Purpose:

Tumors are the result of an evolutionary process [1]. Clonal trees describe tumor evolution by defining ancestor and descendant relationships between mutations. The comparison of clonal trees plays an important role in benchmarking software that generates trees from bulk sequencing data and in generating consensus trees from pools of tree candidates. Several techniques have been developed to measure the distance between trees, these techniques provide only a single number [2]. However, these techniques provide only a single number, and reveal little about how the structure of the trees determines their distance. To fill this gap, we have implemented a visualization tool that, given two input trees, outputs a visualization comparing them according to a distance measure selected by the user. The visualizations highlight the areas of the trees that contribute the most to the distance between them. We currently support four such distance measures, each described below.

Parent-Child

Overview: The parent-child distance counts the number of parent-child pairs that appear in one tree and not the other. It is the only of our four distance measures that we visualize using edge coloration rather than node coloration.

Details: A mutation *a* is a parent of the mutation *b* if the node containing *a* is the parent of the node containing *b*. The parent-child distance between two trees is the number of parent-child relationships that appear in one tree and not the other.

We visually encode parent-child distance using edge colorings. The contribution of an edge is equal to the number of parent-child pairs lying along that edge that do not appear in the other tree. We color edges with greater contribution blue, and edges with a lower contribution yellow. Figure 1 illustrates this approach.

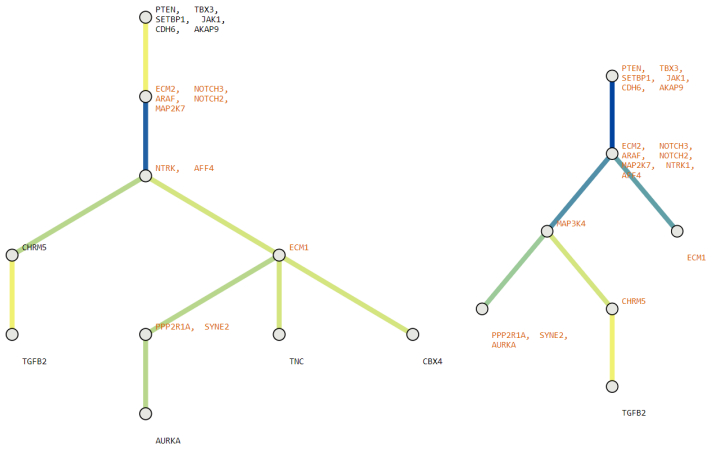

Fig 1. The visualization of the parent-child distance of two sample trees, taken from triple-negative breast cancer data.

Ancestor-Descendant

Overview: Ancestor-descendant counts the number of ancestor-descendant pairs that appear in one tree and not the other. We visualize it using node colorings for the most highly contributing nodes.

Details: The ancestor-descendant distance generalizes parent-child by allowing a contributing pair of nodes to be connected by a directed path rather than an edge. In our implementation, each ancestor-descendant pair contributes 1 if it does not appear in the other tree, and 0 otherwise. The contribution of each mutation is determined by the number of times it appears as an ancestor or a descendant in a contributing ancestor-descendant pair. The contribution of each node is equal to the sum of the contributions of the mutations that lie at that node. Nodes with greater contribution are colored dark blue, and nodes with smaller contribution are colored yellow.

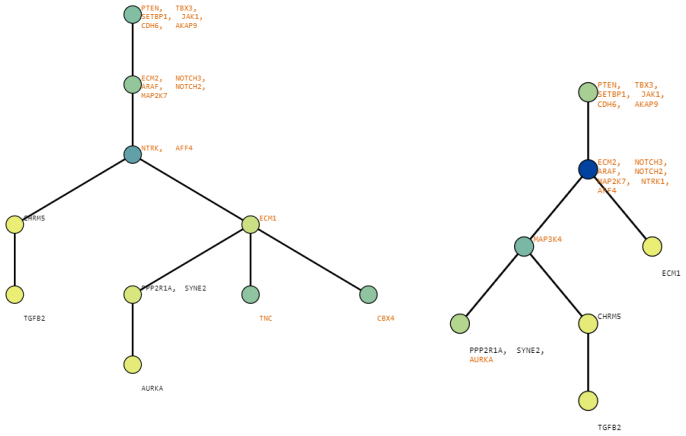

Fig 2. The visualization of the ancestor-descendant distance of two sample trees, taken from triple-negative breast cancer data.

Common Ancestor Set (CASet)

Overview: CASet emphasizes mutation differences close to the root of the tree. We color a node according to the amount that mutations lying at it contribute to the distance.

Details: If mutations *i* and *j* appear in a tree, the *common ancestor set* of *i* and *j* (denoted $C(i,j)$) consists of the set of mutations that are ancestors of both *i* and *j*. The *Jaccard distance* between two sets *A* and *B* is given by

$$\mathrm{Jacc}(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

Note that *Jacc*(*A*,*B*) yields a number between 0 and 1 giving the proportion of elements not shared between *A* and *B*. If *Jacc*(*A*,*B*) = 0, then *A* = *B*, and if *Jacc*(*A*,*B*) = 1, then *A* and *B* do not share any elements.

The *CASet distance* between trees T_1 and T_2 with a combined set of mutations *M* is computed as follows:

$$\mathrm{CASet}(T_1,T_2) = \frac{1}{\binom{|M|}{2}} \sum_{\{i,j\}\subseteq M} \mathrm{Jacc}(C_1(i,j), C_2(i,j))$$

Here, $C_1(i,j)$ and $C_2(i,j)$ are the common ancestor sets of *i* and *j* in T_1 and T_2, respectively. Thus, the CASet distance measure looks at each pair of mutations present in the trees and computes its common ancestor set in each tree. It takes the Jaccard distance between the two resulting common ancestor sets, and then averages this across all mutation pairs. The CASet distance measure emphasizes tree differences close to the root of the tree, as mutations near the root will appear in a large number of common ancestor sets.

From a visualization perspective, the difficulty lies in visually encoding information about the distance between two trees. Our tool assigns a contribution value to each mutation in each tree. The contribution of each appearance of a mutation is equal to the total amount that it contributes as an ancestor in a common ancestor set. More precisely, let $cont_1(p)$ denote the contribution of the mutation *p* in the tree T_1, and let $cont_2(p)$ denote the contribution of *p* in the tree T_2. Then

$$cont_1(p) = \frac{1}{\binom{|M|}{2}} \sum_{\{i,j\} \subseteq M} \frac{|C_1(i,j) \cap \{p\} - C_2(i,j) \cap \{p\}|}{|C_1(i,j) - (C_1(i,j) \cap C_2(i,j))|} \text{Jacc}(C_1(i,j), C_2(i,j))$$

$$cont_2(p) = \frac{1}{\binom{|M|}{2}} \sum_{\{i,j\} \subseteq M} \frac{|C_2(i,j) \cap \{p\} - C_1(i,j) \cap \{p\}|}{|C_2(i,j) - (C_1(i,j) \cap C_2(i,j))|} \text{Jacc}(C_1(i,j), C_2(i,j))$$

Note that the numerator in the really complicated fraction in $cont_1(p)$ is equal to 1 if and only if p appears in $C_1(i,j)$ and not $C_2(i,j)$. Similarly, the numerator in the complicated fraction in $cont_2(p)$ is equal to 1 if and only if p appears in p appears in $C_2(i,j)$ and not $C_1(i,j)$. The contribution of a node is equal to the sum of the contributions of the mutations lying at that node.

Example: If a mutation appears near the root in the left tree, and near a leaf in the right, it will appear in many common-ancestor sets of mutations in the left tree, and few on the right. Each time such a mutation appears in a common ancestor set of two mutations lower down on a tree, the distance is increased.

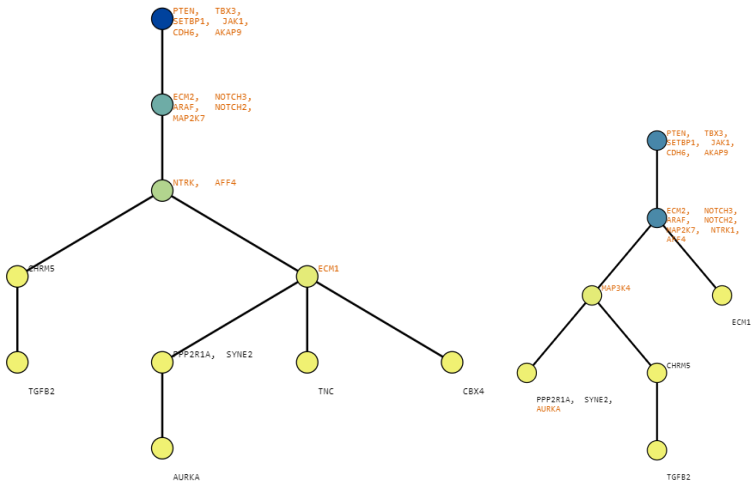

Fig 3. The visualization of the CASet distance of two sample trees, taken from triple-negative breast cancer data.

Distinctly Inherited Set Comparison (DISC)

Overview: The DISC distance measure highlights mutation differences further down in the tree than the CASet distance measure. We color a node according to the amount that mutations lying at it contribute to the distance.

Details: If mutations i and j both appear in the tree T_k, the distinctly inherited ancestor set $D(i,j)$ consists of the set of mutations that are ancestors i and not ancestors of j. The DISC distance measure is the average Jaccard distance between all corresponding inherited ancestor sets between trees T_1 and T_2:

$$\mathrm{DISC}(T_1, T_2) = \frac{1}{|M|(|M|-1)} \sum_{(i,j)\in M^2, i\neq j} \mathrm{Jacc}(D_1(i,j), D_2(i,j))$$

Note that nodes close to the root will appear in many ancestor sets, so they won't appear in very many distinct ancestor sets, so DISC de-emphasizes mutation differences near the root. We compute the contribution of a mutation p to the trees T_1 and T_2 as follows:

$$cont_1(p) = \frac{1}{|M|(|M|-1)} \sum_{(i,j)\in M^2, i\neq j} \frac{|D_1(i,j)\cap\{p\} - D_2(i,j)\cap\{p\}|}{|D_1(i,j) - (D_1(i,j)\cap D_2(i,j))|} \mathrm{Jacc}(D_1(i,j), D_2(i,j))$$

$$cont_2(p) = \frac{1}{|M|(|M|-1)} \sum_{(i,j)\in M^2, i\neq j} \frac{|D_2(i,j)\cap\{p\} - D_1(i,j)\cap\{p\}|}{|D_2(i,j) - (D_1(i,j)\cap D_2(i,j))|} \mathrm{Jacc}(D_1(i,j), D_2(i,j))$$

An example of our visualization in action:

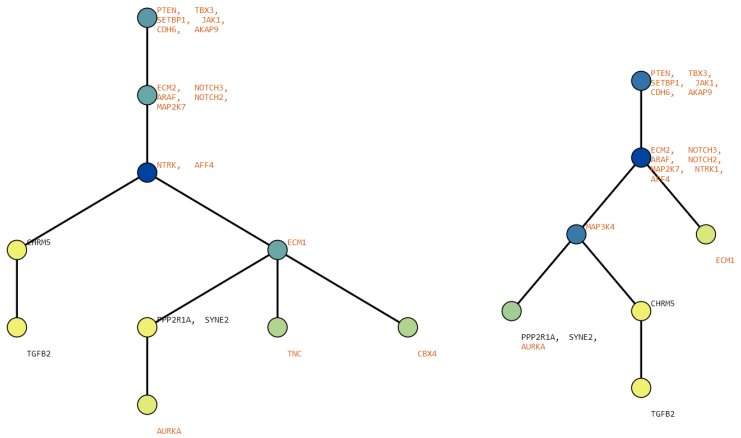

Fig 4. The visualization of the DISC distance of two sample trees, taken from triple-negative breast cancer data.

References

[1] Nowell PC. The clonal evolution of tumor cell populations. Science. 1976 Oct 1;194(4260):23-8. doi: 10.1126/science.959840. PMID: 959840.

[2] Zach DiNardo, Kiran Tomlinson, Anna Ritz, Layla Oesper, Distance measures for tumor evolutionary trees, Bioinformatics, Volume 36, Issue 7, 1 April 2020, Pages 2090–2097, https://doi.org/10.1093/bioinformatics/btz869