Appendix for Gap Graph Paper

Jakob Fredslund

Figure 5 shows a small gap graph. In interval I_1, s_2 is the only sequence with gaps while s_1 and s_3 have nucleotides. By assumption these nucleotides are evolutionarily related since they are aligned with each other, and thus a, which lies on the path between s_1 and s_3, must also have nucleotides in that interval, I_1. By consequence, since a is the closest relative of s_2, the indel that caused s_2's gaps must have occurred on the edge between a and s_2. We cannot yet say whether this indel stretched beyond the interval I_1, but we know that an indel occurred on this particular edge in the evolutionary tree.

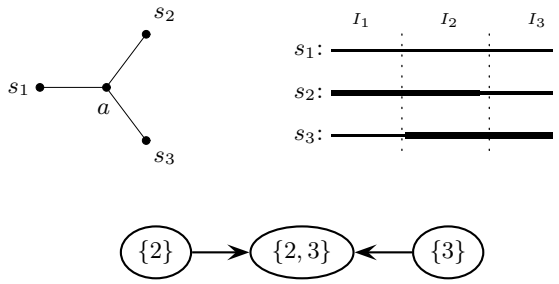


Figure 5. Small gap graph (interval indices omitted in vertices). Thick lines indicate gaps.

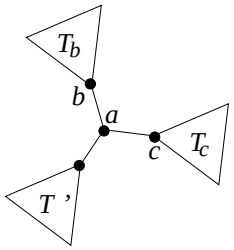
Carrying this argument to the corresponding gap graph in Figure 5, we observe that we can with certainty say that the leaf vertex $\{2\}$ corresponds to an indel occurring on the edge between the leaf s_2 and the closest relative of s_2, a. We generalize this argument in two lemmas.

Lemma 3. *Let $(\mathcal{T}')_I$ be a vertex in a gap graph where $\mathcal{T}'$ is a subtree of the evolutionary tree $\mathcal{T}$. Then the closest relative of $\mathcal{T}'$ in $\mathcal{T}$ has nucleotides in the interval I.*

Proof. Let $a \in \mathcal{T}$ be the closest relative of $\mathcal{T}'$. By construction of the gap graph, all leaves in $\mathcal{T}'$ has gaps in interval I, and $\mathcal{T}'$ is maximal: $\mathcal{T}'$ cannot be extended without including a leaf that has nucleotides in I. Thus, if a is a leaf it has nucleotides in I.

If a is an internal node it has three edges: one to the root of $\mathcal{T}'$, one to a node b, and one to a node c. Let $\mathcal{T}_b$ be the subtree rooted at b obtained by cutting the edge from a to b. Similarly, let $\mathcal{T}_c$ be the subtree rooted at c obtained by cutting

the edge from a to c. Consider the subtree $\mathcal{T}'_b = T_b \cup \{a\} \cup \mathcal{T}'$, rooted at a. Since $\mathcal{T}'_b \supset \mathcal{T}'$ there must exist a leaf $l_b \in \mathcal{T}'_b$ with nucleotides in interval I. All leaves in $\mathcal{T}'$ have gaps in I and a is not a leaf, so $l_b \in \mathcal{T}_b$. A similar argument shows that there must exist a leaf $l_c \in \mathcal{T}_c$ with nucleotides in I. Since the nucleotides of l_b and l_c are evolutionarily related, being aligned over each other in interval I, all nodes on the path between l_b and l_c must also have nucleotides in I. Since a lies on this path, a must have nucleotides in I. □



Lemma 4. (Leaf vertices) *Any leaf vertex $\{l\}_I$ in a gap graph corresponds to an indel that occurred on the edge between l and the closest relative of l (which is unique, since l is a leaf) in the evolutionary tree $\mathcal{T}$.*

Proof. Let l be a leaf node in $\mathcal{T}$, and let a be the closest relative of l. Since $\{l\}$ is a subtree of $\mathcal{T}$, lemma 3 states that a has nucleotides in interval I, and so the indel that caused the gaps of l in I must have occurred on the edge between a and l. □

Lemma 4 interprets leaf vertices in a gap graph. We next discuss vertices with no edges to other vertices. We call such a vertex $(\mathcal{T}')_I$ an *orphan*: its subtree $\mathcal{T}'$ is unrelated to the subtrees in all vertices in neighboring intervals (i.e., $\mathcal{T}' \cap \mathcal{T}_w = \emptyset$ for the subtree $\mathcal{T}_w$ of any vertex w in an interval adjacent to I).

Lemma 5. (Orphans) *Any orphan $(\mathcal{T}')_I$ in a gap graph corresponds to an indel that occurred on the edge between the root of $\mathcal{T}'$ and the closest relative of $\mathcal{T}'$ in the evolutionary tree $\mathcal{T}$.*

Proof. Let r be the root of $\mathcal{T}'$ and let a be the closest relative of $\mathcal{T}'$. Since $(\mathcal{T}')_I$ is an orphan, none of $\mathcal{T}'$'s leaves have gaps in adjacent intervals. Thus, the indels that caused the gaps in $\mathcal{T}'$ do not extend beyond interval I. Pick one of these indels $\theta_{I,\mathcal{T}_m}$, where $\mathcal{T}_m \subseteq \mathcal{T}'$. Since $\theta_{I,\mathcal{T}_m}$ does not cross a cousin edge, Theorem 1 states the existence of a gap graph vertex in interval I with the subtree, $\mathcal{T}_m$. We already have the vertex $(\mathcal{T}')_I$ in interval I, and since vertices in the same interval are disjoint we must have that $\mathcal{T}_m = \mathcal{T}'$. Consequently $\theta_{I,\mathcal{T}_m}$ explains *all* gaps in $\mathcal{T}_m = \mathcal{T}'$. Thus, one indel occurring on the edge between a and r explains the gaps in all the leaves of $\mathcal{T}'$. □

For an orphan, none of the leaves of its subtree have gaps in adjacent intervals. Consider a vertex $(\mathcal{T}')_I$ where *some but not all* of the leaves of its subtree $\mathcal{T}'$ have gaps in an adjacent interval. Such a vertex is called a *patriarch*. A patriarch has at least one edge (zigzag or in-going), but its subtree also has at least one leaf which is not included in a subtree in any vertex in an adjacent interval. We say that such a leaf has *solo gaps*: it has gaps in interval I but nucleotides on both sides (Figure 6 shows some examples of patriarch vertices). Next we give a lemma that interprets patriarchs in a gap graph.

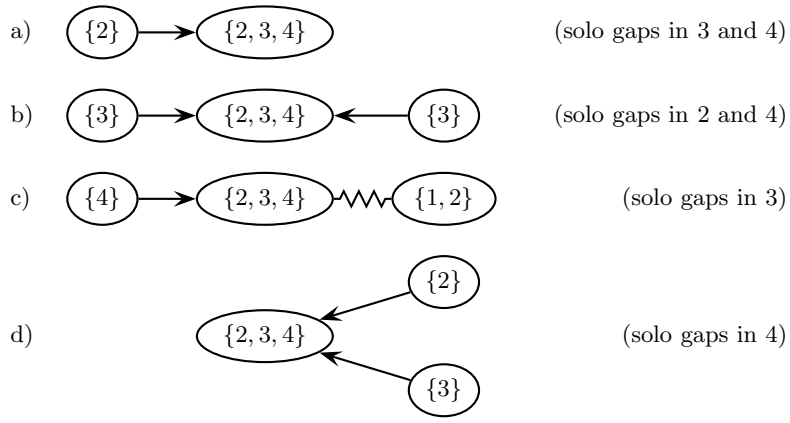


Figure 6. In all four gap graphs the vertex $\{2, 3, 4\}$ is a patriarch: at least one, but not all, of its leaves has solo gaps.

Lemma 6. (Patriarchs) *Any patriarch $(\mathcal{T}')_I$ in a gap graph corresponds to an indel that occurred on the edge between the root of $\mathcal{T}'$ and the closest relative of $\mathcal{T}'$ in the evolutionary tree $\mathcal{T}$.*

Proof. Assume on the contrary that $(\mathcal{T}')_I$ is decomposed. Let c_l and c_r be the alignment columns to the immediate left and right of interval I, respectively. Since $(\mathcal{T}')_I$ is a patriarch there exists a leaf $l \in \mathcal{T}'$ which has nucleotides in both c_l and c_r. Let $\theta_{I_l,\mathcal{T}_l}$ be the indel explaining the gaps of l. By Lemma 2 this indel must extend beyond I, and thus $I_l \supset I$. That means that l has gaps in one or both of c_l and c_r, which is a contradiction. Therefore $(\mathcal{T}')_I$ is not decomposed, and so the gaps of $\mathcal{T}'$ resulted from one indel occurring on the edge between the root of $\mathcal{T}'$ and the closest relative of $\mathcal{T}'$. □

Once a patriarch has been confirmed, its edges are removed and new edges are created directly between its former neighbors on either side (if any are needed). A patriarch has no out-edges, and its subtree is larger than the subtrees of its in-edge neighbors; thus, by the same argument explained in connection with Figure 2, the intervals represented by the patriarch's neighbors are in fact consecutive and so the neighbors may be connected directly if they share leaves. A special case is when the patriarch has two cousins; in this situation the edges cannot be removed. This is to keep the information that no indels with larger subtrees than the subtree of the patriarch can "pass" the patriarch and cause gaps on both sides. Looking again at Figure 6, the edge of examples **a)** and **d)** can be removed since the patriarch in neither case has neighbors on both sides to compare. In example **b)**, the edges disappear and the two $\{3\}$-vertices are merged. In example **c)** the edges are removed, and since the subtrees of the neighboring vertices, $\{4\}$ and $\{1, 2\}$, do not share leaves, no new edges are created.

Some vertices are connected to other vertices only on one side. We call such vertices *end vertices*. Like leaf vertices, orphans, and patriarchs, end vertices can immediately be interpreted as indels as the next lemma shows.

Lemma 7. (End vertices) *Any end vertex $(\mathcal{T}')_I$ in a gap graph corresponds to an indel that occurred on the edge between the root of $\mathcal{T}'$ and the closest relative of $\mathcal{T}'$ in the evolutionary tree $\mathcal{T}$.*

Proof. Assume on the contrary that $(\mathcal{T}')_I$ is decomposed. Without loss of generality let $(\mathcal{T}')_I$ have edges to other vertices on its right side only. Theorem 2 states that the indels that explain the gaps of $\mathcal{T}'$ do not all extend in the same direction. Thus there exists an indel $\theta_{I_l,\mathcal{T}_l}$ with $I_l \supset I$ and $\mathcal{T}_l \subset \mathcal{T}'$ that extends to the left of interval I. Consequently there exists a vertex $(\mathcal{T}^*)_{I^*}$ in an adjacent interval I^* to the left of I with $\mathcal{T}^* \supseteq \mathcal{T}_l$. Since $\mathcal{T}_l$ is also a subset of $\mathcal{T}'$, the vertex $(\mathcal{T}^*)_{I^*}$ has an edge to $(\mathcal{T}')_I$, but that is a contradiction. Thus $(\mathcal{T}')_I$ is not decomposed and the gaps of $\mathcal{T}'$ are explained by one indel occurring on the edge between the root of $\mathcal{T}'$ and the closest relative of $\mathcal{T}'$. □

Besides the lemmas given here a few other reduction rules help reduce the gap graph further. The preprocessing phase makes use of these theoretical results by going over the gap graph in a series of passes, each pass reducing the graph by local applications of the lemmas and reduction rules. Since the removal of edges following a patriarch confirmation may turn previously undecidable vertices into decidable ones, several passes are performed.