frontiers

AUTO-TUNE: FINDING AN OPTIMAL DISTANCE THRESHOLD FOR INFERRING HIV TRANSMISSION CLUSTERS

Steven Weaver [1*], **Vanessa Davila Conn** [2], **Hannah Verdonk** [1], **Joel Wertheim** [3], **and Sergei L. Kosakovsky Pond** [1]

[1] *Center for Viral Evolution, Temple University, Philadelphia, PA, USA*
[2] *Center for Research in Infectious Diseases, National Institute of Respiratory Diseases, Mexico City, Mexico*
[3] *Department of Medicine, University of California, San Diego, CA*

Correspondence*:
Steven Weaver
sweaver@temple.edu

ABSTRACT

Choosing an appropriate distance threshold is an important part of inferring a transmission network to determine the relative growth of clusters within a localized epidemic. This distance threshold determines how close two consensus sequences must be in order for a link to be created between them in the network. Using a distance threshold that is too high can result in a network with many unnecessary links, making it difficult to interpret and analyze. On the other hand, using a distance threshold that is too low can result in a network with too few links, which may not capture key insights into rapidly growing clusters among patients with shared attributes that could benefit from public health intervention measures.

Here, we present a heuristic scoring approach for tuning a distance threshold by associating each tested threshold against the maximal number of clusters created across all thresholds and the difference between the ratio (R_{12}) of the largest cluster in the network to the second largest cluster at each iteration. The number of clusters is normalized between $[0, 1]$ then gated via a Gompertz function transform. Meanwhile, the distribution of all R_{12} ratios are converted to Z scores, and normalized relative to the largest positive Z score across all candidate distances. The priority score is the sum of aforementioned two components.

Published research using the HIV-TRACE software package frequently use the default threshold of 1.5% for HIV pol gene sequences. We apply our scoring heuristic to outbreaks with different characteristics, such as regional or temporal variability, and demonstrate the utility of using the scoring mechanism's suggested distance threshold to identify clusters exhibiting risk factors that would have otherwise been more difficult to identify. For example, while we found that a 1.5% distance threshold is typical for US-like epidemics, recent outbreaks like the CRF07_BC subtype among men who have sex with men (MSM) in China has been found to have a lower optimal threshold of 0.5% to better capture the transition from injected drug use (IDU) to MSM as the primary risk factor. Alternatively, in communities surrounding Lake Victoria, where there has been sustained transmission for several years, we found that a larger distance threshold is suitable to

capture a more risk factor diverse populace with sparse sampling over a longer period of time. Such identification may allow for more informed intervention action by respective public health officials.

Keywords: molecular epidemiology, HIV, network, transmission cluster, surveillance

1 INTRODUCTION

Choosing an appropriate distance threshold is an important part of using a transmission network to track the spread of a contagious disease. This distance threshold determines how close two individuals must be in order for a link to be created between them in the network.

Using a distance threshold that is too small can result in a network with many unnecessary links, making it difficult to interpret and analyze. On the other hand, using a distance threshold that is too large can result in a network with too few links, making it difficult to accurately track the spread of the disease.

To ensure that the transmission network is useful and informative, it is important to carefully consider the appropriate distance threshold. This may vary depending on the specific disease and the context in which it is spreading. For example, a highly contagious respiratory illness may require a smaller distance threshold than a less contagious illness that is primarily spread through direct contact.

In general, the goal is to strike a balance between having enough links to accurately track the spread of the disease, while not having so many links that the network becomes difficult to interpret. This can be achieved through careful analysis and consideration of the specific disease and context.

Overall, choosing an appropriate distance threshold is an important step in using a transmission network to track the spread of a contagious disease. It can help ensure that the network is useful and informative, and can ultimately aid in efforts to control and prevent the spread of the disease.

2 METHODS

2.1 Scoring Heuristic Procedure

Network threshold selection procedure proceeds as follows:

1. For each candidate threshold d_L, in increasing order, ranging from the smallest genetic distance in the dataset, up to either the largest distance or a predetermined maximal threshold, we compute two network statistics: R_{12}, the ratio of the largest cluster to the second largest cluster, and C – the number of clusters in the network.
2. A priority score is assigned to each d_L. This score measures two properties of the threshold: Does R_{12} jump at d_L? How far is the number of clusters C at d_L from the maximal number of clusters over all threshold values? Let there be N overall d_L candidate values, and assume we are examining the ith candidate, d_L^i with $W < i \leq N - W$ (W is a positive integer defined below).
   a. The R_{12} jump is computed by looking at the normalized ratio of the mean R_{12} values computed over the leading window $d_L^{i+1} \ldots d_L^{i+W}$ and the trailing window $d_L^{i-W} \ldots d_L^{i-1}$. The width of the window, W, is defined as $((\lfloor \frac{N}{100} \rfloor, 3), 30)$. The distribution of ratios is converted to Z scores, and normalized relative to the largest positive Z score across all candidate distances, yielding the jump component of the score.

b. The number of clusters, C_i at threshold d_L^i is first normalized to $[0, 1]$ through $\frac{C_{max}-C_i}{C_{max}-C_{min}}$ and next gated via a Gompertz function transform $1 - e^{-e^{-25x+3}}$ This function provides an ad hoc means for penalizing having too few clusters relative to the maximum over all ranges. For example, a threshold that yields 95% of the maximal number of clusters receives a score of 0.996, while a threshold that yields 85% - a score of 0.376.

c. The priority score for d_L^i is the sum of the two components defined in (a) and (b).

3. The threshold with the highest priority score will be selected as the suggested automatic distance threshold, if the score is high enough (1.9 or more), and either of the two conditions hold.
   a. No other thresholds have priority scores of 1.9 or higher
   b. If other thresholds have priority scores of 1.9 or higher, then the range of thresholds represented by these options is small (no more than $log(N)$ times the mean step between successive d_L^i).
4. If no single threshold can be selected in step 3, then the one with the highest priority score is suggested, and an inspection of the plot like the one on the analyze page is recommended to ensure that the threshold is sensible.

2.2 Assortativity

Degree-weighted homophily (DWH) is a measure of similarity between nodes in a network based on their attributes (such as demographic characteristics or behaviors) and their degree (i.e., the number of connections they have to other nodes in the network). It is used to quantify the extent to which nodes with similar attributes tend to be connected to each other more frequently than would be expected by chance. DWH is calculated as the ratio of the observed number of connections between nodes with similar attributes to the expected number of connections between such nodes, based on their degree.

In mathematical terms, it is defined as:

$$DWH = \frac{W_M + W_C - 2W_X}{\frac{d_{in}}{nodes_{in}^2} + \frac{d_{out}}{nodes_{out}^2}} \tag{1}$$

Where

- W_M : Weight of in-group connections
- W_C : Weight of out-group connections
- W_X : Weight of cross-group connections
- d_{in} : In-group degree
- d_{out} : Out-group degree
- $nodes_{in}$: number of in-group nodes
- $nodes_{out}$: number of out-group nodes

DWH ranges from -1 to 1. A DWH value of 0 indicates that there is no more homophily than expected with chance, while a value of 1 indicates that there is perfect homophily (e.g. Birds always link to birds, and only birds). A value of -1 is achieved for perfectly disassortative networks (e.g. Bird never linking with another bird).

DWH is used in social network analysis and in the study of how different attributes are related to the formation of connections between individuals. It is used as a way to measure the similarity of attributes

between individuals in a network. Additionally, randomization is performed by shuffling attribute labels among nodes, then performing DWH computation. This is useful in creating a null distribution of DWH scores under random mixing. A panmictic range is reported by shuffling attributes multiple times and reporting the minimum and maximum score.

2.3 Implementation

The software implementation involves a step-by-step process that utilizes the HIV-TRACE suite of packages. It starts with calculating pairwise distances wth the tn93 tool and a supplied multiple sequence alignment. This generated pairwise distances are supplied to the hivnetworkcsv script while providing the -A keyword argument. A brief outline of the software's implementation are as follows

1. Calculate pairwise distances: The user first calculates the pairwise distances using the tn93 fast pairwise distance calculator, providing the necessary threshold value and the input FASTA file. The command for this step is

```
tn93 -t 0.030 pol.fasta > pairwise_distances.15.tn93.csv
```

   Please note that the threshold should include the maximal range one is intending to test.

2. Compute distance threshold scores: The hivnetworkcsv script is then executed with the required input file, format, and autotune option to generate a tab-separated output file, as shown below

```
hivnetworkcsv -i pairwise_distances.15.tn93.csv -f plain -A 0 > autotune_report.tsv
```

3. Visualize the report: Users can upload the generated autotune_report.tsv file to `http://autotune.datamonkey.org/analyze` for visualization and further analysis of the data. This web-based platform provides an interactive environment to explore scores and other metrics across the range of tested outputs.

4. Run HIV-TRACE: Once AUTO-TUNEd threshold(s) are settled upon after review, the user runs the HIV-TRACE command with the appropriate input FASTA file, distance threshold, and other required arguments. The output is saved as a JSON file. An example command is

```
hivtrace -i ./INPUT.FASTA -a resolve -r HXB2_prrt -t < autotune_threshold > -m 500 -g .05
    ↪> hivtrace.results.json
```

2.3.1 Optional : Compute Assortatviity Metrics

5. Annotate results: The hivnetworkannotate script is used to annotate the results obtained from the HIV-TRACE step with attributes. The script takes the JSON results file, node attributes file, schema file, and a resolve flag as input.

```
hivnetworkannotate -n hivtrace.results.json -a node_attributes.json -g schema.json -r
```

   For more information, users can refer to the hivnetworkannotate documentation.

6. Analyze the results with DWH: After the results file has been annotated, the user can proceed to the assortativity page, http://autotune.datamonkey.org/assortativity, for further analysis of the output.

   AUTO-TUNE is readily accessible on GitHub as part of the hivclustering repository (https://github.com/veg/hivclustering). It is integrated into the command-line interface of the software as the -A or –auto-profile argument. hivclustering is a key component of the HIV-TRACE suite of tools, a resource for the inference, analysis and visualization of HIV transmission networks.

The Degree Weighted Homophily (DWH) calculation tool, an integral component of the assortativity step, is developed using TypeScript, a statically typed superset of JavaScript that ensures robustness and scalability. In an effort to promote accessibility and ease of integration, the DWH tool is packaged and distributed through the Node Package Manager (NPM), enabling researchers and developers to conveniently incorporate this advanced analytical tool into their own projects and workflows. DWH can be used in-browser or as a command line tool, allowing researchers and developers to employ the tool in an interactive command-line interface or integrate it into larger software applications, thus catering to a diverse array of technical needs and preferences. Instructions for usage and installation is found on Github (https://github.com/veg/dwh).

The described workflow offers a systematic approach to analyze potential distance thresholds for one's data with AUTO-TUNE, from calculating pairwise distances to visualizing and annotating results.

2.4 Visualization

Visualizations of AUTO-TUNE results are accessible at http://autotune.datamonkey.org/analyze. It is a dynamic and interactive web-based platform that offers visualization and analysis of results generated by AUTO-TUNE. The website provides a comprehensive view of the data by generating various plots across candidate distance thresholds. These include a score plot, allowing users to identify trends and anomalies across the full range of thresholds. Additionally, it generates a graph showing the number of clusters across candidate thresholds, one of the components that contribute to the score. The site also includes an R1/R2 plot that displays the ratio of the largest cluster to the second largest cluster across candidate thresholds, which is the other metric that contributes to the scoring heuristic.

An assortativity tool is available at http://autotune.datamonkey.org/assortativity, and is an advanced analytical tool engineered to facilitate the calculation of Degree Weighted Homophily (DWH) values. It utilizes the DWH NPM package to generate a tabular representation of DWH values corresponding to each value for a selected attribute annotation, providing an exhaustive examination of the interrelationships for the field. A notable feature is the computation of the panmictic range, which involves a label permutation test to generate the null distribution of DWH values. This feature establishes a comparative baseline that aids in determining the significance of homophily versus what would be expected by chance. Lastly, the site also provides a plot of the fraction of pairwise connections, normalized by degree, for each value pertinent to the selected field. This visual depiction facilitates an intuitive comprehension of the distribution and interconnections within the dataset.

The site aims to offer a user-friendly interface for data visualization, playing an important role in interpreting and understanding AUTO-TUNE's output data. The visualization code is available on Github (https://github.com/stevenweaver/autotune-app/).

2.5 Comparisons with previously published analyses

In conducting our comparisons with the established clustuneR method, we procured our datasets from Wolf et al. (2017) and Vrancken et al. (2017) utilizing the identical approach delineated in Chato et al. (2020). These datasets, namely Middle Tennessee, Seattle, and Alberta, were processed using the workflow prescribed in Section 2.3. This enabled us to determine an optimal threshold for each dataset using our proposed method, AUTO-TUNE. We further executed the command as detailed in step 4 of Section 2.3, deploying thresholds previously established as optimal by Chato et al. (2020).

To perform comparisons, we computed the average degree-weighted homophily score over a set of three-year sliding windows. Specifically, the homophily among nodes was calculated for a collection of date ranges as follows:

$$\bar{H} = \frac{1}{N} \sum_{i=1}^{N} H(w_i) \tag{2}$$

where $\bar{H}$ represents the average degree-weighted homophily score, N is the total number of sliding windows, $H(w_i)$ is the homophily score for the i-th window, and the windows w_i correspond to the date ranges, e.g., '2012-2015', '2013-2016', '2014-2017', etc. This methodology allowed us to compare the "best thresholds" derived from our proposed AUTO-TUNE method against those defined as optimal in Chato et al. (2020).

Second, we set out to compare the thresholds obtained in original investigations with those obtained by AUTO-TUNE. To select the data sets for this analysis, we conducted a scientific literature search to identify studies focused on HIV networks for public health purposes. We then filtered the studies that utilized HIV-TRACE to infer genetic networks and had publicly available sequences. Thus, we attempted to include studies from different countries and regions, enabling us to assess the performance of our method across various epidemic contexts, risk groups, and network sizes in real-data sets that used variable clustering thresholds.

In order to evaluate the influence of sampling density on the genetic distance threshold as determined by AUTO-TUNE, we implemented a strategy of random subsampling from the original dataset sourced from Rhee et al. (2019). This study was selected due to its satisfactory AUTO-TUNE score when utilized in its entirety, as well as its inherent design as a Geographically-Stratified set of 716 Pol Subtype/CRF (GSPS) reference sequence dataset. The dataset, which comprises 6034 samples gathered between 1959 and 2016, was subjected to random subsampling ten times at proportions of 25%, 50%, and 75% of the original sample size. For each subsample, the optimal threshold and associated scores were determined via AUTO-TUNE.

3 RESULTS

3.1 Comparison with clustuneR

We compared results to clustuneR, which employs the recency of sample collection or diagnosis as individual-level weights in a predictive model to estimate the growth of HIV clusters.

3.2 Comparison with Prior Publications Citing HIV-TRACE

3.3 CRF07_BC Network

3.4 Effect on Sampling Density

4 DISCUSSION

AUTO-TUNE operates solely utilizing genetic sequence data to ascertain a decisive threshold. It employs a scoring heuristic, which is based on the number of clusters produced by a pairwise distance threshold and the ratio of the largest cluster to the second largest across a range of possible thresholds using sliding windows.

A key advantage of this approach is its autonomy from supplementary data. When a patient tests positive for HIV, data collection protocols can greatly vary, and additional data are not always available or consistent. However, by leveraging only genetic sequence data, AUTO-TUNE eliminates the need for such information.

Consequently, AUTO-TUNE's performance is consistently controlled, irrespective of the fluctuations seen in data collection protocols after a positive HIV diagnosis. This level of adaptability demonstrates its suitability for integration into various contexts related to HIV, and possibly other viral, cluster detection and response. This versatility underscores the strong methodological foundation of AUTO-TUNE and its potential utility.

4.1 When a Score is Below Two

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

The Author Contributions section is mandatory for all articles, including articles by sole authors. If an appropriate statement is not provided on submission, a standard one will be inserted during the production process. The Author Contributions statement must describe the contributions of individual authors referred to by their initials and, in doing so, all authors agree to be accountable for the content of the work. Please see here for full authorship criteria.

FUNDING

Details of all funding sources should be provided, including grant numbers if applicable. Please ensure to add all necessary funding information, as after publication this is no longer possible.

ACKNOWLEDGMENTS

This is a short text to acknowledge the contributions of specific colleagues, institutions, or agencies that aided the efforts of the authors.

SUPPLEMENTAL DATA

Supplementary Material should be uploaded separately on submission, if there are Supplementary Figures, please include the caption in the same file as the figure. LaTeX Supplementary Material templates can be found in the Frontiers LaTeX folder.

DATA AVAILABILITY STATEMENT

Data are available at GenBank accession numbers JX160108-JX161480,JX498971-JX498972,JX498976-JX498990,JX498992-JX499018,KU190031-KU190839,KY34691-KY37792,KY883695-KY883762,KY888784-KY888875,KY921717-KY921757,MG434786-MG435347,MG435358-MG436769,MH352627-MH355541,MK25MK25548,MN424584-MN427369,MT336755-MT336776,MT368043-MT369927.

REFERENCES

Bbosa, N., Ssemwanga, D., and Kaleebu, P. (2020). Short Communication: Choosing the Right Program for the Identification of HIV-1 Transmission Networks from Nucleotide Sequences Sampled from Different Populations. *AIDS research and human retroviruses* 36, 948–951. doi:10.1089/AID.2020.0033

Brenner, B. G., Ibanescu, R.-I., Osman, N., Cuadra-Foy, E., Oliveira, M., Chaillon, A., et al. (2021). The Role of Phylogenetics in Unravelling Patterns of HIV Transmission towards Epidemic Control: The Quebec Experience (2002-2020). *Viruses* 13, 1643. doi:10.3390/v13081643

Chato, C., Kalish, M. L., and Poon, A. F. Y. (2020). Public health in genetic spaces: a statistical framework to optimize cluster-based outbreak detection. *Virus Evolution* 6, veaa011. doi:10.1093/ve/veaa011

Dalai, S. C., Junqueira, D. M., Wilkinson, E., Mehra, R., Kosakovsky Pond, S. L., Levy, V., et al. (2018). Combining Phylogenetic and Network Approaches to Identify HIV-1 Transmission Links in San Mateo County, California. *Frontiers in Microbiology* 9, 2799. doi:10.3389/fmicb.2018.02799

Ding, X., Chaillon, A., Pan, X., Zhang, J., Zhong, P., He, L., et al. (2022). Characterizing genetic transmission networks among newly diagnosed HIV-1 infected individuals in eastern China: 2012–2016. *PLOS ONE* 17, e0269973. doi:10.1371/journal.pone.0269973. Publisher: Public Library of Science

H, Y., H, W., Y, X., L, H., Y, L., Q, L., et al. (2021). Acquisition and transmission of HIV-1 among migrants and Chinese in Guangzhou, China from 2008 to 2012: Phylogenetic analysis of surveillance data. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases* 92. doi:10.1016/j.meegid.2021.104870. Publisher: Infect Genet Evol

Holmes, E. C., Zhang, L. Q., Robertson, P., Cleland, A., Harvey, E., Simmonds, P., et al. (1995). The molecular epidemiology of human immunodeficiency virus type 1 in Edinburgh. *The Journal of Infectious Diseases* 171, 45–53. doi:10.1093/infdis/171.1.45

Kosakovsky Pond, S. L., Weaver, S., Leigh Brown, A. J., and Wertheim, J. O. (2018). HIV-TRACE (TRAnsmission Cluster Engine): a Tool for Large Scale Molecular Epidemiology of HIV-1 and Other Rapidly Evolving Pathogens. *Molecular Biology and Evolution* 35, 1812–1819. doi:10.1093/molbev/msy016

Liu, M., Han, X., Zhao, B., An, M., He, W., Wang, Z., et al. (2020). Dynamics of HIV-1 Molecular Networks Reveal Effective Control of Large Transmission Clusters in an Area Affected by an Epidemic of Multiple HIV Subtypes. *Frontiers in Microbiology* 11, 604993. doi:10.3389/fmicb.2020.604993

Rhee, S.-Y., Magalis, B. R., Hurley, L., Silverberg, M. J., Marcus, J. L., Slome, S., et al. (2019). National and International Dimensions of Human Immunodeficiency Virus-1 Sequence Clusters in a Northern California Clinical Cohort. *Open Forum Infectious Diseases* 6, ofz135. doi:10.1093/ofid/ofz135

Sivay, M. V., Hudelson, S. E., Wang, J., Agyei, Y., Hamilton, E. L., Selin, A., et al. (2018). HIV-1 diversity among young women in rural South Africa: HPTN 068. *PloS One* 13, e0198999. doi:10.1371/journal.pone.0198999

Vrancken, B., Adachi, D., Benedet, M., Singh, A., Read, R., Shafran, S., et al. (2017). The multi-faceted dynamics of HIV-1 transmission in Northern Alberta: A combined analysis of virus genetic and public health data. *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases* 52, 100–105. doi:10.1016/j.meegid.2017.04.005

Weaver, S., Shank, S. D., Spielman, S. J., Li, M., Muse, S. V., and Kosakovsky Pond, S. L. (2018). Datamonkey 2.0: A Modern Web Application for Characterizing Selective and Other Evolutionary Processes. *Molecular Biology and Evolution* 35, 773–777. doi:10.1093/molbev/msx335

Wolf, E., Herbeck, J. T., Van Rompaey, S., Kitahata, M., Thomas, K., Pepper, G., et al. (2017). Short Communication: Phylogenetic Evidence of HIV-1 Transmission Between Adult and Adolescent Men

Who Have Sex with Men. *AIDS research and human retroviruses* 33, 318–322. doi:10.1089/AID.2016.0061

Yan, H., He, W., Huang, L., Wu, H., Liang, Y., Li, Q., et al. (2020). The Central Role of Nondisclosed Men Who Have Sex With Men in Human Immunodeficiency Virus-1 Transmission Networks in Guangzhou, China. *Open Forum Infectious Diseases* 7, ofaa154. doi:10.1093/ofid/ofaa154

TABLES

Table 1. clustuneR Comparison

Dataset	clustuneR		AUTO-TUNE		
	Threshold	Avg. Homophily	Threshold	Avg. Homophily	Max Score
Middle Tennessee	0.0160	0.0079	0.01431	0.0147	1.25807
Seattle	0.0160	0.0259	0.01354	0.0348	1.53325
Northern Alberta	0.0104	-0.0536	0.01099	-0.0448	1.01678

Table 2. Threshold Comparison with Prior Publications Citing HIV-TRACE

PMID	Country	Collection Date	Threshold Used	AUTO-TUNE
29975689	South Africa	2011-2015	2.5%	2.584%
30574123	USA	1997-2008	2%	1.848%
32500089	China	2008-2015	0.5%	0.675%
32693608	Uganda	2009-2016	1.5%	1.707%
33281803	China	2000-2016	0.5%/0.7%	0.676%
33901684	China	2008-2012	1.5%	1.215%
34452506	Canada	1996-2017	1.5%/2.5%	0.547%
31041344	USA	1997-2017	1.5%	0.927%

Table 3. CRF07_BC DWH and Panmictic Range at Different Thresholds

	Threshold 1.5%		Threshold 0.76%		Threshold 0.19%	
Record	DWH	Panmictic Range	DWH	Panmictic Range	DWH	Panmictic Range
MSM	0.211	[-0.105, -0.205]	0.237	[-0.120,-0.240]	0.292	-0.146 -0.280
Hetero	0.133	[-0.092, -0.190]	0.185	[-0.100,-0.211]	0.25	-0.093 -0.256
PWID	0.168	[-0.001, -0.089]	0.401	[-0.005,-0.081]	0.445	-0.012 -0.129

5 FIGURE CAPTIONS

Figure 1. Hola