expect to conduct the studies on the samples, based on the studies already conducted in the world. Additional research will be performed in order to determine which tools exactly will be used combined with which the methods in order to receive the most objective output. These sequences combined with sequences from other Balkan and European countries would be used in generating a phylogenetic tree. One of the question that we hope we will find answer to is what is the geographic evolution of the HIV-1 virus in our country and why are there so few infections on the territory of Macedonia. In the end there would be general comparison with the results and conclusions of the Albanian, Bulgarian and other European studies.

VI. CONCLUSION

The work presented here explains the usage of the bioinformatics software tools when dealing with HIV virus. The study gives an insight which software package to be used with the blood samples collected from the Macedonian HIV patients.

When it comes to subtyping the RIP tools are the best candidates, since they are part of a reliable set of tools. Regarding alignment there are two tools ([6] and [25]) part of the Los Alamos HIV databases [3], one of the best online software packages, which are pretty easy to use and give excellent output. Out of the 3 used tools, the MEGA software generates a more accurate output, presents better user interface, faster performance, gives better customization when generating phylogenetic tree and offers better export methods. This makes MEGA the preferred tool for building phylogenetic trees.

REFERENCES

[1] Kilmarx P. Acquired immunodeficiency syndrome. In: Heymann DL, editor. Control of communicable diseases manual, 19th Edition. Washington, D.C.: APHA Press; 2008.

[2] Toshiyuki Goto, Masuyo Nakai, Kazuyoshi Ikuta , "The life-cycle of human immunodeficiency virus type 1", Micron, Volume 29, Issues 2–3, April–June 1998, pp. 123–138

[3] http://www.hiv.lanl.gov/content/sequence/HIV/MAP/landmark.html

[4] André F. Santos, Marcelo A. Soares, HIV Genetic Diversity and Drug Resistance, Viruses. 2010 February; 2(2): 503–531.

[5] David J. Lipman, Stephen F. Altschul, John D. Kececioglu, "A tool for multiple sequence alignment", Proc. Nail. Acad. Sci. USA Vol. 86, June 1989 , pp. 4412-4415

[6] Gaschen B, Kuiken C, Korber B, Foley B. Retrieval and on-the-fly alignment of sequence fragments from the HIV database.Bioinformatics 2001.17(5):415-8

[7] Marco Salemi, Tulio de Oliveira, Massimo Ciccozzi, Giovanni Rezza, Maureen M. Goodenow, "High-Resolution Molecular Epidemiology and Evolutionary History of HIV-1 Subtypes in Albania", January 2008

[8] Marco Salemi, Maureen M. Goodenow, Stefania Montieri, Tulio de Oliveira, Maria Mercedes Santoro, Danail Beshkov, Ivailo Alexiev, Ivailo Elenkov, Ivan Elenkov, Tsvetana Yakimova, Tonka Varleva, Giovanni Rezza and Massimo Ciccozzi, "The HIV Type 1 Epidemic in Bulgaria Involves Multiple Subtypes and Is Sustained by Continuous Viral Inflow from West and East European Countries", AIDS RESEARCH AND HUMAN RETROVIRUSES, Volume 24, Number 6, pp. 771-779, 2008

[9] T. de Oliveira,K. Deforche, S. Cassol, M. Salminem, D. Paraskevis, C. Seebregts, J. Snoeck, E. J. van Rensburg ,A. M. J. Wensing, D.A. van de Vijver, C. A. Boucher , R. Camacho , and A-M Vandamme, An Automated Genotyping System for Analysis of HIV-1 and other MicrobialSequences. 2005; 21(19): 3797-3800

[10] http://www.hiv.lanl.gov/

[11] RIP tool - http://www.hiv.lanl.gov/content/sequence/RIP/RIP.html

[12] Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ and Higgins DGBioinformatics 2007 23(21): 2947-2948

[13] Swofford, D. L. 2003. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

[14] Rega subtyping tool - http://dbpartners.stanford.edu/RegaSubtyping/

[15] BioEdit - http://www.mbio.ncsu.edu/bioedit/bioedit.html

[16] MrBayes - http://mrbayes.sourceforge.net/

[17] Lucas GM, Chaisson RE, Moore RD, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. Annals of Internal Medicine, 1999, 131(2):81-87

[18] MEGA-CC: Computing Core of Molecular Evolutionary Genetics Analysis Program for Automated and Iterative Data Analysis

[19] Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput ; Nucleic Acids Res. 32(5):1792-1797

[20] http://www.clustal.org/omega/

[21] http://evolution.genetics.washington.edu/phylip/progs.algs.heur.html

[22] http://www.atgc-montpellier.fr/bionj/

[23] Gouy M., Guindon S. & Gascuel O. (2010) SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution* 27(2):221-224

[24]http://www.hiv.lanl.gov/content/sequence/FORMAT_CONVERSION/form.html

[25]http://www.hiv.lanl.gov/content/sequence/QUICK_ALIGN/QuickAlign.html