



4-20-2013

## A Polyglot Approach to Bioinformatics Data Integration: Phylogenetic Analysis of HIV-1


Steven Reisman

Catherine Putonti  
*Loyola University Chicago*

George K. Thiruvathukal  
*Loyola University Chicago, gkt@cs.luc.edu*

Konstantin Läufer  
*Loyola University Chicago, klaeuf@gmail.com*

Follow this and additional works at: [https://ecommons.luc.edu/cs\\_facpubs](https://ecommons.luc.edu/cs_facpubs)

 Part of the [Bioinformatics Commons](#), [Computational Biology Commons](#), [Genomics Commons](#), [Programming Languages and Compilers Commons](#), and the [Software Engineering Commons](#)

### Recommended Citation

S. Reisman, C. Putonti, G. K. Thiruvathukal, and K. Läufer. A Polyglot Approach to Bioinformatics Data Integration: Phylogenetic Analysis of HIV-1: Research Poster. 2nd Greater Chicago Area System Research Workshop (GCASR), May 3, 2013, Evanston, IL, USA.

This Presentation is brought to you for free and open access by the Faculty Publications and Other Works by Department at Loyola eCommons. It has been accepted for inclusion in Computer Science: Faculty Publications and Other Works by an authorized administrator of Loyola eCommons. For more information, please contact [ecommons@luc.edu](mailto:ecommons@luc.edu).



This work is licensed under a [Creative Commons Attribution-NonCommercial-No Derivative Works 3.0 License](#).



# A Polyplot Approach to Bioinformatics Data Integration: Phylogenetic Analysis of HIV-1

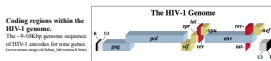
Catherine Putonti <sup>1,2,3</sup>, Steven Reisman <sup>1,2</sup>, George Thiruvathukal <sup>3</sup>, Konstantin Läufer <sup>3</sup>  
 1 Department of Biology, 2 Bioinformatics, 3 Computer Science, Loyola University Chicago, Chicago, IL

## Abstract

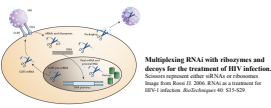
RNA-interference has potential therapeutic use against HIV-1 by targeting highly-functional mRNA sequences that contribute to the virulence of the virus. Empirical work has shown that within cell lines, all of the HIV-1 genes are affected by RNAi-induced gene silencing. While promising, inherent in this treatment is the fact that RNAi sequences must be highly specific. HIV, however, mutates rapidly, leading to the evolution of viral escape mutants. In fact, such strains are under strong selection to include mutations within the targeted region, evading the RNAi therapy and thus increasing the virus' fitness in the host. Taking a phylogenetic approach, we have examined 4000+ HIV-1 strains obtained from NCBI's database for each of the HIV genes, identifying conserved regions at each hypothetical and operational taxonomical unit within the tree. Integrating the wealth of information available from each genome's record, we are able to observe how conserved regions vary with respect to their distribution throughout the world. This was made possible through the development of a new software tool, developed such that similar analyses can be conducted for any species or gene of interest, not just HIV-1. In addition to the phylogenetic signal which we can recognize from the HIV-1 genomes examined, we can also identify how selection varies across the genome. Taking this evolutionary approach, we have detected regions ideal for targeting by RNAi treatment.

## HIV-1 Genomics & Therapy

Given its medical importance to humans, thousands of HIV-1 isolates have been completely sequenced. Over 4000 genomes are available through public database.

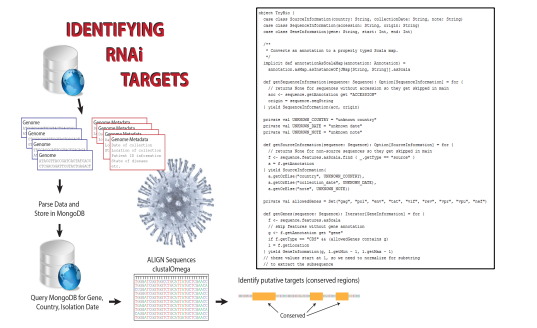


RNAi therapeutics have been proposed by targeting HIV RNAs or host transcripts required for viral replication. Numerous siRNAs have been able to achieve viral inhibition both *in vitro* and *in vivo*. Challenges present themselves as a result of the evolution of escape variants.



## Computational Approach

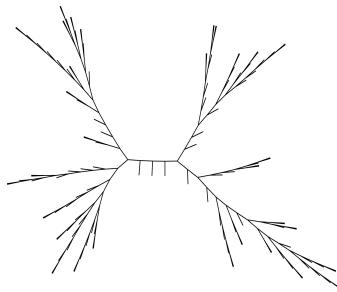
Using the programming languages Scala and Python, we have developed a tool which, although here applied to HIV-1 sequence data, can be used for the analysis of any species. Sequence metadata is also stored and made accessible to user queries via a mongoDB database. Users can select any country, range of dates, and individual coding region to analyze and identify conserved short subsequences for putative use as siRNAs.



## Background

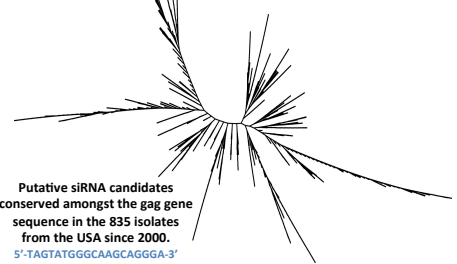
With the decreased cost and increased throughput of genome sequencing, it has now become feasible to sequence numerous genomes from different patients worldwide as well as over the progression of a disease. In addition to large nationally supported, publicly accessible data repositories, a number of databases specialized for a specific infectious organism have been developed. These data repositories include more than just sequence data; they also include metadata about the sequence. Diagnostics and treatments for infectious viruses, bacteria, and fungi are moving towards a more molecular approach. Molecular diagnostics and therapies necessitate knowledge of the genomic sequence of the infectious agent. The success of this approach hinges on the fact that the diagnostic/therapy developed must target an exact nucleotide sequence present within the infectious agent's genome. Thus, molecular diagnostics and therapies want to target the most conserved regions, or those regions acquiring the least accumulation of mutations. Targeted regions, however, are not immune to mutation. Species with genomes including a mutation or mutations within the target region will evade detection and thus thrive. Therefore, constant monitoring of genomic changes over time and from a wide population is necessary.

This tree shows the evolutionary relationship amongst 131 sequences of the gag gene from strains isolated from individuals in China since 2000.



Putative siRNA candidates, conserved amongst all of these isolates ( $\geq 15$  nt)  
 5'-TAGTATGGCCAGCAG-3'  
 5'-GGGCAATGGTACATCA-3'  
 5'-GGAGCCACCCACAGATTAAA-3'  
 5'-TTAAATAAATAGTGAAGAATGTATAGCCC-3'

This tree illustrates the evolutionary relationship amongst 835 gag gene sequences from strains isolated from individuals in the USA since 2000.



Putative siRNA candidates conserved amongst the gag gene sequence in the 835 isolates from the USA since 2000.  
 5'-TAGTATGGCCAGCAGGA-3'

Targeting only the more recent strains, the strains more likely to be in the circulation, looking at sequences isolated since 2008 (40 strains) produces three additional siRNA candidate sequences:  
 5'-AATGGATGGGTAAGA-3'  
 5'-GAACCAAGGGGAAGTGCATAGC-3'  
 5'-GTAATAAATGGTACACAGAAC-3'

Results

## Research Objectives

The aim of the proposed research will be to develop software for assisting visualizing variable rates of evolution across genomic sequences for which there are numerous sequenced strains and associated metadata (sample collection information, patient data, etc.). This effort will include the development of novel data structures (framework) to store the sequences and metadata, and algorithms to facilitate analysis of the data. We have begun this effort by focusing on analysis of HIV-1 genomes, using the data available through NCBI's HIV Sequence Database ([www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)). In addition, a database will be designed to allow high accessibility to the genomic data.

## Future Directions

At present we are still developing the querying functionality, in addition to sequence post-processing. The functionality has been developed to minimize human intervention. However, to reduce false positives and annotation issues (inherent in the underlying data), post-processing scripts have been developed. Furthermore, we are developing this framework such that it is flexible and can thus be applied to any species of interest and any data resource available. In doing so, the tool can be utilized for any number of genomic and evolutionary studies.

## Acknowledgements

This research is supported by a Research Support Grant from Loyola University Chicago (GT). Funding for SR to attend and present in October 2012 at the MEEGID conference (New Orleans, LA) was provided by Loyola Undergraduate Research Opportunities Program's Travel Grant.