Interactive Visualization of Microbial Community Dynamics

Kris Sankaran
Stanford University
Department of Statistics
krissankaran@stanford.edu

ABSTRACT
Modern microbiome studies often revolve around the dynamics of microbial communities, motivating the development of visualization techniques well-suited to hierarchical time series. We report preliminary experiments that apply the timeboxing, sparklines, and degree-of-interest principles to microbial time series, using data from [5]. Our implementation is available at `http://github.com/krisrs1128/treelapse_expers`. The problems detailed in this report have the potential to be an interesting point of contact between the data visualization and microbiome communities, both prompting novel visual design questions and opening the door for feature-rich microbiome studies.

Author Keywords
tree layout, hierarchical structure, time series, Degree-of-Interest

ACM Classification Keywords
H.5.2 Information Interfaces and Presentation: Miscellaneous—*Optional sub-category*

INTRODUCTION

In this paper, we describe several experiments in designing visualizations to facilitate the analysis of microbiome data. Due to the complexity of the associated research questions, this domain presents several challenges from an information visualization perspective. As explained below, resolving these questions requires appropriate representation, navigation, and comparison of collections of hierarchical time series. While we use the language around microbiome analysis to ground our discussion, we note that any visualization principles developed during this case study could have value beyond the microbiome literature, as hierarchical time series appear widely, as fMRI signals across related brain regions or stock prices across similar industries, for example.

Our first basic contribution is to formulate ways in which existing visualization methods can be interwoven to allow study of these hierarchical time series. Specifically, in order to enable navigation and comparisons across several large trees, we adapt ideas from Degree-of-Interest (DOI) trees and Treeversity, while to better study time series along trees, we combine TimeBoxes with generalized linking [3, 7, 10, 1]. Our second contribution is the description of the unmet visualization needs the rapidly moving, and relatively new field of microbiome research.

The microbiome is a term for ecological communities made up of microbes. Characterizing the behavior of these communities can have important health, environmental, and industrial implications. For example, certain human microbiome signatures have been associated with disease, and some microbiome properties have influenced the development of biofuels. As the field matures, more sophisticated statistical and data analytic methods are necessary, as most low-hanging fruit (e.g., the association between a single microbe and a disease), have been picked.

Many microbiome studies revolve around the dynamics of microbial communities, attempting to elucidate their response to environmental changes. Microbes can be arranged on a taxonomic tree, and it is often the case that microbes in closely related subtrees have similar time series of abundances. If a pattern is found in a microbes abundance series, It is interesting to determine the largest subtree in which that pattern occurs. Hence the need for tree representations in visualization.

For an example study involving dynamics, a typical question is: after an antibiotic shock, does the initial community rebound, or does a new microbial configuration take its place, and for how long is the shock visible in microbial abundances? Further, it is useful to compare these dynamics across samples with different covariates. Continuing the antibiotics example, some researchers hypothesize that certain microbial communities tend to be more resilient to antibiotic treatments than others, and that the composition of these initial communities might be related to host physiology.

A few bioinformatics workflows, which take raw sequencing reads and create interpretable microbial count matrices, have become widely adopted in the microbiome community [13, 2]. These workflows offer some visualization and data analysis capabilities, and these methods have become a de facto standard. Most of these analysis are based on dimensionality reduction techniques – see for example [13, 18] – though alternatives are becoming available [6]. The long-term goal of this project is to provide alternatives that are guided by

Submitted for review to CHI 2009.

visual design principles, and which are created with time series data in mind.

RELATED WORK

A number of methods have been proposed for analysis and comparison of tree and time structured data. We list some of these here, those used in our experiments are explained in detail in Section .

The first relevant task in analyzing taxonomically arranged microbial abundances is tree navigation. The focus + context idea is relevant in this setting, since it is hard to gather local information from a full-tree view, and this idea guided the development of DOI trees [3]. DOI trees adjust node layout according to space constraints and a separately specified DOI distribution placed over nodes. [9] enriched the interface, providing breadcrumbs to jump back to ancestor nodes and search for matching descendants. [4] used a DOI tree as a foundation for analysis of time series of variables that traveled across edges in a tree (the career of individual politicians in a political hierarchy).

A separate task is the comparison of different trees; TreeJuxtaposer, TreeVersity and TreeVersity2 are methods developed to fulfill this need [14, 7, 8]. TreeJuxtaposer described a focus + context technique to highlight topological changes in phylogenetic tree structure between different genomic alignment algorithms; these types of trees are in fact available in microbiome studies, though we choose to focus on taxonomic trees instead, since each it is easier to interpret nodes in this case. TreeJuxtaposer also emphasized computational speed, since the phylogenetic trees under consideration there were on the order of tens of thousands of leaves.

TreeVersity introduced the bullet glyph for comparing quantitative variables living at tree nodes. This glyph was overlaid on the nodes of a consensus tree to indicate the direction and magnitude of change from one tree to another (tall and green means the values from the year after was much larger, deep and red means the value dropped). TreeVersity2 abandoned the tree representation entirely, instead vertically laying out bar charts in a way reflecting hierarchical structure. That is, each of the vertically arranged viewed was a small multiple whose bar chart represented changes between two years at a specific depth in the tree; if this node had two children, then two bar charts were positioned just below it, exploiting the fact that the change at the parent is the sum of changes of children.

For the dynamics component of microbiome analysis, we look to methods for studying large collections of similarly shaped time series. [10] introduced timeboxes as a way of visually querying time series data by sketching a general shape with brushes. [16] combined this sketching idea with a tunable clustering parameter, so the querying could be performed on cluster centroids.

The benefit of querying is that it allows the display of individual series without the occlusion associated with plotting all simultaneously. As an alternative to querying, a collection of time series can also be studied by laying out sparklines [15]. In fact, [12] adapted sparklines to interactively display series along tree structures. Further, [11] has shared (but not published) work on visualization of hierarchical time series, basically by grouping related sparklines.

METHODS

Rather than attempting to identify a single representation for hierarchical time series, we experiment with several variations, each accentuating a different visual comparisons. We propose DOI abundance trees and DOI sankey diagrams for comparing microbial abundances across taxonomic subgroups and samples, respectively, while fixing time. For time series analysis, we combine timeboxes with generalized linking to see which taxonomic groups have certain abundance dynamics.

Our DOI implementation parallels the one in [9], though we have not implemented many of the features described there. The core algorithm is split into two separate modules. The first is the calculation of a DOI function from nodes to non-positive integers, specifying the degree-of-interest at each node. Following `http://prefuse.org/gallery/treeview/`, when a node is clicked, it and all its ancestors are given a maximal DOI of 0. Any other node v is given a DOI $f(v) = -k$, where k is the minimum distance from v to any node with DOI equal to 0.

The second part of the algorithm is the calculation of a tree layout which focuses on nodes with high DOI. The last-clicked node is positioned at the center of the screen. If the full tree fits on the screen, no further steps are taken. Otherwise, groups of sibling nodes are removed one at a time – those with the lowest average DOI are removed first – until the specified nodes fit within space constraints.

To reflect our application, we encode microbial abundances in node size and edge width. Nodes and edges with zero microbes within a certain taxonomic group are made small and colored in black; we choose not to remove them by default, as absense is informative. Nonetheless, we allow filtering of nodes based on minimum abundance using a side panel. In principle node abundance could be used to adjust each nodes DOI. Further, we allow search for individual family nodes, accentuating all branches matching the search criterion in red while the user enters text. We also allow the user to change the x and y-spacing of nodes; finding the optimal spacing for these nodes is an open questions.

The DOI sankey is another variation of DOI trees to reflect our scientific application. In this display, each edge is split into several colors, reflecting the source of the sample. The widths of each color within edges encodes abundance for samples. An alternative display, analogous to the bullet glyph in TreeVersity, would be to include a histogram of sample abundances at each node; it is unclear which of the two views will ultimately be more interpretable. The DOI sankey retains filtering and resizing eatures from the original DOI tree, but we have not implemented a search mechanism for this view yet.

The previous views are limited to comparisons that ignore time – we simply average abundances across all timepoints. To allow comparisons across timepoints, we propose leveraging timeboxes and linked brushing. One concrete implementation is to use timeboxes to select time series representing abundances for individual taxonomic groups. These selected series are highlighted along with corresponding nodes in a linked taxonomic tree. We intentionally use a very light grey color for unselected nodes and series, so that it is easier to see actual selections. This is particularly important for maintaining visibility of selected nodes and series for low abundance taxa, though arguably it is not enough.

Alternatively, treeboxes can be drawn over tree nodes, and the series associated with the union of these nodes are highlighted below. Slowly dragging these treeboxes over regions in the tree can reveal gradual shifts in time series dynamics and the composition of microbes within the treebox changes. We avoid drawing both time and treeboxes simultaneously, as we do not want to different types of queries to compete with one another.

Finally, we considered drawing sparklines along tree tips, as in [12], with node sizes and edge widths reflecting abundance average over time. Adjusting brush width along the nodes changes the time window over which the abundances are averages, narrowing or widening edges appropriately. A filtering slider is provided, allowing the removal of nodes below a depth. Further, we include a simplified version of timeboxes – it only allows a single brush at a time – in order to highlight sparklines for queried series.

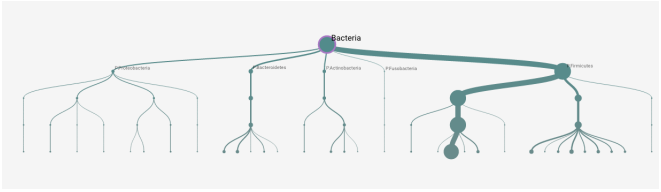

Figure 1. The starting DOI tree display. It is clear that most microbes in this sample belong to the Firmicutes phylum.

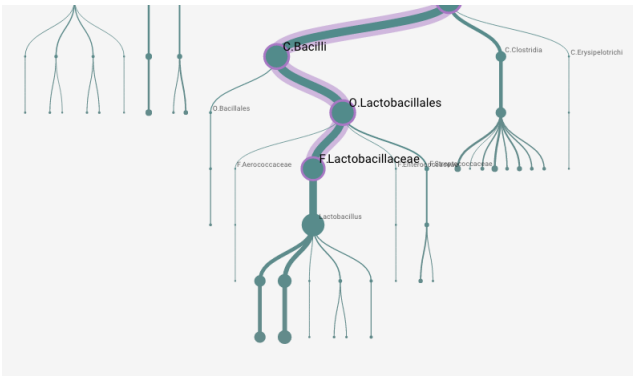

Figure 2. Descending down the Firmicutes phylum, the DOI tree recalculates a layout to display abundances at lower taxonomic orders. The navigation can be done smoothly and interactively.

RESULTS

We use this section to evaluate the utility and limitations of the methods described above, using data from [5]. The goal of this study was to identify relationships between microbial composition and preterm birth. Data at several body sites was collected on 40 individuals across about 20 timepoints (before and after delivery) per person, though the exact sampling times differs across individuals. Data and previous analysis are publically available at `http://statweb.stanford.edu/~susan/papers/PNASRR.html`. For all but the DOI sankey, we limit analysis to the vaginal site in subject 10101; for the DOI sankey, we average across individuals within prespecified Community State Types (CSTs) defined in [5] – these are the centroids of a clustering algorithm.

The DOI tree with encoded abundances makes it easy to see that the vaginal microbiome for subject 10101 is composed mostly of Lactobacillus, with some Clostridia; see Figures 1 and 2. This is consistent with the literature on the vaginal microbiome. However, this view does not allow comparison across timepoints or types of samples. No markers are given for elided subtrees, but this is less of a problem in this data set, since each microbe is at depth 8 (kingdom → species-level genomic variant) in the tree.

The DOI sankey allows comparisons across CSTS. It is relatively easy to see that many smaller abundance phyla are dominated by CST 4. This is consistent with the interpreta-

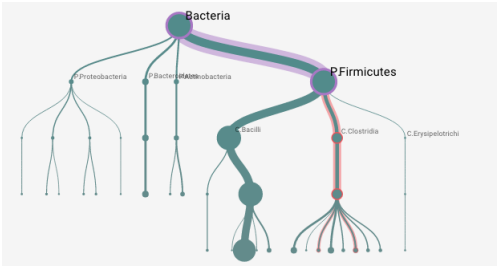

Figure 3. Searching for a term highlights the path towards a matching node in red.

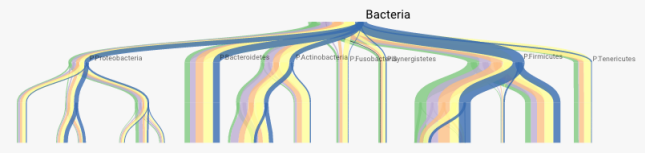

Figure 4. The starting view in the DOI sankey. Subtrees can be opened and closed as in the DOI tree.

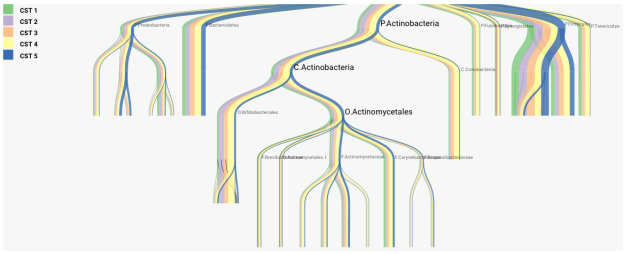

Figure 5. Navigating down the Actinomycetes phylum suggests that it is only highly abundant for samples in CST 4.

tion in [5], which also found that samples in this more diverse state were associated with preterm birth in the host mother. This view doesn't allow us to see any time series for community types, or drill into the raw data either. We would never know if one of the CSTs was strongly linked with one individual, for example, and this kind of information is important in scientific interpretation. A further difficulty is that edges encoding high abundance taxonomic groups sometimes overlap with neighbors, especially when the user chooses a narrow x-width. In some cases, small taxonomic groups are contained in wide edges.

The time and treeboxes make it easy to identify a taxonomic groups with large increases and decreases in abundance. For example, the Lactobacillus and some Clostridia appear to start with lower abundance, but both spike up near the end of the series; see Figure 6. This is perhaps due to delivery, though the display does not make this clear. Also, it can be hard to identify smaller scale changes – the series for species are too close to zero, and their nodes are hard to see. Rescaling the view, or adopting DOI-style focus and context could alleviate this problem.

The sparklines view is useful in its ability to allow comparisons across low abundance microbes. However, to display all the sparklines at once at a certain taxonomic level re-

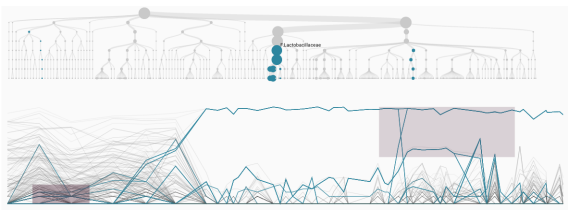

Figure 6. Linking timeboxes and tree display makes it easy to easy to see that a few species from Lactobacillus and Clostridium go from relatively low to very high abundance over the course of the study.

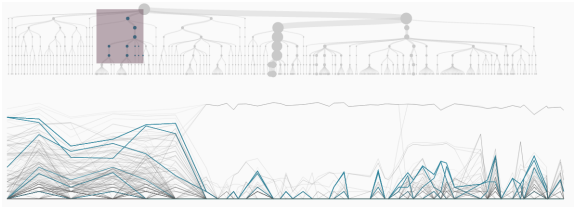

Figure 7. The treeboxes give a quick view into the shared dynamics of a taxonomic group.

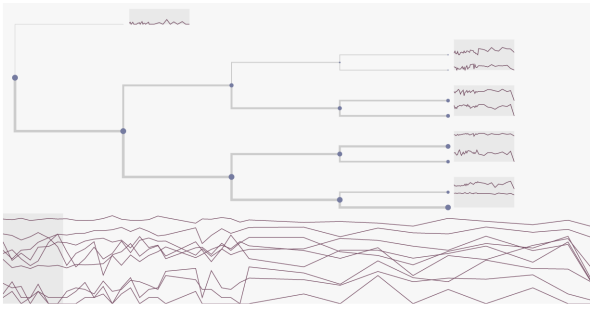

Figure 8. The sparklines lay out time series for leaf nodes in a phylogenetic (not taxonomic) tree.

quires a large space, and the series become impossible to discern. Further, the tree agglomeration operation changes wide swaths of the tree at once, making it impossible to keep track of local changes. As mentioned below, combining sparklines with the DOI principle may be a way to alleviate this problem. Similarly, the changes in node size in response to brushing along sparklines are hard to use in comparisons, because they require the user remember sizes before the brushing operation. An alternative would be to show several temporal frames of the same tree (indeed, this would be the true name for "treelapse").

FUTURE WORK

While this work presents a step in the right directory – even if only in its reference to research in the visualization community – much work needs to be done both in refining existing displays and designing new displays to help navigation of complex microbiome data.

Towards refining existing visualizations, we consider

- In the DOI figures, add markers for elided subtrees, and breadcrumbs to link to ancestors.
- In the original DOI display, try to display sparklines along nodes with high DOI, as a form of semantic zooming.
- In the DOI sankey, consider histograms at each node, rather than full edges encoding abundance.
- In the DOI sankey, develop an approach to hiding clusters on demand, so trees can be viewed in isolation, as in TreeVersity.

We note further that many of the labels and displays may be hard to read, motivating some form of user study.

From a broader visualization design perspective, we identify the following issues,

- The DOI sankey will not be useful without some form of aggregation, but it would be valuable to link to raw data from these aggregated displays, somehow.
- The time and treeboxes are useful in building focus among time series, but perhaps including some form of tree navigation would be useful, to allow focus in tree-views. In particular, the current display does not scale well to very large trees.
- None of the approaches described in this work facilitate comparisons across both time and sample types simultaneously.

We finally note that, to be truly useful to microbiome researchers, it is not sufficient to build a website with a visualization for a single data set. Ideally, there would be a way of generating these interactive visualizations easily after either uploading data or calling function in a statistical programming language. We are considering the development of an HTMLwidgets package to interface phyloseq, an R package for microbiome analysis, with these interactive visualizations [17, 13].

CONCLUSION

We have shared first steps towards better microbiome data visualization. The problems detailed in this report have the potential to be an interesting point of contact between the data visualization and microbiome communities, both prompting novel visual design questions and opening the door for feature-rich microbiome studies.

REFERENCES

1. R. A. Becker and W. S. Cleveland. Brushing scatterplots. *Technometrics*, 29(2):127–142, 1987.
2. J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Pena, J. K. Goodrich, J. I. Gordon, et al. Qiime allows analysis of high-throughput community sequencing data. *Nature methods*, 7(5):335–336, 2010.
3. S. K. Card and D. Nation. Degree-of-interest trees: A component of an attention-reactive user interface. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, pages 231–245. ACM, 2002.
4. S. K. Card, B. Suh, B. A. Pendleton, J. Heer, and J. W. Bodnar. Time tree: Exploring time changing hierarchies. In *Visual Analytics Science And Technology, 2006 IEEE Symposium On*, pages 3–10. IEEE, 2006.
5. D. B. DiGiulio, B. J. Callahan, P. J. McMurdie, E. K. Costello, D. J. Lyell, A. Robaczewska, C. L. Sun, D. S. Goltsman, R. J. Wong, G. Shaw, et al. Temporal and spatial variation of the human microbiota during pregnancy. *Proceedings of the National Academy of Sciences*, 112(35):11060–11065, 2015.
6. A. M. Eren, Ö. C. Esen, C. Quince, J. H. Vineis, H. G. Morrison, M. L. Sogin, and T. O. Delmont. Anvio: an advanced analysis and visualization platform for omics data. *PeerJ*, 3:e1319, 2015.
7. J. A. G. Gómez, A. Buck-Coleman, C. Plaisant, and B. Shneiderman. Treeversity: Comparing tree structures by topology and node's attributes differences. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pages 275–276. IEEE, 2011.
8. J. Guerra-Gomez, M. L. Pack, C. Plaisant, and B. Shneiderman. Visualizing change over time using dynamic hierarchies: Treeversity2 and the stemview. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2566–2575, 2013.
9. J. Heer and S. K. Card. Doitrees revisited: scalable, space-constrained visualization of hierarchical data. In *Proceedings of the working conference on Advanced visual interfaces*, pages 421–424. ACM, 2004.
10. H. Hochheiser and B. Shneiderman. Interactive exploration of time series data. In *Discovery Science*, pages 441–446. Springer, 2001.
11. R. Hyndman. Visualizing and forecasting big time series data. jan 2015. http://robjhyndman.com/seminars/visualizing-and-forecasting-big-time-series-data/.
12. N. Li, Z. Jiang, Z. Liu, and X. Meng. A method of hierarchical time-series data visualization. In *Proceedings of the 6th International Symposium on Visual Information Communication and Interaction*, pages 113–114. ACM, 2013.
13. P. J. McMurdie and S. Holmes. phyloseq: an r package for reproducible interactive analysis and graphics of microbiome census data. *PloS one*, 8(4):e61217, 2013.
14. T. Munzner, F. Guimbretière, S. Tasiran, L. Zhang, and Y. Zhou. Y003. tree juxtaposer: Scalable tree comparison using focus+ conext with guaranteed visibility. *ACM Transactions on Graphics*, 22(3).
15. E. R. Tufte and P. Graves-Morris. *The visual display of quantitative information*, volume 2. Graphics press Cheshire, CT, 1983.
16. Y. Uchida and T. Itoh. A visualization and level-of-detail control technique for large scale time series data. In *Information Visualisation, 2009 13th International Conference*, pages 80–85. IEEE, 2009.
17. R. Vaidyanathan, J. Cheng, J. Allaire, Y. Xie, and K. Russell. htmlwidgets: Html widgets for r. *R package version 0.3*, 2, 2014.
18. Y. Vázquez-Baeza, M. Pirrung, A. Gonzalez, and R. Knight. Emperor: a tool for visualizing high-throughput microbial community data. *Structure*, 585:20, 2013.