

Application of W-curves and TSP to Clustering HIV1 Sequences

Douglas Cork^{1,2,3}, Steven Lembark⁴, Nelson Michael^{1,5}, Jerome Kim^{1,5}

US Military HIV Research Program¹; Henry M. Jackson Foundation for the Advancement of Military Medicine², Rockville, MD. 20850, BCPS Dept., Illinois Institute of Technology³, Chicago, IL. 60616, Workhorse Computing⁴, Woodhaven, NY., Walter Reed Army Institute of Research⁵, Rockville, MD.

Abstract - *The high mutation rate in HIV-1 makes it difficult to treat and analyze. Monitoring the evolution of drug resistance requires frequent re-sequencing, but comparing and visualizing the progress is difficult. One difficulty is simply locating the areas of interest: gaps and crossover mutations make it difficult to isolate clinically significant sequences for comparison. Effectively displaying the results of comparisons grouped according to multiple regions is also a problem. Our comparison algorithm based on the W-curve helps automate the comparison process, producing results suitable for clustering via a modified solution to the Traveling Salesman Problem (“TSP”). Appropriate color-coding of the TSP results allows us to display the results of multiple comparisons effectively for single samples or time-series. The results can be useful for providing guidance in treatment, analyzing the membership in anonymous study populations, tracking the evolution of drug resistance in populations, or rates of co-infection within study groups.*

Key words: HIV-1 Genomic Clustering, W-curve, Traveling Salesman Problem, Drug Resistance

1 Disclaimer

The opinions or assertions contained herein are the private views of the author, and are not to be construed as official, or as reflecting true views of the Dept. of the Army or the DoD.

2 Introduction

One step in managing HIV infection response is monitoring individuals and study populations for evolution of drug resistance. Given *in vivo*

replication kinetics with more than 10^9 new cells infected every day, each and every possible single-point mutation occurs between 10^4 and 10^5 times per day in an HIV infected individuals [1]. This high level of variation leaves identifying the markers for drug resistance difficult, requiring multiple manual steps in many cases. Especially for population studies, an automated process that can identify and compare drug resistance sites would be an enormous help. The W-curve's scoring process produces a set of localized comparison values that can be used as landmarks that can be used to automate the process of locating relevant areas for comparison. Clustering new samples with known ones using the Traveling Salesman Problem (“TSP”) can automate the grouping of new samples into drug-resistance categories. Combining them leads to an automated process for tracking medication in individuals or analysis of groups.

3 Background

Analyzing the data in many HIV studies is made difficult by a combination of HIV's genetics and the sources of data. Patients in many HIV clinics are anonymous, often because they are illegally engaged in prostitution, illegal drug use, or homosexual sex. Studies of groups look for new strains and how they spread and have to check for changing members of the sample population; individual patients have to be sequenced frequently to evaluate appropriate drugs.

HIV's high mutation and crossover rates combined by discontinuous sequences for the drug responses make the analysis difficult. The high mutation in areas between the regions of interest confound any analysis based on whole genes or regions:

there is simply too much white noise on the genome to use large portions of it for comparison. High rates of mutation and co-infection leave many individuals with multiple strains of HIV, further confounding the analysis, and HIV strains have frequent crossover mutations, making things even worse. As an example, studies of cross-clade neutralization produce effectively random results [2].

Treatment of HIV is still largely a manual process: patients have to be sampled and sequenced frequently and doctors have to make informed decisions on how to deal with evolution of the strains infecting them. Presenting sequence comparison results is particularly important: physicians have to evaluate the similarity of a single individual to multiple known drug-resistant strains.

In the US 98% of the infections are type-B, and the single FDA-approved program for treatment using genetics treats type-B only. The U.S. Army has to treat soldiers who acquire HIV all over the world and only 85% of their cases are type-B, leaving them without any good options for 15% of their cases. Being able to at least classify the non-B cases and evaluate their treatment outcomes effectively would be a huge help.

4 Methodology

4.1 W-curves

The W-curve was originally designed as the basis for a graphical tool for visualizing very large regions of

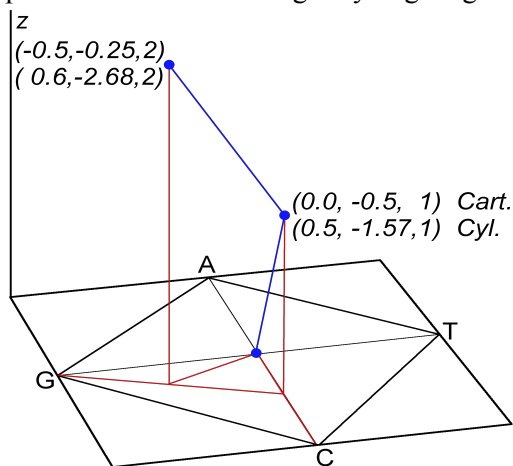


Figure 1: W-curve generation showing initial points for curve "CG" [4].

DNA [3]. Over time it has been adapted for numerical comparison as well [4,5,6]. By abstracting the genetic sequences into a three-dimensional space, W-curves offer a wider range of comparisons rather than comparisons based on searching, aligning and tree-building (assuming a mutation model) with uni-dimensional strings of characters. W-curves make it easier to find patterns in sequences or locate common features between genes. Their 3D nature also makes it possible to align smaller features than are possible with string based techniques. The design of our comparison utilities builds on these capabilities to provide a technique for matching the small regions of HIV-1's CD4 epitopes along its *gp120* gene [7].

The order of bases along the corners is significant: number of hydrogen bonds (2 or 3) and chemical structure (purine or pyrimidine) share quadrants around the square. This means that most synonymous SNP's in the gene sequences will leave the curves in the same quadrant. This keeps our same-quadrant measure small for SNP's.

We use a two-pass process for comparing W-curves. The first pass produces an array of alignment regions with starts, stops, and a difference measure that we call "chunks". SNP's increase the difference measure, gaps show up as differences in the starting values on successive chunks of the comparison, indels as successive chunks with no change in the relative offset. The second pass summarizes the chunks into whatever measures are useful, for example by averaging the differences over the length of a sequence.

The advantage of chunking the results first is that similar chunks can be used to locate landmarks for aligning sequences with one another. Areas with small differences provide an automated means of locating the offsets between start and stop values in the sequences. Given a library of sequences with known landmarks such as points of known drug resistance, we can score a single incoming sequence against all of them.

4.2 TSP and Clustering

The Traveling Salesman Problem ("TSP") is quite easy to describe but difficult to solve. The problem is to take a list of distances between cities and make a tour of them which visits each city once

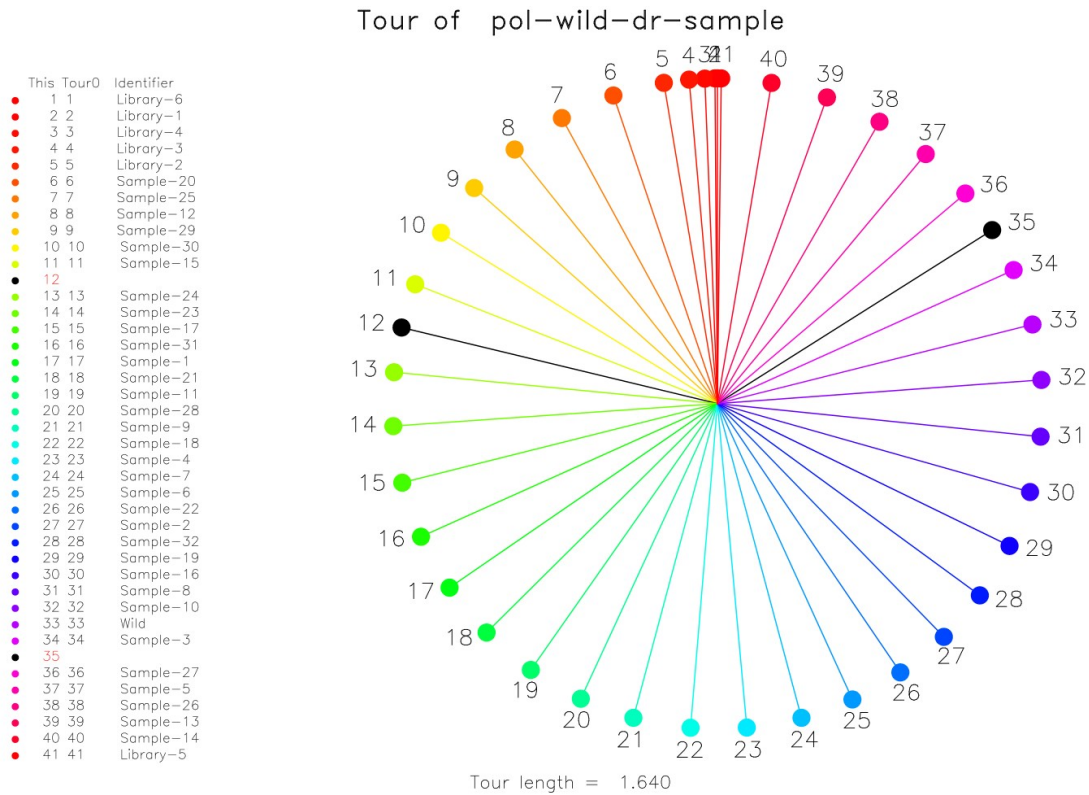


Figure 2: Example of tour generated by R's TSP package comparing POL sequences of the wild type with a library of simulated drug-resistant (“dr”) and random sample strains of HIV-1 [6]. The library strains cluster tightly together at the top (red) with the wild strain (#33) outside the drug-resistant cluster to the right. The approach shown here works well if individuals are sampled over time: pre-assigning the colors from a first tour makes it relatively easy to watch if samples from individuals drift into different groups.

with the least total distance. Substitute cost and the algorithm will find the “cheapest” route through all of the cities. As it turns out, this problem is NP-complete, requiring analysis of all routes to guarantee the least distance. Much work has gone into developing heuristics for solving this problem and there are fast algorithms for approximating the solution.

The utility of TSP problems is that an optimal tour will cluster the closest cities together. If the difference measures are for genes, they can similarly be clustered on any region of interest. A number of techniques for determining inter- and intra-clade distances have been developed. One technique developed by Climer and Zhang at Washington University is to add a fixed number of “dummy

cities” to the list [8]. Each dummy has a small distance to all other cities (we use 2^{-20}). The non-zero distance leaves these cities in the intra-cluster gaps. We display the resulting tours as color-coded pinwheel diagrams. Appropriate color-coding makes these relatively easy to analyze individually or compare to one another.

4.3 Generating and Analyzing TSP Clusters with R

The R statistical package includes a TSP library, available from CRAN. We have used it here to generate an approximate solution for clustering the genes. In our case an optimal tour is not required: any good approximation will cluster the genes properly. Our approach starts with a square

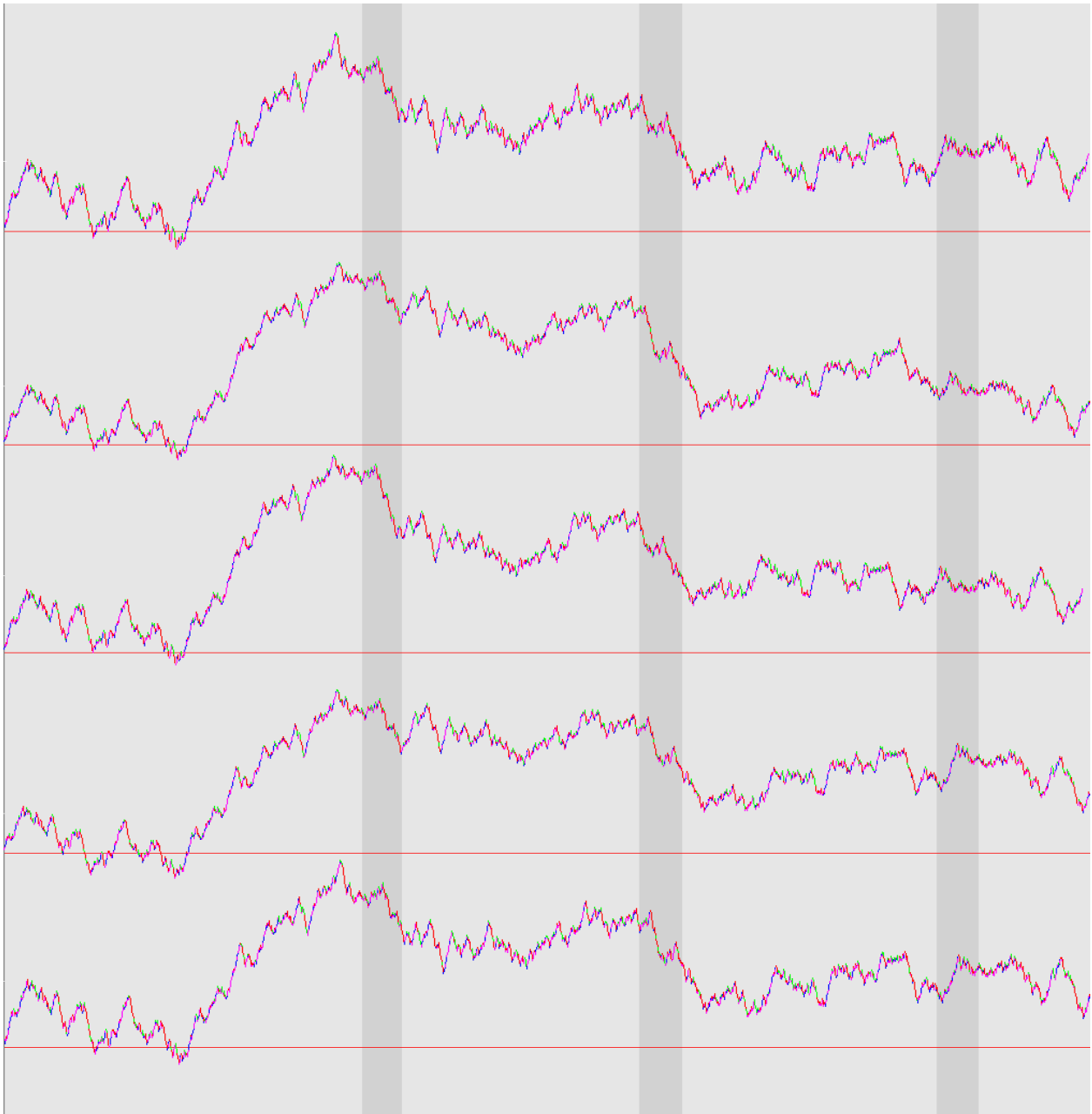


Figure 3: Example W-curves POL gene from wild (top) and example drug resistant HIV-1 strains [10]. Local features of the W-curve geometry can provide landmarks for isolating smaller regions of the sequences for comparison [6]. Highlighted regions show some areas with more easily visible variations between the curves.

distance matrix having zeros on the diagonals. The TSP package does not require a symmetric matrix but in our case we use one.

The colors shown in Figure 2 are assigned by generating a list of 1024 colors and rotating the tour

so that it starts with the first library sequence. After that the colors are assigned by taking the fractional tour length times 1024. For example, if the total tour length were 20 and one of the sequences fell at a cumulative distance of 9 then it would be assigned a color of $9 / 20 * 1024$, or 460.

We have found that this works better than simply assigning the colors sequentially from one to the sample size. Assigning the colors by position along the tour gives some additional visual feedback from the similarity of colors. This can be seen in Figure 2 where the closely related “library” sequences are all red. We have found this color scheme also works well for comparing tours generated from different sections of the same genes: even if the nodes appear in different orders on different graphs the similar colors grouping together help identify the clusters.

A slightly different color scheme can be used to observe the evolution of drug resistant strains in a population. In that case coloring the time scale allows us to watch which direction the group is evolving. With the library sequences colored red and the population samples green through blue over time, the migration of blue dots towards library samples is easy to pick out. The library samples can be drug resistant, or susceptible to different drugs. Either way, the progression across clusters is relatively easy to view.

This approach also works well for integrating samples of an individual over time. We can display results of comparing various regions of an individual to a library of sequences with known clinical results. With the library samples in one color and the individual's sequence of samples colored over time we can see how the various samples migrate between clusters. This provides a nice way to integrate treatment information about an individual that may rely on samples of unrelated genes.

4.4 Combining the TSP and W-Curve

The TSP approach shown here will work for any comparison mechanism: ClustalW, Fasta, or Blast provide a suitable square comparison matrix and generate the graphic results from R. Our use of the W-curve has an advantage due to chunks: we can automate scoring discontinuous regions that may have differing alignments. This matters with HIV-1 where the high rate of SNP's leaves too much white noise in larger areas and the high rate of gaps makes locating the often small, discontinuous areas causing drug resistance difficult. The curve's geometry also provides us a more feature-rich environment in

which to compare the curves. Figure 3 shows the wild (HXB2 standard) and five drug resistant sequences [9,10]. Differences in geometry are visible, even when viewed at different scales [6]. The geometric representation also offers us more options for approximate matching using discrete spatial mathematics than string comparison techniques allow.

5 Conclusions

The W-curve provides us with a way of using landmarks to identify regions of interest and score only the relevant portions of sample sequences. The TSP with Climer & Zhang's boundary techniques offers a fast, effective way to cluster genes. The R statistical package provides us with the tools to analyze and color-code the results for analysis. Taken together this provides us with a useful tool for comparing the status and evolution drug response for HIV-1.

6 Acknowledgments

Military HIV-1 Research Project (MHRP)/Henry Jackson Foundation (HJF) and Workhorse Computing. This work was funded by MHRP (Military HIV-1 Research Project) cooperative agreement (W81XWH-07-2-0067-P00001) between the Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc., and the U.S. Department of Defense (DOD). The cohort work was supported through NICHD by R01 HD34343-03 and partly by a cooperative agreement between the Henry M. Jackson Foundation for the Advancement of Military Medicine and the U.S. Dept. of Defense. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

The authors wish to thank Pranamee Sarma and Scott Zintek for their assistance preparing the W-curve graphics used in this paper.

7 References

- [1] Coffin JM, (1995) HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. *Science* 267, 483-489.
- [2] Brown BK, Sanders-Buell E, Rosa Borges A, Robb ML, Birx DL, et al. (2008) Crossclade neutralization patterns among HIV1 strains from the six major clades of the pandemic evaluated and compared in two different models. *Virology* 375: 529–538.
- [3] Wu D, Roberge J, Cork DJ, Nguyen BG, Grace T (1993) Computer visualization of long genomic sequences, in *Visualization 1993*, IEEE Press, New York City, New York, CP 33:308–315.
- [4] Cork DJ, Lembark S, Tovanabutra S, Robb ML, Kim JH (2010) W-curve Alignments for HIV1 Genomic Comparisons. *PLoS ONE* 5(6): e10829.
doi:10.1371/journal.pone.0010829
- [5] Cork D, Lembark S, Tovanabutra S, Sanders-Buell E, Brown B, Robb M, Wiczorek L, Polonis V, Michael N, Kim J. Application of W-curves and TSP to Clustering HIV-1 Sequences by Epitope. 847-853
<http://www.cs.uga.edu/~hra/2010-proceedings-final/biocomp/volume-i.pdf>
- [6] Supplemental data and code used to generate the W-curves and TSP results shown here are available at
<http://www.bioinformatics.org/wcurve/>.
- [7] Zhou T, Xu L, Dey B, Hessel AJ, Van Ryk D, Xiang SH, Yang X, Zhang MY, Zwick MB, Arthos J, Burton DR, Dimitrov DS, Sodroski J, Wyatt R, Nabel GJ, Kwong PD (2007) Structural definition of a conserved neutralization epitope on HIV-1 gp120. *Feb 15;445(7129):732-7*.
- [8] Climer S, Zhang W (2004) Take a walk and cluster genes: a TSPbased approach to optimal rearrangement clustering. *ACM International Conference Proceeding Series*; Vol. 69, p. 22.
<http://portal.acm.org/citation.cfm?id=1015419>
- [9] Leitner T, Korber B, Daniels M, Calef C, Foley B. (2005) HIV-1 Subtype and Circulating Recombinant Form (CRF) Reference Sequences., <http://hiv.lanl.gov>
- [10] Stanford HIV RT and Protease Sequence Database. See supplemental data for specific resistant drug studies [6].
<http://hivdb.stanford.edu/pages/genotype-rx.html>