Ancestral pangenomes and their phylogenetic reconstruction

Xintong Zhou[1] and David Sankoff[1]

Uiversity of Ottawa
sankoff@uottawa.ca

Abstract. Looking past questions of gene content, we focus on structural variants of the genomes within a pangenome and seek to find a phylogeny where all the ancestral nodes, including the root, are also pangenomes. Representations of pangenomes generally search for compact structures that emphasize common regions or common duplications among the constituent genomes, but necessarily sacrifice some other aspects of gene order. Since the gene order of a monoploid genome is basically just the set of all the gene adjacencies it is composed of, we will consider a pangenome as being made up all the adjacencies of genes appearing in at least one of its constituent genomes. Our key combinatorial tool, *phylogenetic validation*, does not involve optimization, but is simply a filter that removes any adjacencies present in input (extant) pangenomes which are unlikely to have been present in any ancestor, inspired by Dollo's law of irreversible changes. In simulartions, this tool turns out to be extraordinarily efficient in retrieving only adjacencies in the original ancestor.

Keywords: pangenome · adjacencies · phylogenetic validation

1 Introduction

Pangenome graphs [1] represent a set of variant genomes of a species as a directed graph with some device for handling differences in gene content and gene orders among these variants. Differences in gene content are usually discussed statistically in terms of core versus non-core genes, while structural variants due to insertion, deletion and duplication (tandem or otherwise) are amenable to several kinds of graphical representation. However, in the models in this paper, for the purposes of focusing on the variability in gene order, we simply assume that gene content is identical across all the genomes, and does not involve paralogy.

The strategy in previous approaches to comparing DAG or DG representations of gene order between two species has been to extract a linear order from each of the two graphs, doing as little violence as possible to the information contained in each of them, in such a way that these two linear orders are optimally similar in terms of rearrangement distance [2, 3].

In the context of the phylogenetics of a number of species each represented by a pangenome, it does not seem appropriate to simply reduce each pangenome to a linear order and then proceed with a traditional phylogenetic analysis of these linearized genomes. After all, it is not a new idea that an ancestral population may be more or less heterogeneous with respect to the genomes of individuals or groups. This is explicit in the modern recognition of incomplete lineage sorting [4] but it was understood earlier, such as in the description of species as clouds or quasispecies of more or less closely related individuals [5]. In this paper, then, we explore the notion of phylogenetic analysis of pan-genomes, where the root ancestor and all the intermediate ancestors are also pangenomes. In this initial study, we model the pangenomes and their evolution in the simplest terms.

2 Definitions

Structure. A pangenome consists of a set of related genomes $G = \{g_1, ..., g_\gamma\}$, which are unichromosomal and linear, and which all contain the same n genes. A gene x is denoted by the set of gene ends $\{x^h, x^t\}$, where h (heads) and t (tails) are assigned arbitrarily. The distinct identity of each genome g_i resides in its gene order, which is a set Adj_{g_i} or simply Adj_i, of $n+1$ "adjacencies", ordered pairs containing two ends from two different genes (x, y), plus two terminal pairs representing the ends of the genome $(0, x^h)$ (or $(0, x^t)$ and $(0, y^h)$ (or $(0, y^t)$, such that all $2n$ gene ends are in exactly one pair of the $n+1$ adjacencies in Adj_i. This set contains all information about the structure of a genome.

Evolution. Evolution of the pangenome proceeds by a number of inversions (reversals) affecting each of its constituent genomes independently.

An inversion in genome g_i replaces two adjacencies from from Adj_i by two new pairs, with reversed order, as illustrated in Figure 1.

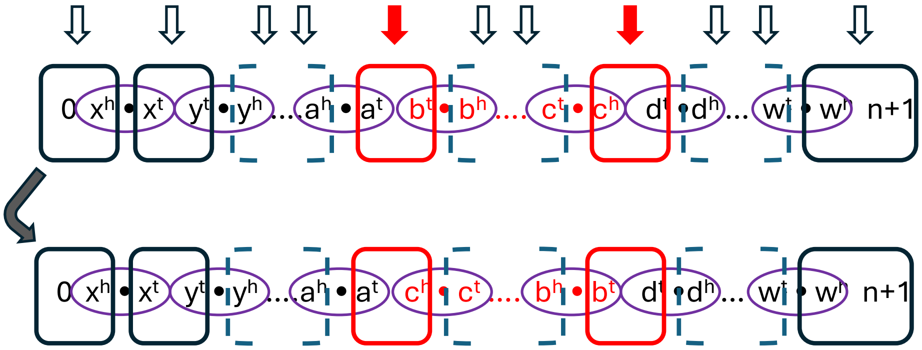

0 x^h • x^t y^t • y^h ... a^h • a^t b^t • b^h .... c^t • c^h d^t • d^h ... w^t • w^h n+1
0 x^h • x^t y^t • y^h ... a^h • a^t c^h • c^t .... b^h • b^t d^t • d^h ... w^t • w^h n+1

Fig. 1. Inversion involving adjacencies between genes a and b and between c and d. Genes enclosed in ovals, adjacencies in rectangles. Dashed incomplete rectangles contain just one member of an adjacency. Arrows indicate potential break points

Phylogeny. We will consider evolution on a phylogenetic tree including of a root pangenome vertex, three descendant pangenome vertices, each of which may represent a modern (extant) pangenome (terminal vertex - degree 1) or an intermediate ancestor vertex (degree 3). Each intermediate ancestor has two descendants, each of which is either another intermediate ancestor pangenome or a terminal pangenome. Although there is a natural temporal orientation, namely from the root towards the modern genomes, in simulating data, topologically the tree is a binary branching tree, with the root having degree 3, like all other non-terminal vertices.

Measurement. We write $pairs(g_i) = |Adj_i|$ and measure the similarity between two genomes g_i and g_j as $pairs(g_i \cap g_j) = |Adj_i \cap Adj_j|$, and eventually between two pangenomes Y and Z, $pairs(Y \cap Z) = |(\cup Adj_{g_i \in Y}) \cap (\cup Adj_{g_j \in Z})|$, as the number of adjacencies they contain in common. The count of adjacencies takes into account neither the relative order of the two genes in the genome nor the heads/tails identity of the gene ends involved.

3 Generating divergent modern pangenomes from an ancestral pangenome

3.1 Generation

The ancestor. The ancestral pangenome X is simulated by independently generating three genomes X_1, X_2, X_3 from the sequence $1, 2, 3, \cdots, 100$, using r random reversals of lengths sampled from a negative binomial with mean 10 and variance 400, appropriately truncated. We write $X_i \in X$ for $i = 1, 2, 3$, and the set of adjacencies in X is $Adj(X) = \cup_{i=1}^{3} Adj_i$. We explore parameter values $r = 5, 10, 20, 40, 80$ and 120.

The first generation. Three descendant genomes A_i, B_i, C_i are generated from each genome X_i in pangenome X using r random reversals of lengths sampled from the same negative binomial distribution as before. Then the descendant pangenomes are $A = \{A_1, A_2, A_3\}, B = \{B_1, B_2, B_3\}$ and $C = \{C_1, C_2, C_3\}$ and, for example, $Adj(A) = \cup_{i=1}^{3} Adj_{A_i}$.

Further branching. If a descendant pangenome D is itself to be considered an (intermediate) ancestor of two other pangenomes F and G, the three genomes in D each produce two further descendants, one which becomes part of F and the the other part of G.

4 Inference

4.1 Phylogenetic validation and the pangenomic median

Our key reconstruction technique is based on Dollo's idea that, in certain biological contexts, phylogenetic characters, such as the adjacencies we study here, are gained only once and can never be regained if they are lost [7]. This is realized in an unrooted tree by the property that the set of vertices containing the character are connected. It is a necessary and sufficient condition, valid both for terminal vertices (or degree 1) and internal (ancestral) ones (degree 3 in an unrooted binary tree).

Formally, the connectedness condition for a character can be satisfied by a set of non-terminal nodes of a tree or, trivially, by a single terminal node. For phylogenetic reconstruction, however, we require that an adjacency be present in the genomes associated with at least two terminal vertices, so that by connectedness we can reconstruct that it must have been present in their most recent common ancestor. Otherwise, if it were present only in one terminal set, it could not be inferred as present in any of the non-terminal vertices.

In an unrooted binary branching tree, each non-terminal vertex subtends three subtrees, as in Figure 2. For an adjacency to be used in constructing a phylogenetic tree and the output sets, clearly each adjacency must be present in at least two of the three subtrees, as illustrated in Figure 1. More precisely, each adjacency must be present at least in one terminal vertex set in at least two of the three subtrees. We call these adjacencies "phylogenetically validated". The possibility that an adjacency originates twice or more over the phylogenetic time span, so that connectedness is not assured, is not zero, but very small, at least for small or moderate rates of evolution, so that errors in the validation process would be rare.

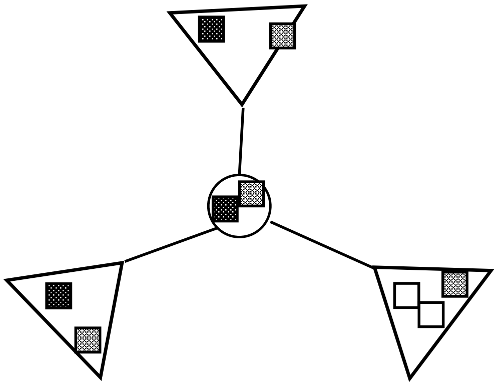


Fig. 2. Necessary condition for adjacencies to appear at an internal vertex associated with an ancestral pangenome of a binary branching phylogenetic tree. Light shaded adjacency (small square) appears in all three trees (triangles) subtended by the internal vertex (circle). Dark shaded adjacency appears in only two of the trees. Unshaded adjacency appears in only one subtree so does not affect internal vertex. The shaded adjacencies are "phylogenetically validated" with respect to the internal vertex. The unshaded one is not validated. Adapted from [8]

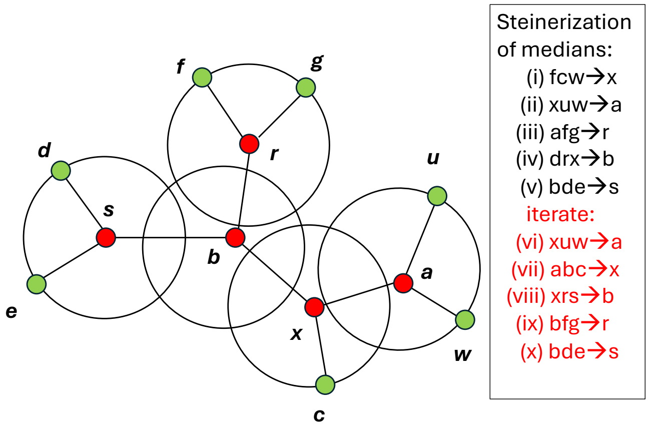
The median. The pairs in common between pangenomes A and B are $Adj_A \cap Adj_B$, between pangenome B and C are $Adj_B \cap Adj_C$ and between C and A are $Adj_C \cap Adj_A$. Then all the pairs in at least two of the three pangemomes are

$$\boxed{\text{X'=(Adj}_A \cap Adj_B) \cup (Adj_B \cap Adj_C) \cup (Adj_C \cap Adj_A).} \tag{1}$$

Equation (1) is an expression of the phylogenetic validation criterion, excluding adjacencies that are in only one of the pangenomes A, B or C, as well as adjacencies that are in none of them.

Steinerization The phylogenies we are modeling and inferring are situated in historical time, with an original "root" vertex representing ancestor X at time zero and all edges directed away from this vertex.

In solving the small phylogeny problem through an iterative "steinerization" process, originally introduced in [9], we first select any three modern pangenomes each located on a different subtree subtended by X. All gene pairs occurring in at least two of these three are considered to form a first estimate



Steinerization of medians:
(i) fcw→x
(ii) xuw→a
(iii) afg→r
(iv) drx→b
(v) bde→s
iterate:
(vi) xuw→a
(vii) abc→x
(viii) xrs→b
(ix) bfg→r
(x) bde→s

Fig. 3. Calculating the ancestral pangenomes through steinerizing based on the medians.

5 Simulations

The steinerizing process calculates the ancestral (root and intermediate ancestors) pangenomes by iterating the median problem for all non-terminal vertices until convergence which we illustrate in Figure 3 for seven terminal vertices and four non-terminal vertices.

For our simulations, however, we used the smaller tree in Figure 4 with only six terminal vertices and three non-terminal vertices. For $r = 5, 10, 20, 40, 80$ and 120, we generated genomes X_1, X_2 and X_3 from the sequence $1, 2, \ldots, 100$ using r inversions for each. We set the ancestor pangenome $X = (X_1, X_2, X_3)$ and generated the intermediate ancestors A, B and C as described in Section 3.1 above. From these ancestors we then generated the "extant" pangenomes R, S, V, W, T, Z.

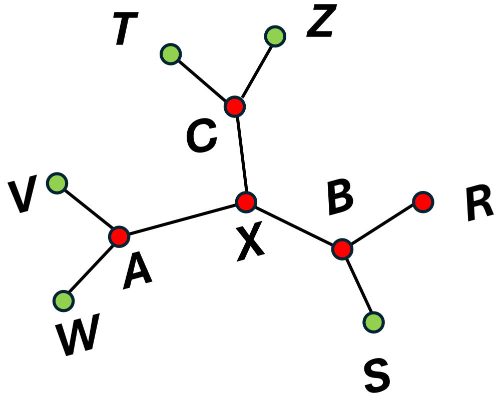


Fig. 4. Phylogeny with ancestor pangenome X used in simulations.

With these simulated data, we could then reconstruct estimated intermediate ancestor pangenomes A', B', C' in terms of the adjacencies in the extant pangenomes that were filtered through the phylogenetic validation criterion.

Finally we used the intermediate ancestors to construct X'. The entire inference procedure was interated as in Figure 3 until convergence. Each simulation

inversions	$\|Adj_X \cap Adj_{X'}\|$	$\|Adj_X \backslash Adj_{X'}\|$	$\|Adj_{X'} \backslash Adj_X\|$
5	125	2	4
10	137	14	8
20	121	69	12
40	44	195	9
80	3	276	3
120	0	290	1

Table 1. Results of the inference process. The parameter r measuring the rate of evolution ranges from 5 per time period to 140. The first two columns show that up to $r = 10$, almost all of the adjacencies in X are recovered in X'.

was repeated 100 times and the mean numbers of adjacencies is reported in Table 1 and Figure 5. These results are remarkable in that for $r = 5$ and even $r = 10$, almost all the adjacencies in X are recovered in X', and few extraneous adjacencies manage to make it into X'. On the other hand, it is clear that increasing the inversion rate to 40 or higher will defeat the method.

The power of the phylogenetic validity filter is clear from Table 1, when we we see the hundreds of adjacencies that are filtered away either in the reconstruction of A', B' and C', or in the final reconstruction of X'.

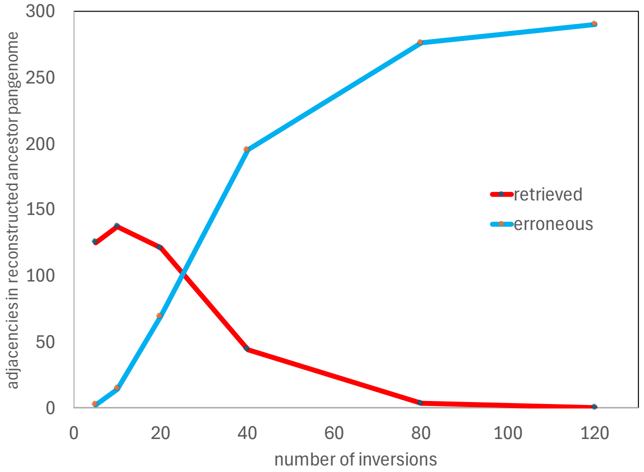


Fig. 5. Effect of evolutionary rate on number of reconstructed adjacencies.

inversions	union of $(R, S, T, V, W, Z) \setminus X'$	union of $(A, B, C) \setminus X'$
5	229	79
10	424	155
20	780	318
40	1293	600
80	1630	810
120	1691	855

Table 2. Adjacencies in extant pangenomes and in intermediate ancestors that are filtered out by the phylogenetic validity criterion.

6 Large phylogeny

In the pangenomic context, there is a major difference with other phylogenetic problems in the small phylogeny context, namely the use of phylogenetic validation instead of some optimization criterion. In the large phylogeny case, however, there is little difference in the basic intractibility of the problem, necessitating exhaustive approaches, heuristics and the like. Here we have six terminal vertices, and only 105 different possible phylogenies. In the present study, however, we simply evaluated one additional tree using the same data

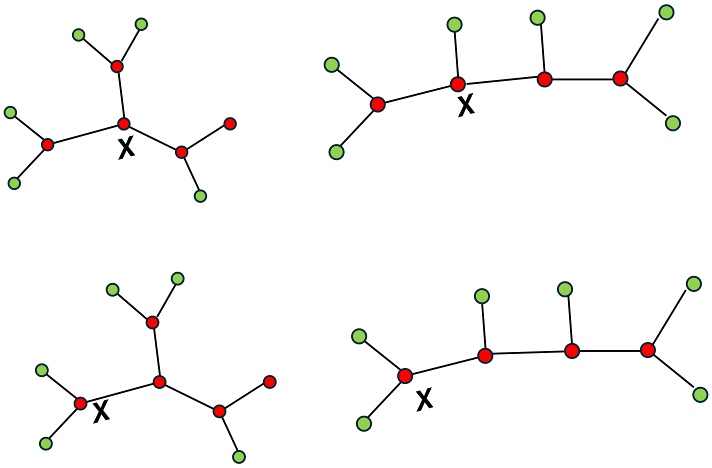


Fig. 6. example

The results of using the second tree to reconstruct the ancestor at the origin of the data were equivocal - slightly fewer original adjacencies recovered and also slightly more extraneous adjacencies in X'. The potential of the method for for large phylogenies awaits further investigation.

7 Discussion and Conclusions

The most striking result from this work is the power of the phylogenetic validation criterion based on Dollo's principle to weed out the massive amounts of recently generated adjacency data to preserve the original gene order information in the original pangenome.

Our model is extremely simple. Moreover no suitable data exists to our knowledge for even a more relaxed and parameterized model. Nevertheless, we submit that we have showed proof of principle for a new approach to ancestral pangenome reconstruction, which is itself a new objective..

References

1. Eizenga JM, Novak AM, Sibbesen JA, Heumos S, Ghaffaari A, Hickey G, Chang X, Seaman JD, Rounthwaite R, Ebler J, Rautiainen M (2020) Pangenome graphs. *Annual Review of Genomics and Human Genetics* **21**:139-62.
2. Zheng C, Lenert A, Sankoff D (2005) Reversal distance for partially ordered genomes. *Bioinformatics*: **21**:502-8.
3. Zheng C, Sankoff D (2005) Genome rearrangements with partially ordered chromosomes.*International Computing and Combinatorics Conference*.52-62.
4. Maddison WP, Knowles L, Lacey T. (2006) Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology* **55**: 21-30.
5. Eigen M, Schuster P (1977) A principle of natural self-organization. *Naturwissenschaften* **64**:541-65.
6. Bergeron A, Mixtacki J, Stoye J (2006) A unifying view of genome rearrangements. In: Bücher P, Moret BME (eds) *Algorithms in Bioinformatics,* Lecture Notes in Computer Science **4175**: 61-16.
7. Dollo L. (1893) Les lois de l'évolution. *Bulletin de la Société Belge de Géologie, de Paléontology et d'Hydrolgie* **7**; 164-166.
8. Xu Q, Sankoff D (2023) Gene order phylogeny via ancestral genome reconstruction under Dollo. Lecture Notes in Bioinformatics **13883**: 100-111.
9. Sankoff, D., Cedergren, R., Lapalme, G.: Frequency of insertion-deletion, transversion, and transition in the evolution of 5S ribosomal RNA. Journal of Molecular Evolution 7, 133–149 (1976)