GFAKaleidos: a tool for computing and comparing pangenome graph statistics

Yixin XIANG, Keyu LIU, Leqi WANG and Jianyu ZHOU

Frontiers of Computer Science, DOI: 10.1007/s11704-025-50324-0

Problems & Ideas

- With the growing availability of pangenome graphs across, statistical analyses of these large graphs become challenging. Unlike simple script-based tasks on linear genomes, these analyses require efficient algorithms to compute graph metrics within a limited time and computing resources.

- Ideas: Develop a tool that integrates the computation of key graph-based statistics on a web server, offering pangenome graph analysis in a timely manner.

Main Contributions

- Contributions:
  - Computing comprehensive statistics in three pangenome graph models ;
  - Supporting the integration and comparison of multiple GFA files, which is necessary for evaluating graph-building algorithms;
  - Providing Online Interaction;
  - Enabling Offline Deployment.

Catagory	Statistic	Description
GFA file	File size	The size of GFA files, measuring the compression efficiency of the graph.
	# of segments	The number of segments.
	# of links	The number of links.
	# of paths	The number of paths within graphs, representing the number of haplotypes.
	# of single direction segment	This means that segment A only exists as A- or A+.
	# of bidirectional direction segment	This means that segment A only exists as A- and A+.
	Growth	Estimates how the pangenome expands as additional genome assemblies are added [1].
Vertices	# of vertices	The number of vertices in the graph.
	Vertex size	The distribution of the number of bases in vertex labels across the graph.
	Total length	Total number of bases in vertex labels across the graph.
	N50	A larger N50 indicates that the main path in the graph is longer, suggesting a higher quality of pangenome construction.
	L50	The minimum number of vertices in the pangenome graph whose cumulative length accounts for at least half of the total pangenome size.
	Degree distribution	The degree distribution of vertices, with high-degree vertices often serving as mutation hotspots.
	Dead ends	The number of vertices that end without connecting to other vertices.
	Start ends	The number of vertices with in-degree is 0.
	Coverage	The number of bases/vertices/edges covered by different amounts of paths.
Edges	# of edges	The number of edges in the graph.
	# of loops	The number of loops in the graph, representing unmutated tandem repeats.
	Loop length	The Length distribution of loops in the graph.
	# of cycles	The number of cycles in the graph. Cycles represent tandem repeats but also increase the graph's path complexity.
	Minimum weight cycle	The minimum total length (sum of base counts on vertex labels) of cycles .
	Cycle distribution	The length distribution of cycles in the graph.
Subgraphs	Cuts	The number of cut points or bridges used to assess the connectivity.
	(Weak) Connected components	The number of connected components in the graph, representing co-occurring and co-localized gene families.
	Strongly connected components	The number of strongly connected components in the graph, representing genome regions of high complexity.
	Superbubbles	The number of superbubbles in the graph, representing polymorphisms [2].
	Simple bubbles	The number of simplebubbles in the graph.
	Nested bubbles	The distribution of bubble nesting depths in the graph [3].
	Bubble chains distribution	The length distribution of bubble chains.