COLT-Viz: Interactive Visualization of Antibody Lineage Trees

Chenfeng He[1,*], Ben S. Wendel[2,*], Jun Xiao[3], Keke Chen[4], Ning Jiang[1,5,$]

Authors and Affiliations:

[1]Department of Biomedical engineering, Cockrell School of Engineering, The University of Texas at Austin, Austin, TX 78712, USA

[2]McKetta Department of Chemical Engineering, Cockrell School of Engineering, The University of Texas at Austin, Austin, TX 78712, USA

[3]ImmuDX LLC, Austin, TX 78750, USA

[4]Department of Computer Science and Engineering, Wright State University, Dayton, OH 45431, USA

[5]Institute for Cellular and Molecular Biology, University of Texas at Austin, Austin, TX 78712, USA.

[*]These authors contributed equally.

[$]Correspondence should be addressed to:

Ning Jiang, Ph.D.

Email: jiang@austin.utexas.edu

Phone: 512-471-4860

Abstract

Many tools have been developed to visualize phylogenetic trees, which is a traditional technique for evolutionary tree analysis. However, due to the unique characteristics of antibody lineage trees, the phylogenetic method cannot adequately construct proper tree structures for antibody lineages, and many other tools have been developed to address this problem. However, there still lacks of an adequate tool to visualize the resulted antibody lineage structures that are more complicated than phylogenetic trees. In addition, high-throughput sequencing-based antibody repertoire profiling enables the counting of the number of transcripts associated with individual antibody sequences, thus more dimensions need to be encoded in the tree structure visualization. Further, users may wish to manually adjust the tree structure for a special context. When doing so, they may wish to maintain some biological constraints that are applicable in antibody lineage tree structure, such as isotype switching constraints or different sampling constraints. Here, we report an interactive visualization tool (COLT-Viz) designed to display the number of RNA copies, number of somatic hypermutations, and sample collection time associated with each antibody sequence as well as the distance to neighboring sequences for each antibody sequence in the lineage. COLT-Viz also allows users to interactively visualize and edit antibody lineage structures while giving users the option to automatically check biological constraints on the edited structures to ensure accuracy. COLT-Viz takes JSON text format as input files and can easily be used to visualize networks with or without the biological constraints. We believe the amount of information that can be displayed for complex antibody lineages, the interactive interface, and the option of checking for biological constraints make COLT-Viz a versatile tool for antibody lineage tree visualization that will guide further biological discoveries.

Key words: bioinformatics software, antibody lineage structure, biological constraints, interactive visualization, intuitive user interface

1. Introduction

Lineage analysis is frequently used by researchers to analyze the evolution of different species of organisms or to trace cellular development within an organism. Specifically in the immune system, lineage analysis has been adopted to trace the mutational evolution of antibodies (1–3) and understand selection pressure (4,5).

Typically, an antibody lineage is represented as a tree structure, which describes the possible development path (i.e., edges in a tree) among all unique antibody sequences (i.e., nodes in a tree). Traditionally, researchers used phylogenetic trees for antibody lineages; however, a phylogenetic tree is a binary tree with all nodes presented as leaf nodes, which is not suitable for antibody lineages (3). Much work have been done to construct tree structures specific to antibody lineages (3,6,7); however, there is still a need for a tool to efficiently visualize these trees. The visualization of antibody lineage trees has a few unique needs: (a) there are multiple features specific to antibody sequences (e.g., number of transcripts, number of somatic hypermutations, isotype, etc), integrating all these properties together for tree visualization will facilitate lineage analysis; (b) interactively visualize the antibody structure by showing antibody sequence information and associated properties when users click on a node of interest. This allows user to have access to detailed information of nodes (antibody sequences) without losing the global view of a complex antibody lineage structure; and (c) interactively modify antibody lineage structure. Oftentimes, multiple valid lineage tree structures exist for an antibody lineage. Based on special considerations, researchers may want to manually alter their tree structures by changing or adding connections between nodes. When doing so, they may want to make sure that these modified connections do not violate certain biological constraints. These constraints can either be biological constraints (8) or constraints determined by experiment design (9).

Although many network visualization tools with powerful functions have been developed (e.g., Cytoscape (10), Graphviz (11)), they were not designed for antibody lineage structure visualization and thus lack the consideration of biological properties and constraints. To our knowledge, there is no tool specifically designed to visualize antibody lineage tree structures with the functionalities mentioned above. For example, LymAnalyzer (12) incorporates FigTree (13) to visualize antibody lineage tree structures. However, FigTree is only designed for phylogenetic tree which is not suitable for antibody lineage structure and does not have an

interactive user interface. Another program, Change-O (6), modifies phylogenetic tree analysis for antibody lineage application, but its visualization function can only generate non-interactive figures. In addition, none of these tools can easily be used to visualize the multiple features of antibody lineages, which are unique to antibody lineage tree visualization.

Bearing the drawbacks of current visualization tools in mind, we developed COLT-Viz for antibody lineage tree structure visualization. COLT-Viz enables researchers to interactively explore lineage trees with multiple antibody sequence properties encoded and automatically checked when the tree structure is manually altered. At the same time, researchers have the option to override error messages and proceed with the desired changes. This tool enables users to discover interesting patterns of antibody evolution process and select interesting nodes (antibody sequences) for further evaluation.

Specifically, COLT-Viz projects antibody sampling time, total mutations for a sequence, mutation distances between two neighboring sequences, and antibody RNA copies of unique sequences onto different elements of each node in the lineage visualization, which allows users to intuitively understand the lineage evolution over time, mutation distances from a germline sequence, mutation distances between neighboring nodes, and the clonal size associated with each unique antibody sequence. Furthermore, users can fine tune the lineage by removing edges and reconnecting nodes with all the constraints automatically checked by the system with an option to either override the error message or make other alternative changes. When users edit the connections, COLT-Viz checks two types of constraints: (a) isotype constraint: as described in previous studies (8,14–16), antibody isotype switch can only proceed according to the order of IGH constant genes located on the chromosome (17), thus users may want to maintain this constraint while editing the lineage; (b) time constraint: as described in Schramm et al. (9), longitudinal information offers another dimension in some biological experiments (18). Although in many cases, sequences from an earlier time point may not necessarily be ancestors of sequences from later time point, in some cases, such as transplantation, where donor samples and recipient samples are experimentally separated, this constraint is critical. Thus, users may want to re-enforce the time constraint when manually changing node connections. Therefore, COLT-Viz also checks if any manual changes on the tree structure comply with the time constraint that sequences from an earlier time point should be ancestors of sequences from a later

time point while give users the option of override this error message and proceed with their intended changes.

2. Implementation

COLT-Viz takes the lineage tree from tree-generating tools (e.g., COLT (7), Change-O (6)) as input and outputs the tree visualization. The input of COLT-Viz contains one node file and one edge file. The node and edge files are encoded in JSON (JavaScript Object Notation) format (http://json.org/), which is a lightweight and human-readable text format. JSON expresses data objects in attribute-value pairs, which makes it an ideal data format for visualizing antibody sequences with multiple attributes. Outputs from tree-generating tools/algorithms can be converted to the JSON text format to be compatible with COLT-Viz. Users can also manually edit the JSON text files to change the lineage tree structure. The node file contains a list of nodes, or unique antibody sequences. Each node element includes the node identity, node description, associated sequence abundance (NRNAs, number of RNA copies, and NREADs, number of sequencing reads), antibody isotype (i.e., antibody heavy chain families, IgM, IgD, IgG, IgA, and IgE), timestamp (i.e., sampling time point, for example, when doing research on vaccine, sequences obtained before vaccination are all assigned to a smaller number (e.g. 20) while sequences obtained 7 days post vaccination are all assigned to a larger number(e.g. 40)), and the number of somatic hypermutations from the germline sequence. Each edge in the edge file is defined as a 3-tuple (parent node identity, weight, child node identity), which are also encoded in the JSON format.

Users can explore the tree visualization, e.g., checking the detailed information of each node, and fine tune the tree structure by editing and moving the edges. Before making changes, COLT-Viz automatically checks the constraints and displays a warning message if either the timepoint or isotype constraint is violated. However, user can choose to override this warning message and proceed with the move. Once old edges are removed and new edges are added, the system will automatically update the entire graph. Finally, the updated tree structure can be saved back to the edge files in the JSON format for later usage.

3. Results

3.1 Lineage Tree Visualization

We visually encode multiple properties of nodes and edges into the lineage tree visualization, so users can understand these properties intuitively. Figure 1 shows an antibody lineage we found in a young African child who experienced acute malaria twice in two consecutive malaria seasons from our previous study (19). There are a total of 23 unique sequences obtained from 4 timepoints in this lineage. The height of each node represents the number of RNA copies associated with that unique sequence. Heatmap is used to color nodes – the warmer the color, the greater the number of somatic hypermutations to the germline sequence. The border of each node is colored to represent different timestamps, so users can easily identify the evolution of the antibody lineage over time.

The tree layout algorithm uses the well-known force-directed graph drawing algorithm (20). It progressively changes the nodes' positions, as if there are physical forces between the nodes. Imagine that neighboring nodes are connected by springs: the force of the spring will prevent the nodes from drifting too far apart or collapsing too close together. This algorithm ideally places the root antibody sequence towards the center of the visualization and unfolds the tree outward to maximize the use of the display area. Thus the directionality of the tree is in general from center to periphery. Knowing this makes it easier for user to identify the root node, also if user assign the root as a specific ID (e.g., 0, same as in Figure 1), then the root node should be easily identified in the tree. Then, by following the directionality, users can easily understand the development of the whole lineage.

Using COLT-Viz, features of the antibody lineages that are not obvious by mining the data can be quickly perceived visually. Using the lineage in Figure 1 as an example, users can quickly identify the root sequence and see that, for this lineage, the sequences at later time points have accumulated more somatic hypermutations compared to sequences collected at earlier time points, and they lie in the outer leaves of the lineage. In addition, the relatively even distribution of node heights indicates that the RNA copies of these unique sequences do not differ widely. These visual features help immunologists quickly glean an overview of the evolution of this antibody lineage over time.

3.2 Tree Editing and Constraint Maintaining

COLT-Viz allows users to alter the network connections by removing and adding edges between nodes. When a new edge is added, the system will check two types of constraints: (a) time constraint, an edge x->y is valid only if x's timestamp is earlier (smaller) than y's; and (b) isotype constraint, an edge x->y is valid only if the isotype is either unchanged or following the class-switching rule: a class-switched sequence cannot switch back to IgM. If either of the constraints is violated, COLT-Viz will display a warning. However, users can override this warning and proceed to add the new edge or heed the warning and cancel the new edge. The constraint checking helps avoid artificial mistakes and maintain interesting time features when users manually modify antibody lineage trees.

3.3 Intuitive graphic visualization user interface

COLT-Viz employs an intuitive graphic visualization user interface (Figure 1). Within the antibody lineage structure displaying window, users can drag on any node to reposition the tree. Users can also click on any node to view the full antibody sequence and annotated information in the sequence display window on the right. In addition, users can change the lineage structure by clicking on a node and specifying the new parent node that they wish to establish a link in the edge editing window.

4. Discussion

COLT-Viz is an interactive tool designed for antibody lineage tree visualization and editing. By using COLT-Viz, researchers can understand lineage trees, validate algorithmic results, and fine-tune the tree structures with additional domain knowledge that automated algorithms cannot capture. Further, COLT-Viz allows researcher to identify nodes (antibody sequences) that bear specific features that are only obvious in graphic settings. Constraints are automatically checked to avoid manual editing errors. Although COLT-Viz is specifically designed for antibody lineage analysis, we believe with small tweaks it can be easily applied to any other type of lineage analysis in biomedical research.

198 **Availability of data and software**

199 Project name: COLT-Viz

200 Project home page: https://github.com/immudx/paper (sample data is included).

201 Operation systems: Platform independent

202 Programming language: Java

203 Any restrictions to use by non-academics: None

204

205 **Abbreviations**

206 IgM (D,G,A,E): Immunoglobulin M (D,G,A,E)

207 **Funding**: This work was supported by NIH grants R00AG040149 (N.J.) and by the Welch
208 Foundation grant F1785 (N.J.). NJ is a Cancer Prevention and Research Institute of Texas
209 (CPRIT) Scholar and a Damon Runyon-Rachleff Innovator. BW is a recipient of the Thrust 2000
210 - George Sawyer Endowed Graduate Fellowship in Engineering.

211

212 **Author contributions:** NJ conceived the idea and directed study; CH, BW and NJ participated
213 in the algorithm design. KC and JX wrote the computer software. CH, BW contributed to
214 software testing and optimization. CH, KC, JX and NJ contributed to the writing of the
215 manuscript. All of the authors read and approved the final manuscript.

216 **Conflict of Interest Statement**: Ning Jiang is a scientific advisor of ImmuDX LLC.

217 **Acknowledgements**: Not applicable

218

219

220

221

222

References

1. Jiang N, He J, Weinstein J a, Penland L, Sasaki S, He X-S, Dekker CL, Zheng N-Y, Huang M, Sullivan M, et al. Lineage structure of the human antibody repertoire in response to influenza vaccination. *Sci Transl Med* (2013) **5**:171ra19. doi:10.1126/scitranslmed.3004794

2. Liao H-X, Lynch R, Zhou T, Gao F, Alam SM, Boyd SD, Fire AZ, Roskin KM, Schramm CA, Zhang Z, et al. Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. *Nature* (2013) **496**:469–76. doi:10.1038/nature12053

3. Barak M, Zuckerman NS, Edelman H, Unger R, Mehr R. IgTree©: Creating Immunoglobulin variable region gene lineage trees. *J Immunol Methods* (2008) **338**:67–74. doi:10.1016/j.jim.2008.06.006

4. Yaari G, Vander Heiden JA, Uduman M, Gadala-Maria D, Gupta N, Joel JN, O'Connor KC, Hafler DA, Laserson U, Vigneault F, et al. Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. *Front Immunol* (2013) **4**: doi:10.3389/fimmu.2013.00358

5. Hershberg U, Uduman M, Shlomchik MJ, Kleinstein SH. Improved methods for detecting selection by mutation analysis of Ig V region sequences. *Int Immunol* (2008) **20**:683–694. doi:10.1093/intimm/dxn026

6. Gupta NT, Vander Heiden JA, Uduman M, Gadala-Maria D, Yaari G, Kleinstein SH. Change-O: A toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics* (2015) **31**:3356–3358. doi:10.1093/bioinformatics/btv359

7. Chen K, Sai V, Gogu A, Wu D, Ning J. COLT: Constrained Lineage Tree Generation from Sequence Data. *Proc IEEE Int Conf Bioinforma Biomed 2016* (2016)

8. Horns F, Vollmers C, Croote D, Mackey SF, Swan GE, Dekker CL, Davis MM, Quake SR. Lineage tracing of human B cells reveals the in vivo landscape of human antibody class switching. *Elife* (2016) **5**: doi:10.7554/eLife.16578.001

9. Schramm CA, Sheng Z, Zhang Z, Mascola JR, Kwong PD, Shapiro L. SONAR: A high-

throughput pipeline for inferring antibody ontogenies from longitudinal sequencing of B cell transcripts. *Front Immunol* (2016) **7**: doi:10.3389/fimmu.2016.00372

10. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Res* (2003) **13**:2498–2504. doi:10.1101/gr.1239303

11. Gansner ER, North SC. An open graph visualization system and its applications to software engineering. *Softw Pract Exp* (2000) **30**:1203–1233. doi:10.1002/1097-024X(200009)30:11<1203::AID-SPE338>3.0.CO;2-N

12. Yu Y, Ceredig R, Seoighe C. LymAnalyzer: A tool for comprehensive analysis of next generation sequencing data of T cell receptors and immunoglobulins. *Nucleic Acids Res* (2015) **44**: doi:10.1093/nar/gkv1016

13. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* (2012) **29**:1969–1973. doi:10.1093/molbev/mss075

14. Iwasato T, Shimizu A, Honjo T, Yamagishi H. Circular DNA is excised by immunoglobulin class switch recombination. *Cell* (1990) **62**:143–9. doi:0092-8674(90)90248-D [pii]

15. von Schwedler U, Jack HM, Wabl M. Circular DNA is a product of the immunoglobulin class switch rearrangement. *Nature* (1990) **345**:452–456. doi:10.1038/345452a0

16. Yoshida K, Matsuoka M, Usuda S, Mori A, Ishizaka K, Sakano H. Immunoglobulin switch circular DNA in the mouse infected with Nippostrongylus brasiliensis: evidence for successive class switching from mu to epsilon via gamma 1. *Proc Natl Acad Sci U S A* (1990) **87**:7829–7833. doi:10.1073/pnas.87.20.7829

17. Murphy, K., & Weaver C. *Janeway's immunobiology*. Garland Science

18. Wu X, Zhang Z, Schramm CA, Joyce MG, Do Kwon Y, Zhou T, Sheng Z, Zhang B, O'Dell S, McKee K, et al. Maturation and diversity of the VRC01-antibody lineage over 15 years of chronic HIV-1 infection. *Cell* (2015) **161**:480–485. doi:10.1016/j.cell.2015.03.004

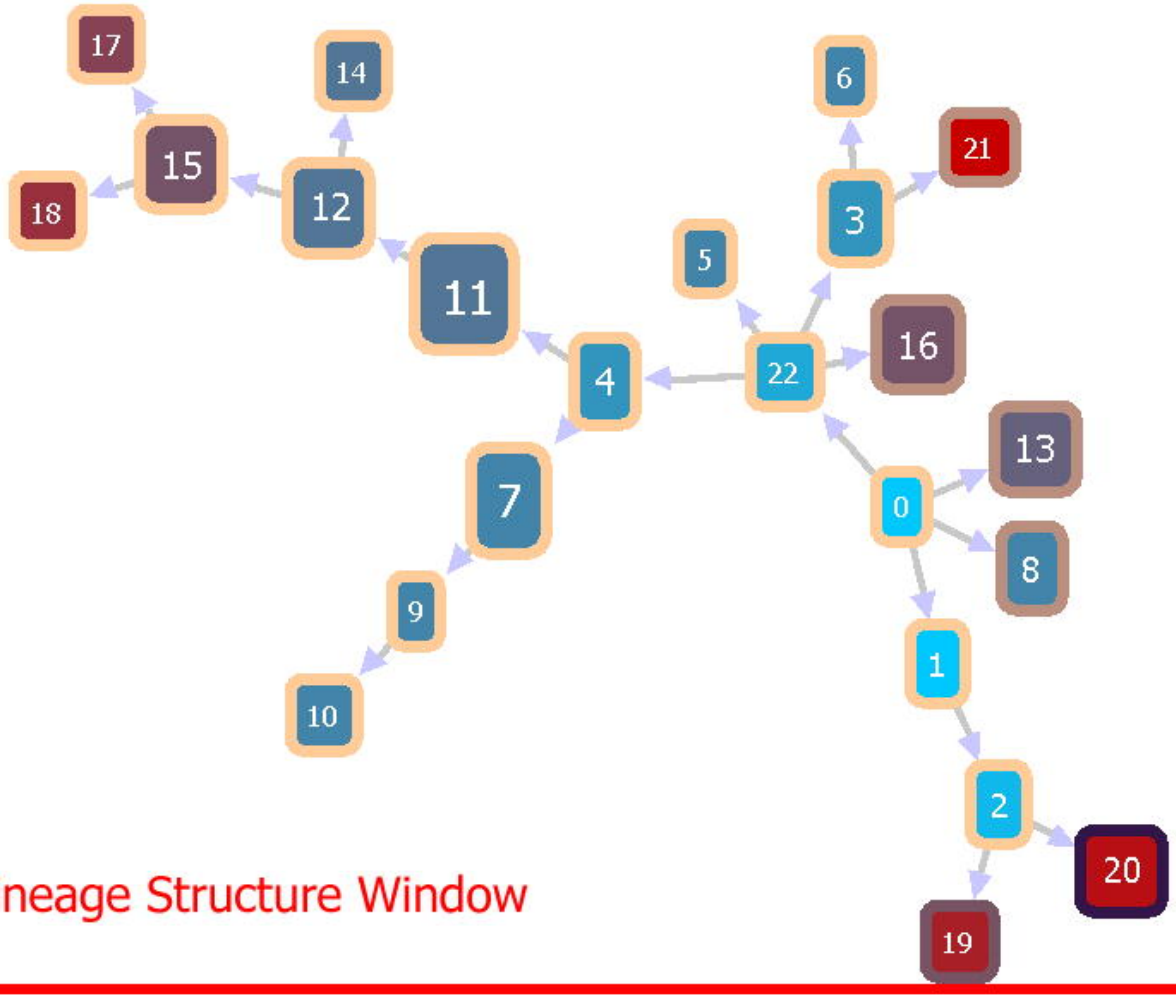
19. Wendel BS, He C, Qu M, Wu D, Hernandez SM, Ma K-Y, Liu EW, Xiao J, Crompton PD, Pierce SK, et al. Accurate immune repertoire sequencing reveals malaria infection driven antibody lineage diversification in young children. *Nat Commun* (2017) **8**:531. doi:10.1038/s41467-017-00645-x

20. Kobourov SG. Spring Embedders and Force Directed Graph Drawing Algorithms. *arXiv Prepr arXiv12013011* (2012)1–23.

Figures:

Figure 1. User interface of COLT-Viz. Within the 'Lineage structure window', height of each node represents the RNA copy of each sequence, color of each node represents the mutation with respect to the germline, color of the border of each node represents the time point and the edge length represents the edit distance between neighboring nodes if user specifies the option. By clicking on a node, the information of this node will be displayed in the 'sequence display window'. User can also edit the connections through 'edge editing window'. The 'lineage structure window' shows an antibody lineage found in a young child who experienced two consecutive malaria infection. 23 sequences were found from 4 time points and visualized using COLT-viz. Four time points are: year 1 pre-malaria infection (light orange node frame for most of the nodes displayed), year 1 acute malaria infection (brown node frame for node 16, 13, and 8), year 2 pre-malaria infection (light purple), and year 2 acute malaria infection (dark purple). Sequence 0 is the root and arrows point to the progeny sequences. Sequences from later time exist in the outer leaves of the lineage, which represents their evolution overtime.

File Settings



Lineage Structure Window

Clicked Node:

NID: 0
LABEL: >|TGCGTGAGTTAT_1_2|iga|1|i6|pre|pbmc
TIMESTAMP: 0
ISOTYPE: IGA
NRNAs 1
NREADs: 2
NMutations: 1
SEQUENCE: ACCTTCACCGGCTACTATATGCACTGGGTGC
GACAGGCCCCTGGACAAGGGCTTGAGTGGATGGGATGGA
CAACCCTAACAGTGGTGGCACAAACTATGCACAGAAGTTTC
AGGGCAGGGTCACCATGACCAGGGACACGTCCATCAGCA
CAGCCTACATGGAGCTGAGCAGGCTGAGATCTGACGACAC
GGCCGTGTATTACTGTGCGACCGAACTGGGGAGTTACTTT
GACTACTGGGGCCAGGGAACGCTGGTCACCGTCTCCTCA

Sequence display window

Change the clicked node's parent to node

Change

Edge Editing Window