

# Walk-preserving shared prefixes

Daniel Doerr

August 16, 2023

## Abstract

We show that collapsing walk-preserving shared prefixes does not affect the set of variants spelled in variation graphs.

A popular data structure in pangenomics are variation graphs [3, 2, 1]. Let  $G = (V, E)$  be a variation graph with node set  $V$  representing the set of *oriented* nodes of  $G$ . Each node  $v \in V$  is associated with a sequence  $\mathbf{t}(v)$  representing a DNA molecule. For any given sequence  $v \in V$  there exists a unique node  $\bar{v} \in V$  with  $\mathbf{t}(\bar{v})$  corresponding to the reverse complementary sequence of  $\mathbf{t}(v)$ . We further assume that  $G$  has a *source* node  $s$  and a *sink* node  $S$ ,  $s, S \in V$ , such that for any  $v \in V$  there exists a walk  $(s, \dots, v, \dots, S)$ . Then the *variants*  $\mathcal{V}_G$  of graph  $G$  is the set of all possible sequences  $\mathbf{t}(s).\mathbf{t}(v_0)\dots\mathbf{t}(v_k).\mathbf{t}(S)$  such that  $(s, v_0, \dots, v_k, S)$  is a walk in  $G$ .

Further,  $\mathbf{p}(\mathbf{t}(u), \mathbf{t}(v))$  denotes the longest common prefix of sequences  $\mathbf{t}(v)$  and  $\mathbf{t}(u)$ . Further, we denote by  $\mathbf{a}(v) := \{u \mid \{\bar{v}, u\} \in E\}$  the *parents* of  $v$  and by  $\mathbf{c}(v) := \{u \mid \{v, u\} \in E\}$  the *children* of  $v$ . A triple  $(u, v, w)$  with  $u \neq w$  and  $\{v, w\} \subseteq \mathbf{c}(u)$  is a *cherry*.

**Proposition 1.** *Let  $(u, v, w)$  be a cherry with  $|\mathbf{p}(\mathbf{t}(u), \mathbf{t}(v))| > 0$ . Let  $G' = (V', E')$  be a graph with*

$$V' = V \setminus \{v, \bar{v}, w, \bar{w}\} \cup \{x, \bar{x}, v', \bar{v}', w', \bar{w}'\}$$

and

$$\begin{aligned} E' = & E \setminus (\{\{v, y\} \mid y \in V\} \cup \{w, z\} \mid z \in V) \\ & \cup \{\{x, p\} \mid p \in \mathbf{a}(v) \cup \mathbf{a}(w)\} \cup \{\{\bar{x}, \bar{p}\} \mid \bar{p} \in \mathbf{a}(\bar{v}) \cup \mathbf{a}(\bar{w})\} \\ & \cup \{\{x, v'\}, \{x, w'\}, \{\bar{x}, \bar{v}'\}, \{\bar{x}, \bar{w}'\}\} \\ & \cup \{\{v', c\} \mid c \in \mathbf{c}(v)\} \cup \{\{w', c\} \mid c \in \mathbf{c}(w)\}, \end{aligned}$$

where  $\mathbf{t}(x) = \mathbf{p}(\mathbf{t}(u), \mathbf{t}(v))$ , and  $\mathbf{t}(v'), \mathbf{t}(w')$  the sequences of  $v$  and  $w$  without  $\mathbf{p}(\mathbf{t}(u), \mathbf{t}(v))$ . Then  $\mathcal{V}_G = \mathcal{V}_{G'}$  holds true in general if and only if  $\mathbf{a}(v) = \mathbf{a}(w)$ .

*Proof.*  $\Rightarrow$  Let  $\mathcal{V}_v := \{\mathbf{t}(s) \dots \mathbf{t}(p).\mathbf{t}(v).\mathbf{t}(c) \dots \mathbf{t}(S) \mid p \in \mathbf{a}(v), c \in \mathbf{c}(v)\} \subseteq \mathcal{V}_G$  and  $V_{v'} := \{\mathbf{t}(s) \dots \mathbf{t}(p).\mathbf{t}(x).\mathbf{t}(v').\mathbf{t}(c) \dots \mathbf{t}(S) \mid p \in \mathbf{a}(x), c \in \mathbf{c}(v')\} \subseteq \mathcal{V}_{G'}$ . By construction, we have  $\mathbf{a}(x) = \mathbf{a}(v)$  and  $\mathbf{c}(v') = \mathbf{c}(v)$  and  $\mathbf{t}(v) = \mathbf{t}(x).\mathbf{t}(v')$ , therefore  $V_{v'} = \{\mathbf{t}(s) \dots \mathbf{t}(p).\mathbf{t}(v).\mathbf{t}(c) \dots \mathbf{t}(S) \mid p \in \mathbf{a}(v), c \in \mathbf{c}(v)\} = \mathcal{V}_v$ . The same logic, applied to all other new variants of  $\mathcal{V}_{G'}$ , completes the proof.

$\Leftarrow$  Assume  $\mathcal{V}_G = \mathcal{V}_{G'}$  and  $\mathbf{a}(v) \subset \mathbf{a}(w)$ . Consider some walk of graph  $G$  of the form  $(s, \dots, p)$ ,  $p \in \mathbf{a}(w) \setminus \mathbf{a}(v)$ , s.t.  $\mathbf{t}(s) \dots \mathbf{t}(p) \notin \{\mathbf{t}(s) \dots \mathbf{t}(p') \mid p' \in \mathbf{a}(v)\}$ . Then the set of sequences  $\{\mathbf{t}(s) \dots \mathbf{t}(p). \mathbf{t}(x). \mathbf{t}(v'). \mathbf{t}(c) \dots \mathbf{t}(S) \mid c \in \mathbf{c}(v)\}$  is a subset of  $\mathcal{V}_{G'}$ , but not of  $\mathcal{V}_G$ , thus contradicting the assumption.  $\square$

## References

- [1] GUARRACINO, A., HEUMOS, S., NAHNSEN, S., PRINS, P., AND GARRISON, E. ODGI: understanding pangenome graphs. *Bioinformatics* 38, 13 (05 2022), 3319–3326.
- [2] HICKEY, G., MONLONG, J., NOVAK, A., EIZENGA, J. M., , LI, H., AND PATEN, B. Pangenome graph construction from genome alignment with minigraph-cactus. *bioRxiv* (2022).
- [3] LIAO, W.-W., ASRI, M., EBLER, J., DOERR, D., HAUKNES, M., HICKEY, G., LU, S., LUCAS, J. K., MONLONG, J., ABEL, H. J., BUONAIUTO, S., CHANG, X. H., CHENG, H., CHU, J., COLONNA, V., EIZENGA, J. M., FENG, X., FISCHER, C., FULTON, R. S., GARG, S., GROZA, C., GUARRACINO, A., HARVEY, W. T., HEUMOS, S., HOWE, K., JAIN, M., LU, T.-Y., MARKELLO, C., MARTIN, F. J., MITCHELL, M. W., MUNSON, K. M., MWANIKI, M. N., NOVAK, A. M., OLSEN, H. E., PESOUT, T., PORUBSKY, D., PRINS, P., SIBBESEN, J. A., TOMLINSON, C., VILLANI, F., VOLLGER, M. R., , BOURQUE, G., CHAISSON, M. J., FLICEK, P., PHILLIPPY, A. M., ZOOK, J. M., EICHLER, E. E., HAUSSLER, D., JARVIS, E. D., MIGA, K. H., WANG, T., GARRISON, E., MARSCHALL, T., HALL, I., LI, H., AND PATEN, B. A draft human pangenome reference. *bioRxiv* (2022).