Graph Layout by Path-Guided Stochastic Gradient Descent

Simon Heumos[1], Andrea Guarracino[2*], and Erik Garrison[3,4]*

[1]Quantitative Biology Center (QBiC) Tübingen, University of Tübingen, Tübingen, Germany
[2]University of Rome Tor Vergata, Via della Ricerca Scientifica 1, Rome, Italy
[3]Genomics Institute, University of California Santa Cruz, Santa Cruz, CA, USA
[4]Biomolecular Engineering and Bioinformatics, University of California Santa Cruz, Santa Cruz, CA, USA
** Contributed equally.*

1 INTRODUCTION

Pangenome graphs built from raw sets of alignments may have complex structures generated by common patterns of genome variation. These structures can introduce difficulty in downstream analyses, visualization, mapping, and interpretation. Graph sorting aims to find the best node order for a 1D and 2D layout to simplify these complex regions. Pangenome graphs (Eizenga et al., 2020) embed pangenomic sequences as paths in the graph, but to our knowledge, no algorithm takes into account this biological information in the sorting. Moreover, existing 2D layout methods struggle to deal with large graphs. For these reasons, we present a new layout algorithm that orders the nodes of a pangenome graph using a path-guided stochastic gradient descent (SGD) approach.

2 IMPLEMENTATION

Our algorithm is inspired by the work of Zheng and colleagues (Zheng et al., 2018), and it is applicable in 1D and in 2D. In our implementation, the algorithm moves a single pair of nodes at a time, optimizing the disparity between the layout distance of a node pair and the actual nucleotide distance of a path traversing these nodes (Fig. 1): 1. The first node of a pair is a uniform path position pick from all nodes. 2. The second node of a pair is sampled from the same path following a Zipfian[1] distribution. 3. The path nucleotide distance of the nodes in the pair guides the actual layout distance update of these nodes. The magnitude of the update depends on the current learning rate of the SGD.

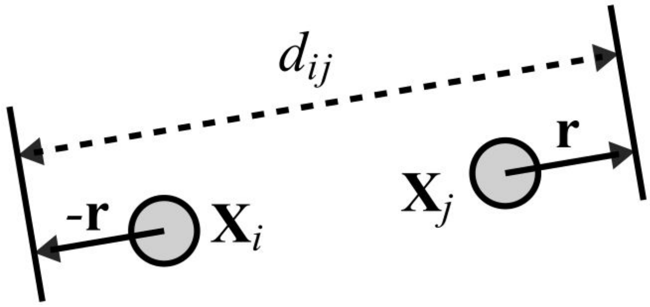

Fig. 1. Outline of the path distance constraint optimization process. (Figure from Zheng et al., 2018)

3 RESULT

Figure 2 shows in 1D an unsorted pangenome graph of a human variable number repeat region. The presence of many and long links which connect distant regions of the pangenome highlights that the node order is not optimal in one dimension. Figure 3 displays the same graph, but sorted by applying our 1D path-guided SGD algorithm. The graph presents shorter links, allowing a clearer visualization. Figure 4 shows a 2D visualization of the previously 1D sorted graph. The coloring of the nodes follows perfectly the node position gradient, reflecting the goodness of the 1D sorting. Moreover, the nodes are well distributed on the 2D plane, without overlapping. The knot highlights the presence of a copy number variation (CNV), also visible as a dense region of links in the 1D plot. This means both visualizations match each other.

4 DISCUSSION

Our multi-threaded implementation[2] presents a working prototype that is based on succinct graph data structures (Eizenga et al., 2020). In progress is the exploration of the path-guided SGD parameter space, in order to get the best layout as quickly as possible. During this process, we also evaluate path-guided metrics in order to measure a graph's stress level. In the future, we want to find out performance boundaries applying the algorithms up to gigabase-scale pangenome graphs. We also plan to compare our proposed 2D graph layouting concept with existing pangenome graph visualization tools. The 1D path-guided SGD implementation is a key step in general pangenome analyses such as the pangenome graph linearization and simplification pipeline smoothxg[3].

ACKNOWLEDGEMENTS

We thank Vincenza Colonna for organizing the Crusco Summer Hackathon and the Forentum Ritrovato museum for hosting it. S.H. acknowledges funding from the Central Innovation Programme (ZIM) for SMEs of the Federal Ministry for Economic Affairs and Energy of Germany.

[1] https://en.wikipedia.org/wiki/Zipf%27s_law

[2] https://github.com/vgteam/odgi/blob/master/src/algorithms/path_sgd.cpp

[3] https://github.com/ekg/smoothxg

REFERENCES

Eizenga, J.M. et al. (2020) Efficient dynamic variation graphs. Bioinformatics, btaa640.

Eizenga, J.M. et al. (2020) Pangenome Graphs. Annual Review of Genomics and Human Genetics, 21, 1.

Wick, E. et al. (2015) Bandage: interactive visualization of *de novo* genome assemblies. Bioinformatics, 31, 20, 3350–3352.

Zheng, J.X. et al. (2018) Graph Drawing by Stochastic Gradient Descent. IEEE Transactions on Visualization and Computer Graphics, 25, 9, 2738–2748.

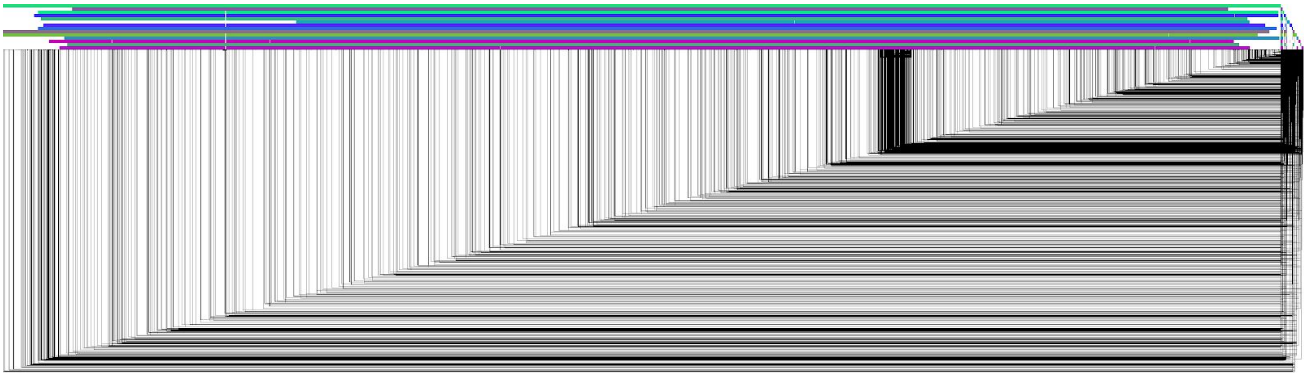

Figure 2: 1D visualization of a pangenome graph built from raw sets of alignments using the edyeet[4] aligner and the variation graph inducer seqwish[5]. The colored bars represent the binned, linearized renderings of the embedded paths versus the pangenome sequence. The black lines under the paths represent the topology of the graph.

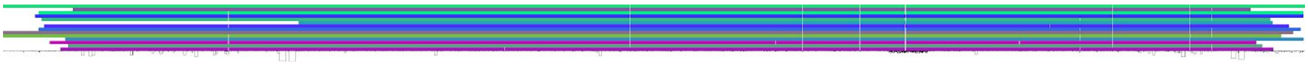

Figure 3: 1D visualization of a sorted pangenome graph built from raw sets of alignments using the edyeet aligner and the variation graph inducer seqwish. It is the same graph as in Figure 2, but after applying our algorithm in one dimension.

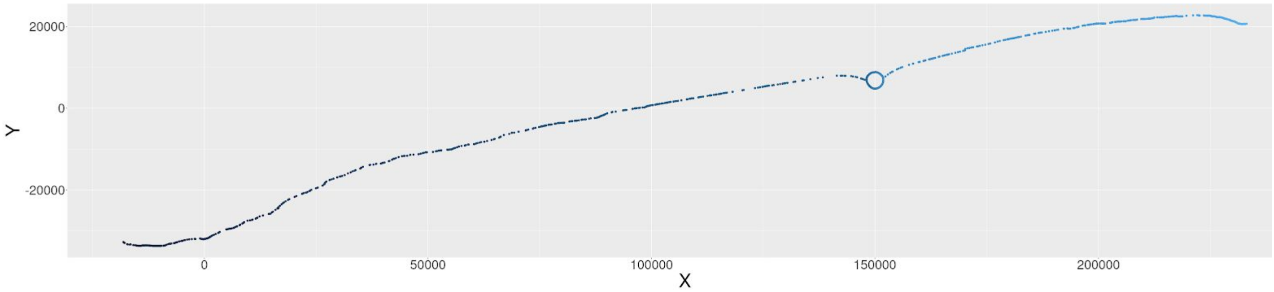

Figure 4: 2D visualization of a 1D sorted pangenome graph built from raw sets of alignments using the edyeet aligner and the variation graph inducer seqwish. It is the same graph as in Figure 3, but displayed in two dimensions. Each dot represents a node. The nodes' x-coordinates are on the x-axis and the y-coordinates are on the y-axis, respectively.

[4] https://github.com/ekg/edyeet

[5] https://github.com/ekg/seqwish