StrainHub: A phylogenetic tool to construct pathogen transmission networks

Adriano de Bernardi Schneider[1,2,✉], Colby T. Ford[2], Reilly Hostager[1], John Williams[2], Michael Cioce[2], Ümit V. Çatalyürek[3], Joel O. Wertheim[1], and Daniel Janies[2,✉]

[1]Department of Medicine, University of California San Diego, San Diego, 92103, United States of America
[2]Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte, 28223, United States of America
[3]School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, 30332, United States of America

Summary: In epidemiology of infectious diseases, networks have been used to reconstruct contacts among patients. Summarizing networks using pathogen metadata (e.g., host species and place of isolation) and a phylogenetic tree is a nascent approach. In this paper, we introduce a tool for reconstructing transmission networks in arbitrary space from phylogenetic information and metadata. Our goals are to provide a means of deriving new insights and infection control strategies based on the dynamics of the pathogen derived from networks and metrics. We created a web-based application, called StrainHub, in which a user can input a phylogenetic tree along with characters derived from metadata using their preferred tree search method. StrainHub generates a transmission network for the user metadata based on character state changes mapped on the phylogenetic tree. The user has the option to calculate centrality metrics on the nodes including betweenness, closeness, degree, and a new metric, source/hub ratio. The outputs include the network with values for metrics on its nodes and the tree with characters reconstructed. All of these results can be exported for further analysis.

Availability: https://github.com/abschneider/StrainHub and strainhub.io

Evolution | Epidemics | Phylogeography | Outbreaks

Correspondence: ***adeberna@ucsd.edu (ABS); djanies@uncc.edu (DJ)***

Introduction

New technologies have been shaping responses to outbreaks of rapidly evolving infectious diseases. High-throughput genetic sequencing has allowed rapid characterization of disease outbreaks such as Ebola, Yellow Fever, and Zika viruses (1–3). In contrast, there have been relatively few advancements in how to interpret genomic data and abstract transmission networks from phylogenetic data and metadata (4).

Genomic data and associated metadata (e.g., host species and place of isolation) from pathogens are observations of the variant biology underlying a disease. The combination of these data collected from related isolates of pathogens allow us to abstract disease transmission patterns as a function of the pathogens evolutionary history. Thus, in order to leverage these data and make phylogenetic transmission graphs accessible to the public health scientists, we built StrainHub, a web-based application.

Methods

The StrainHub application is built with multiple R packages and wrapped using Shiny. In order to build a transmission network, the user provides a phylogenetic tree and associated metadata for all terminal taxa.

Ancestry reconstruction. When parsimony is selected, the ancestral state reconstruction step in StrainHub uses the R package Castor (5). Castor performs the ancestral state reconstruction for discrete traits derived from metadata using the parsimony algorithm described by Sankoff (6). Next, based on the results of ancestral state reconstruction, StrainHub outputs a relationship list of the metadata elements as source and destination.
Users have the option to run phylogeography in BEAST (7) and visualize the transmission network based on the tree edges and trait probability for each node. The relationship list is extracted from the tree nodes of a BEAST phylogeography file using the R package treeio (8) to build a directional network and calculate the metrics as described below. This option is used in lieu of the parsimony ancestral state reconstruction step.

Tree and transmission network visualization. We implemented a tree visualization tab using ggtree, ggplot, and plotly to display the metadata mapped to each taxon within the tree (9, 10). We used the R packages igraph and visNetwork to build the transmission networks. igraph provides functions for generating the backbone of the transmission network and calculates the centrality metrics (11). visNetwork provides the R interface to the "vis.js" javascript charting library, allowing an interactive visualization of the transmission network (12).

Transmission network metrics. Multiple centrality measurements are available to determine the relative importance of nodes within a network with respect to the dynamics of pathogen strains (e.g., as they move from place to place or host to host). In this application, we implemented three different centrality metrics: betweenness, closeness and degree (Table 1). Betweenness measures the number of shortest paths between two other nodes that pass through the node of interest, normalized by the number of all pairs of nodes within the network. The higher betweenness score of a node reflects the importance of that node as a hub in the network

Table 1. Summary of Centrality Metrics applied to measure transmission networks on StrainHub.

Metric	Formula	Meaning
Betweenness Centrality	$C_B(i) = \sum \frac{\sigma_{jk}(i)}{\sigma_{jk}}$ Where σ_{jk} = number of shortest paths.	How frequently a node acts as the intermediary node connecting two other nodes on a shortest path.
Closeness Centrality	$C_C(i) = \frac{1}{\sum d(i,k)}$ Where $d(i,k)$ = distance between nodes.	How close a node is as the starting point to other nodes.
Degree Centrality	$C_D(i) = \frac{\sum a(i,k)}{n-1}$ Where $a(i,k)$ = 1 if and only if i and k are connected, otherwise = 0.	How many times edges originate or end on that node (Outdegree / Indegree).
Source/Hub Ratio	$R_{SH}(i) = \frac{\sum S(i)}{\sum H(i)}$ Where $S(i)$ = to all transitions from node i and $H(i)$ = to all transitions to and from node i.	Number of transitions originated on node over the total number of transitions related to that node. A node scoring SHR close to 1 indicates a source, SHR close to 0.5 a hub and SHR close to 0 a sink for the pathogen.

Formula definitions: i,j and k = nodes i, j and k, respectively; d = distance; σ = shortest path; a = edge.

for traffic of the pathogen. Closeness evaluates a node based on the relative sum of all the shortest paths from that node to all other nodes within network. A higher closeness centrality value is associated with how close a node is as a direct point of transmission to other nodes. Degree is the number of edges a node has with other nodes within the network. The higher the number of edges that a node indicates the higher importance that node in terms of being involved in metadata state transitions irrespective of directionality (13).

We further divided degree centrality, into "indegree" and "outdegree". With these values, we created the "Source Hub Ratio" (SHR) metric. SHR is obtained by calculating the ratio of all transitions from the node (outdegree) over all transitions from and to the node (indegree + outdegree), resulting in a value ranging from 0 to 1. Nodes with SHR values close to 0.5 indicate that the node is a hub, SHR close to 1, indicate it is a source, and SHR close to 0 indicates that it is a sink for the pathogen. The SHR metric reflects the importance of a node within the network as the hub, source or sink of the pathogen, ignoring centrality of the node within network (14). The availability of multiple metrics and multiple metadata types allows the user to have flexibility in assessing hypotheses in different contexts for the spread of diseases.

Shiny. We used the Shiny framework to provide StrainHub a flexible, web-based interface. Shiny allows for the generation of interactive applications that can be hosted as standalone web-applications that are locally installed or that are based on a server for use over the web (15). In ether case, the user interacts with StrainHub via the web browser of choice.

Implementation

StrainHub accepts phylogenetic trees in Newick format and metadata in comma-separated value (CSV) format to run the ancestry reconstruction step.

Alternatively, users can run phylogeography in BEAST and use the maximum clade credibility tree in Nexus format as input to visualize the transmission network.

Three outputs are generated during the transmission network analysis: a transmission network plot; a tree plot with the character of interest mapped to the tips of the tree; and a table with all centrality metrics computed. The transmission network graph and the tree files are user-interactive, and a snapshot for further analysis or publication can be exported in PNG format. The centrality metrics table can be exported as a CSV file.

The web-application can be accessed at strainhub.io and the source code is available at https://github.com/abschneider/StrainHub under a GNU General Public License (GPL) v3.0.

Conclusion

We have created a visual analytic tool that enables the user to interpret and communicate phylogenetic data on the dynamics of pathogens over geography and various hosts. In fact the metadata format can be used for any type of categorical data that user can encode (e.g., food sources, or risk factor, or phenotype). The user use should assume the results are only as strong as the underlying phylogenetic data (i.e. sampling across metadata states and taxa, adequate branch lengths and nodal support to reconstruct an ancestor descendent change). With solid datasets, the data in the transmission networks, such as identification of the hubs and sources for the spread of pathogens, will assist health authorities to allocating resources to parts of the network that will do the most to disrupt the spread of the pathogen.

Acknowledgements

We thank Andrew Frick, Dr. Nídia Trovão, and Dr. Tetyana Vasylyeva for suggestions on methodology during the development of this app.

Funding

This work was supported by the National Institutes of Health (NIH) National Institute of Allergy and Infectious Diseases (grant numbers K01AI110181 and AI135992) to JOW and The United States Defense Threat Reduction Agency (DTRA; contract HDTRA1-16-C-0010) to UVC and DJ. This paper is submitted under DISTRIBUTION STATEMENT A. Approved for public release by DTRA. We acknowledge the support of the Department of Bioinformatics and Genomics, The College of Computing and Informatics, The Graduate School, and The Ribarsky Center for Visual Analytics of the University of North Carolina at Charlotte.

Bibliography

1. Joshua Quick, Nicholas J Loman, Sophie Duraffour, Jared T Simpson, Ettore Severi, Lauren Cowley, Joseph Akoi Bore, Raymond Koundouno, Gytis Dudas, Amy Mikhail, et al. Real-time, portable genome sequencing for ebola surveillance. *Nature*, 530(7589):228, 2016.
2. Nuno Rodrigues Faria, Moritz UG Kraemer, SC Hill, J Goes de Jesus, RS Aguiar, Felipe CM Iani, Joilson Xavier, Josh Quick, L Du Plessis, Simon Dellicour, et al. Genomic and epidemiological monitoring of yellow fever virus transmission potential. *Science*, 361(6405): 894–899, 2018.
3. Nuno Rodrigues Faria, Ester C Sabino, Marcio RT Nunes, Luiz Carlos Junior Alcantara, Nicholas J Loman, and Oliver G Pybus. Mobile real-time surveillance of zika virus in brazil. *Genome medicine*, 8(1):97, 2016.
4. Daniel A. Janies, Laura W. Pomeroy, Chris Krueger, Yuqi Zhang, Izzet F. Senturk, Kamer Kaya, and Umit Catalyurek. Phylogenetic visualization of the spread of h7 influenza a viruses. 2015.
5. Stilianos Louca and Michael Doebeli. Efficient comparative phylogenetics on large trees. *Bioinformatics*, 2017. doi: 10.1093/bioinformatics/btx701.
6. David Sankoff. Minimal mutation trees of sequences. *SIAM Journal on Applied Mathematics*, 28(1):35–42, 1975.
7. Alexei J Drummond, Marc A Suchard, Dong Xie, and Andrew Rambaut. Bayesian phylogenetics with beauti and the beast 1.7. *Molecular biology and evolution*, 29(8):1969–1973, 2012.
8. Guangchuang Yu, David K Smith, Huachen Zhu, Yi Guan, and Tommy Tsan-Yuk Lam. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, 8(1):28–36, 2017.
9. Guangchuang Yu, David Smith, Huachen Zhu, Yi Guan, and Tommy Tsan-Yuk Lam. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, 8:28–36, 2017. doi: 10.1111/2041-210X.12628.
10. Carson Sievert. *plotly for R*, 2018.
11. Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.
12. Almende B.V., Benoit Thieurmel, and Titouan Robert. *visNetwork: Network Visualization using 'vis.js' Library*, 2018. R package version 2.0.4.
13. Linton C Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1 (3):215–239, 1978.
14. Adriano de Bernardi Schneider. *Arboviruses: The Hidden Path of an Imminent Threat*. PhD thesis, The University of North Carolina at Charlotte, 2018.
15. Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. *shiny: Web Application Framework for R*, 2018. R package version 1.1.0.