

# Supplementary material for: Triplet-based similarity score for fully multi-labeled trees with poly-occurring labels

Simone Ciccolella<sup>1</sup>, Giulia Bernardini<sup>1</sup>, Luca Denti<sup>1</sup>, Paola Bonizzoni<sup>1</sup>, Marco Previtali<sup>1</sup>,  
and Gianluca Della Vedova<sup>1</sup>

<sup>1</sup>Department of Computer Systems and Communication, University of Milano-Bicocca, Milan, Italy

## 1 Proof of Lemma 2

Besides the five possible configurations already listed in the proof of Lemma 1, the multi-labeled model admits four additional configurations, due to the extension of the definition of LCA of two labels. In the first three additional cases, two labels are assigned to the same node and the third one to a different one. Without loss of generality, suppose that  $a$  and  $b$  label the same node. These cases are: (i)  $\text{LCA}(a, b) = \text{LCA}(b, c) = \text{LCA}(a, c)$  and the LCA is a node in  $\{a, b, c\}$ , (ii)  $\text{LCA}(a, b) \neq \text{LCA}(b, c) = \text{LCA}(a, c)$  and both LCA are nodes in  $\{a, b, c\}$ , and (iii)  $\text{LCA}(a, b) \neq \text{LCA}(b, c) = \text{LCA}(a, c)$  and only  $\text{LCA}(a, b)$  is a node in  $\{a, b, c\}$ . The remaining case (iv) is the simplest one in which  $a, b, c$  label the same node of  $T$ , implying that  $\text{LCA}(a, b) = \text{LCA}(b, c) = \text{LCA}(a, c)$ .

Figure 2 in the main text reports cases (i) to (iv) from left to right.

## 2 Experimental comparison of MP3

We show here an experimental comparison of the different versions of MP3, demonstrating the reason we decided to use  $\text{MP3}_\sigma$  as our default measure. We show in Figure S1 (a) that  $\text{MP3}_\sigma$  combines the best aspect of  $\text{MP3}_\cap$  and  $\text{MP3}_\cup$  while  $\text{MP3}_G$  is, as expected, the average of the two. The same result can be seen in Figure S1 (b), while Figure S1 (c) display the effect of the sigmoid, where as the trees become less similar the value move towards the union.

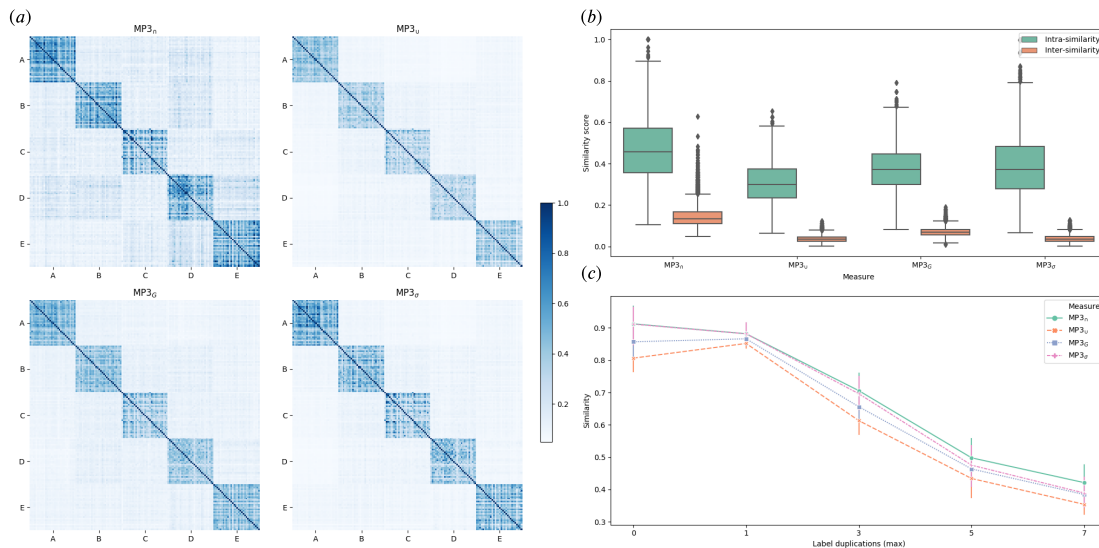


Figure S1: (a) Heatmaps displaying the scores between all the 150 simulate trees from the second experimental setting. (b) Distribution of the similarities between the trees in the same class (Intra-similarity) and in different classes (Inter-similarity) for the 5 classes. (c) Effect of label duplication on the similarity scores. Similarities are the average of 15 trees generate from the same base with the specified value of maximum duplication from the previous experiment.

### 3 Effect of label sliding

We present a comparison of the measure in the case of a label sliding from left to right on the lowest level of a binary tree. The trees are compared against the first tree  $T_0$ . When the label slide to the same subtree no difference is found for all measure, as expected. On the other hand we show a higher decrease in similarity for MP3 with respect to the other measure as the label slides. Interestingly for the last two trees CASet and DISC collapse to the same value showing no difference between the two. MLTD express very little difference between the last four trees.

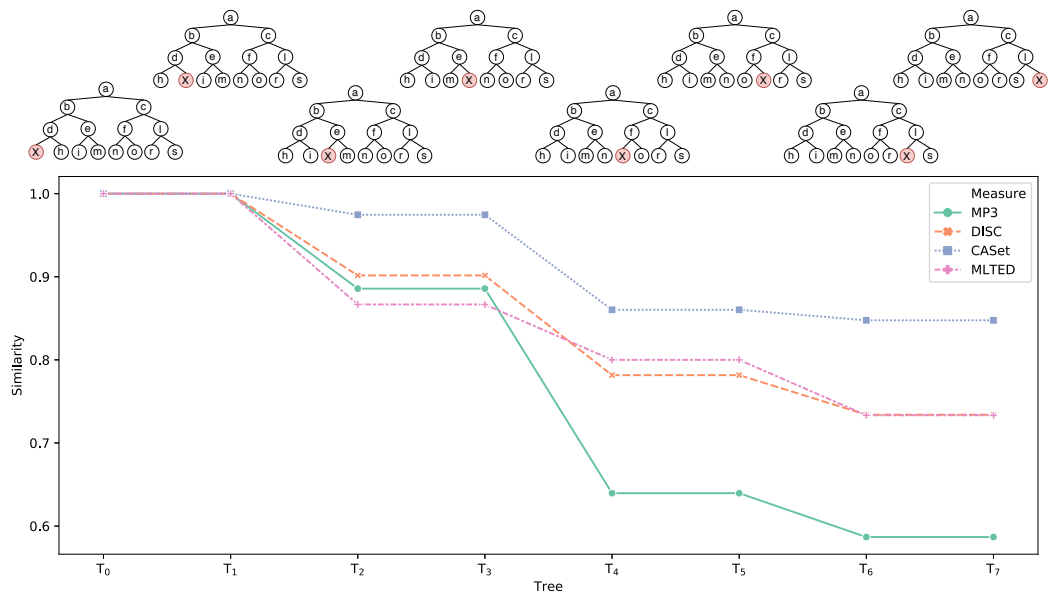


Figure S2: Effect of a label sliding from left to right on the lowest level of a binary tree.

#### 4 Base tree for poly-occurring label experiment

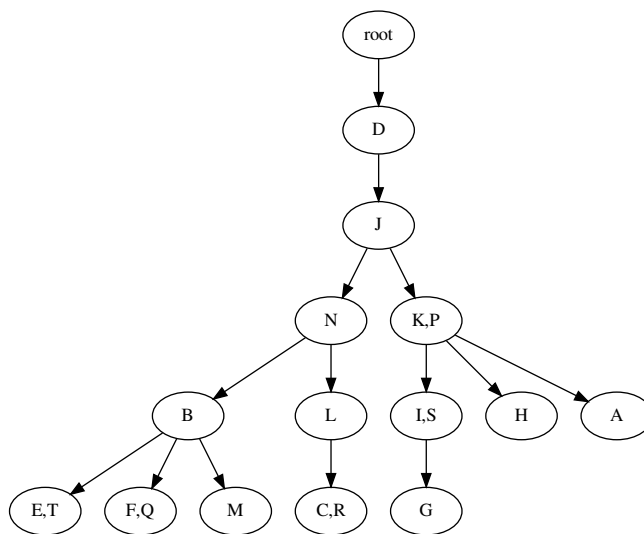


Figure S3: Base tree used for evaluating the effect of poly-occurring labels on the similarity scores.

## 5 Base trees for Exp1 and Exp2

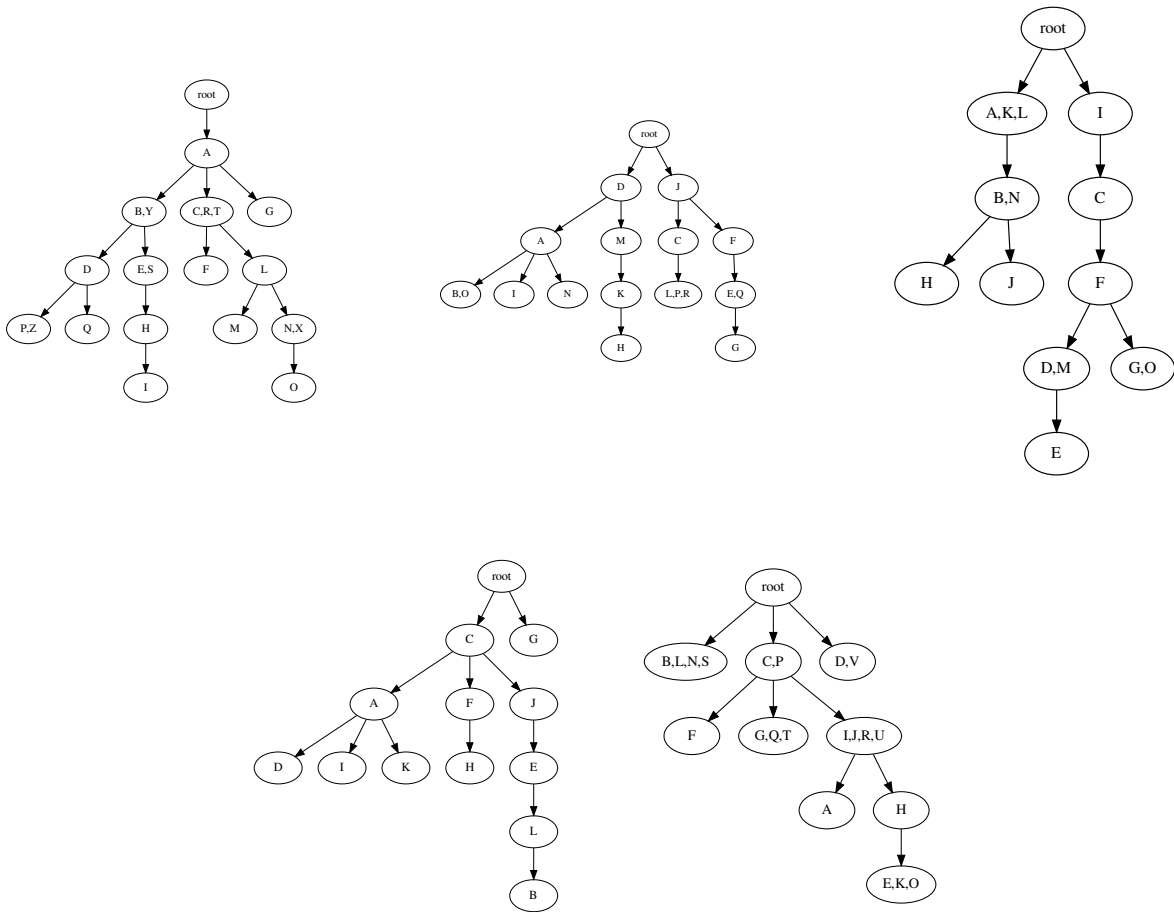


Figure S4: Base trees used in Experiments 1 and 2.

## 6 Base trees for Clustering experiment

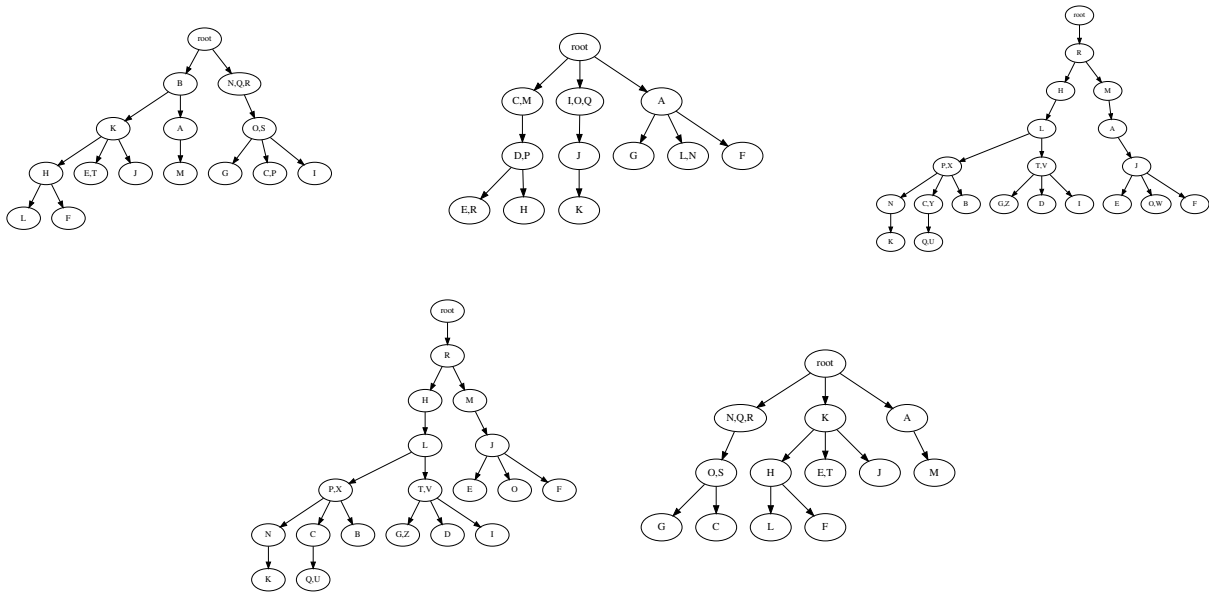


Figure S5: Base trees used in the clustering experiment. Tree 4 is a perturbation of Tree 3 and Tree 5 is a perturbation of Tree 1.

## 7 Trees for real data experiment

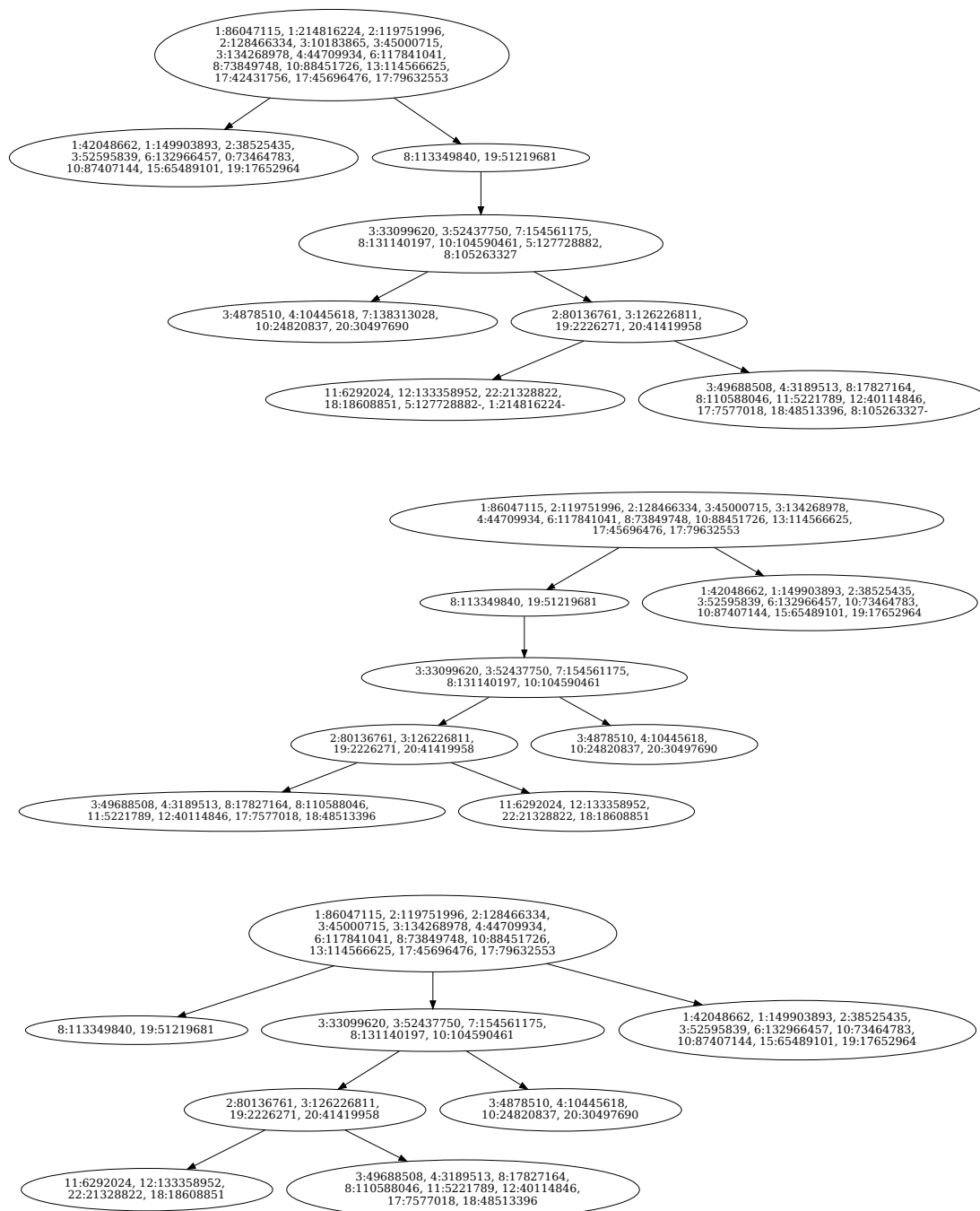


Figure S6: Trees used in the experiment on real data from Gerlinger et al. (2014). The upper tree is the base tree proposed in Gerlinger et al. (2014) for patient RMH002, the second one is the tree inferred by LICHeE, and the last one is the tree inferred by MIPUP (ipd parameter). We crafted the first tree starting from the supplementary material of the corresponding paper whereas we built the other two considering the trees reported in the MIPUP github repository.

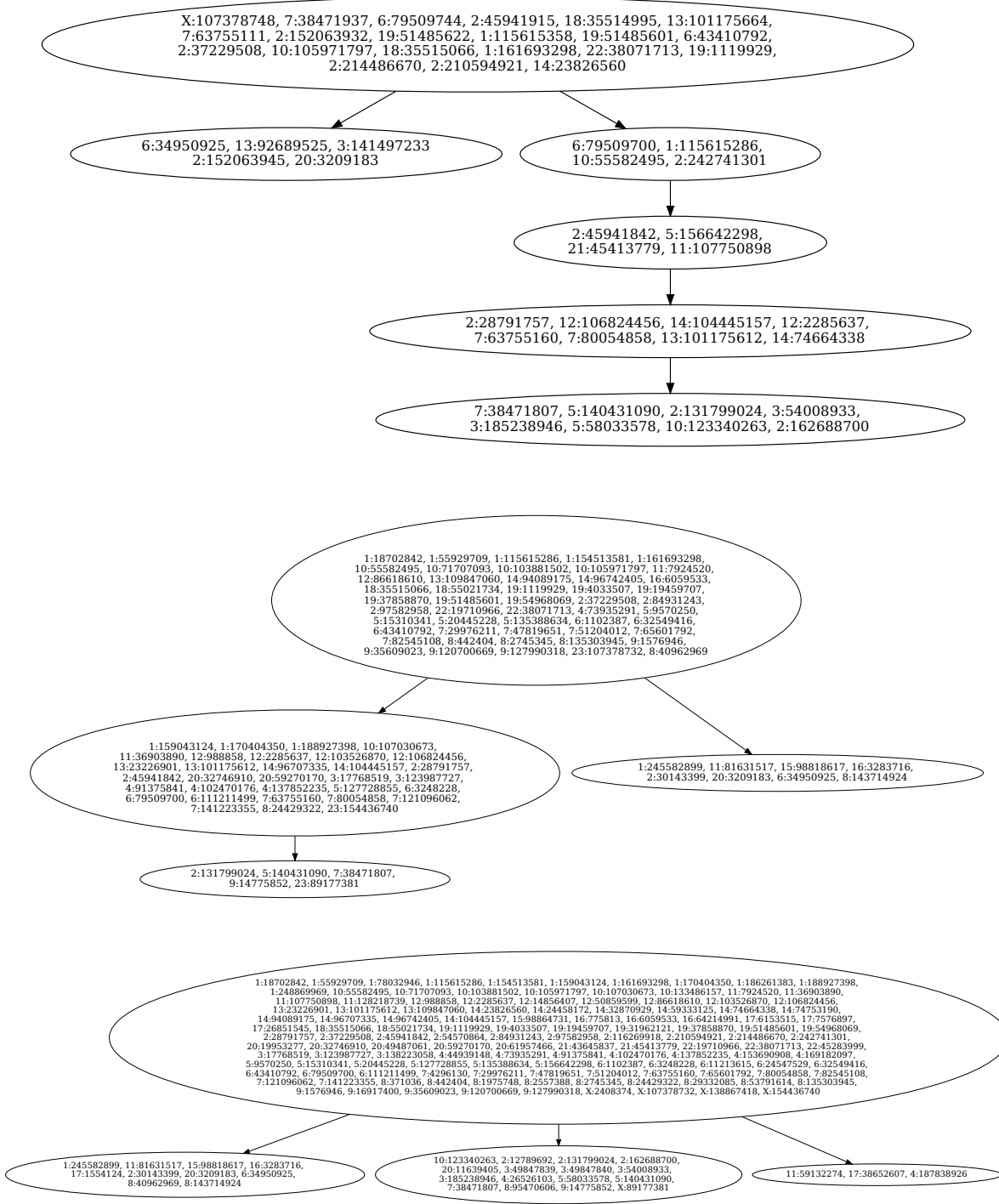
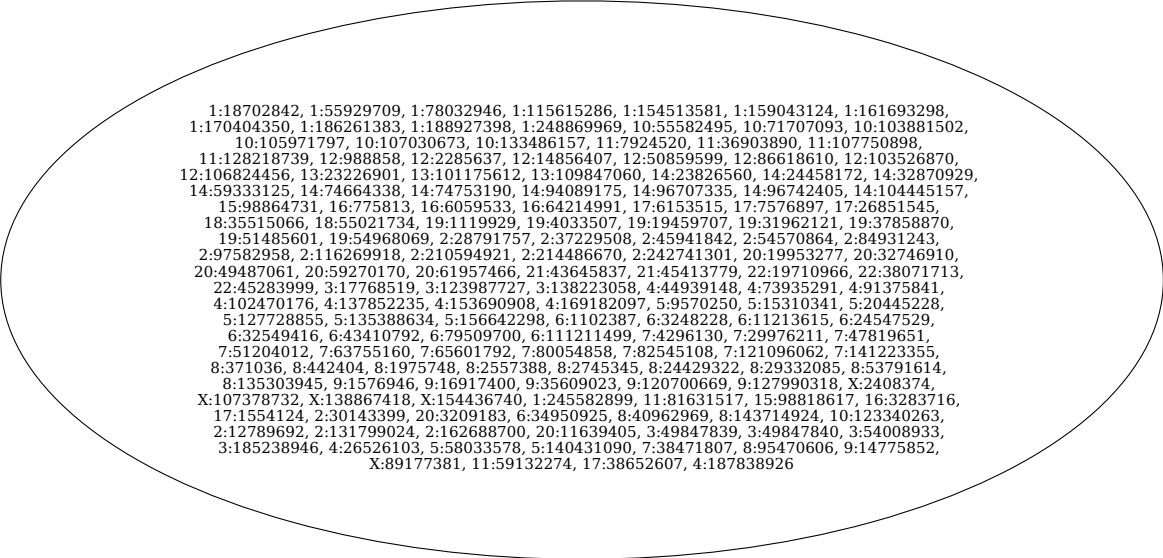


Figure S7: Trees used in the experiment on real data from Eirew et al. (2015). The upper tree is the base tree proposed in Eirew et al. (2015) for case SA501, the second one is the tree inferred by LICHeE, and the last one is the tree inferred by MIPUP. We obtained these trees from the supplementary material of DiNardo et al. (2019).



```

1:18702842, 1:55929709, 1:78032946, 1:115615286, 1:154513581, 1:159043124, 1:161693298,
1:170404350, 1:186261383, 1:188927398, 1:248869969, 10:55582495, 10:71707093, 10:103881502,
10:105971797, 10:107030673, 10:133486157, 11:7924520, 11:36903890, 11:107750898,
11:128218739, 12:988858, 12:2285637, 12:14856407, 12:50859599, 12:86618610, 12:103526870,
12:106824456, 13:23226901, 13:101175612, 13:109847060, 14:23826560, 14:24458172, 14:32870929,
14:59333125, 14:74664338, 14:74753190, 14:94089175, 14:96707335, 14:96742405, 14:104445157,
15:98864731, 16:775813, 16:6059533, 16:64214991, 17:6153515, 17:7576897, 17:26851545,
18:35515066, 18:55021734, 19:1119929, 19:4033507, 19:19459707, 19:31962121, 19:37858870,
19:51485601, 19:54968069, 2:28791757, 2:37229508, 2:45941842, 2:54570864, 2:84931243,
2:97582958, 2:116269918, 2:210594921, 2:214486670, 2:242741301, 20:19953277, 20:32746910,
20:49487061, 20:59270170, 20:61957466, 21:43645837, 21:45413779, 22:19710966, 22:38071713,
22:45283999, 3:17768519, 3:123987727, 3:138223058, 4:44939148, 4:73935291, 4:91375841,
4:102470176, 4:137852235, 4:153690908, 4:169182097, 5:9570250, 5:15310341, 5:20445228,
5:127728855, 5:135388634, 5:156642298, 6:1102387, 6:3248228, 6:11213615, 6:24547529,
6:32549416, 6:43410792, 6:79509700, 6:111211499, 7:4296130, 7:29976211, 7:47819651,
7:51204012, 7:63755160, 7:65601792, 7:80054858, 7:82545108, 7:121096062, 7:141223355,
8:371036, 8:442404, 8:1975748, 8:2557388, 8:2745345, 8:24429322, 8:29332085, 8:53791614,
8:135303945, 9:1576946, 9:16917400, 9:35609023, 9:120700669, 9:127990318, X:2408374,
X:107378732, X:138867418, X:154436740, 1:245582899, 11:81631517, 15:98818617, 16:3283716,
17:1554124, 2:30143399, 20:3209183, 6:34950925, 8:40962969, 8:143714924, 10:123340263,
2:12789692, 2:131799024, 2:162688700, 20:11639405, 3:49847839, 3:49847840, 3:54008933,
3:185238946, 4:26526103, 5:58033578, 5:140431090, 7:38471807, 8:95470606, 9:14775852,
X:89177381, 11:59132274, 17:38652607, 4:187838926

```

Figure S8: Edge case tree used in the experiment on real data from Eirew et al. (2015). We obtained such a tree by collapsing all nodes of the MIPUP’s tree from Figure S7 in a single node.

## 8 Example of computation of MP3

To better understand how to compute MP3, in this section we report a detailed example on trees with repeated labels.

In Figure S9 we present two trees: the tree on the left, that we will refer to as Tree A, is composed of six nodes and does not include any repeated label, whereas the tree on the right, that we will refer to as Tree B, is composed of six nodes and only five distinct labels.

In Figure S10 we report the 20 possible MTTs of Tree A and in Figure S11 we report the 20 possible MTTs of Tree B.

Note that the MTTs of Tree A are unique, whereas the MTTs of Tree B are not. Indeed, since the label  $c$  appears twice, two MTTs of Tree B have the same labels ( $a$ ,  $c$ , and  $f$ ) and, by chance, the same shape (top right of Figure S11). Notice that the same MTT is in the multiset of MTTs of Tree A (top right of Figure S10) but appears only once.

In Figure S12 we report the set of MTTs shared between Tree A and Tree B. This set is composed of seven MTTs and note that the MTT of the triplet  $(a, c, f)$  appears only once. This means that the MTT of that triplet will appear twice in  $M_B(a, c, f)$ , once in the  $M_A(a, c, f) \cap M_B(a, c, f)$ , and once  $M_B(a, c, f) \setminus M_A(a, c, f)$ .

Overall, we have the following (we refer the reader to Section 2.2 of the main document for the notation):

- $|\lambda(A)| = 6$
- $|\lambda(B)| = 5$
- $|I| = 10$ , *i.e.*, the number of the triplets over the multiset  $\{a, b, c, e, f\}$
- $|J| = 35$ , *i.e.*, the number of the triplets over the multiset  $\{a, b, c, c, d, e, f\}$



- $MP3_{\cap} = 7/10 = 0.7$
- $MP3_{\cup} = 7/35 = 0.2$
- $MP3_{\sigma} = 0.2 + \sigma(0.7) \times \min\{0.7 - 0.2, 0.2\} \simeq 0.22384$

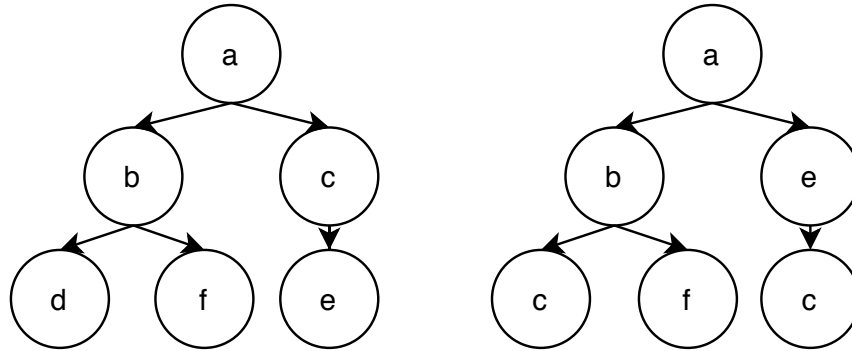


Figure S9: Two tumor progression trees. Tree A (left) is composed of six nodes and six different labels, whereas Tree B (right) is composed of 6 nodes and five distinct labels.

## References

- Zach DiNardo, Kiran Tomlinson, Anna Ritz, and Layla Oesper. Distance measures for tumor evolutionary trees. *Bioinformatics*, 11 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz869. URL <https://doi.org/10.1093/bioinformatics/btz869>. btz869.
- Peter Eirew, Adi Steif, Jaswinder Khattra, Gavin Ha, Damian Yap, Hossein Farahani, Karen Gelmon, Stephen Chia, Colin Mar, Adrian Wan, et al. Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature*, 518(7539):422–426, 2015.
- Marco Gerlinger, Stuart Horswell, James Larkin, Andrew J Rowan, Max P Salm, Ignacio Varela, Rosalie Fisher, Nicholas McGranahan, Nicholas Matthews, Claudio R Santos, et al. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nature genetics*, 46(3):225, 2014.

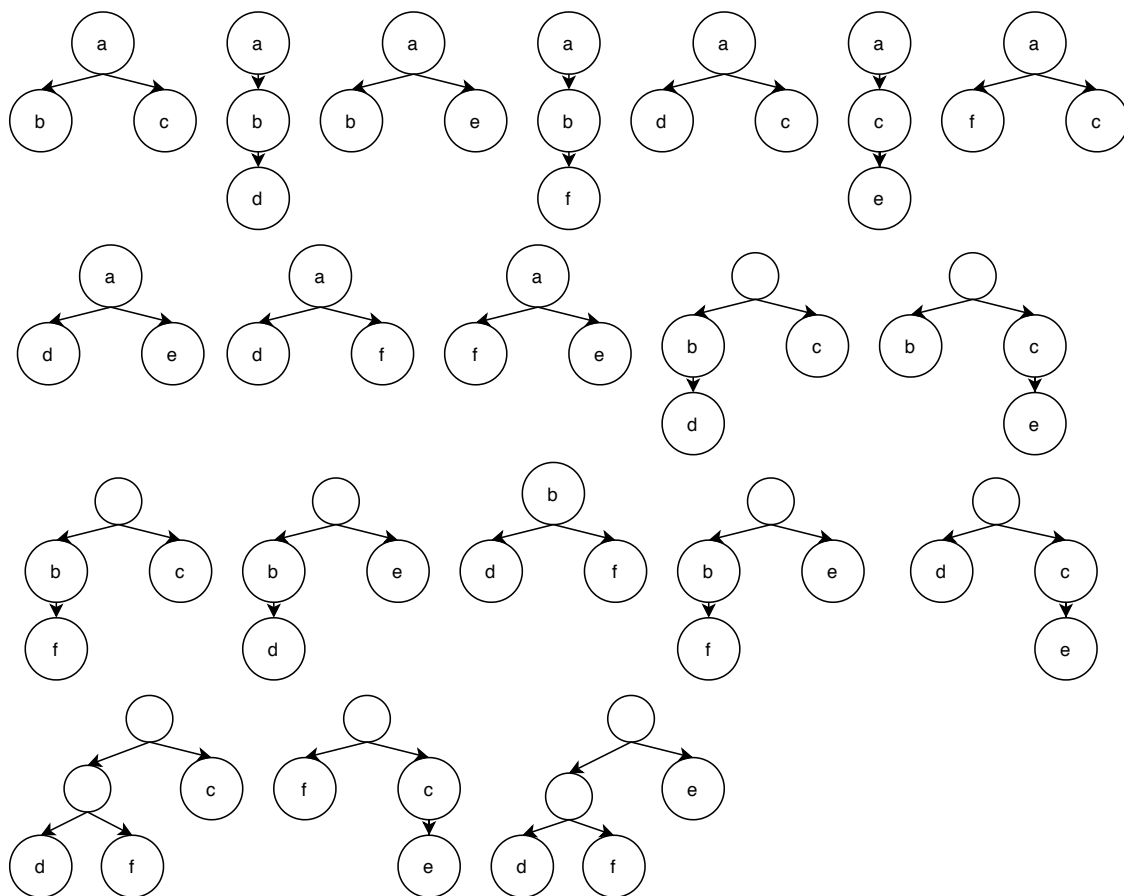


Figure S10: Minimal tree topology of each triplet of nodes in Tree A.

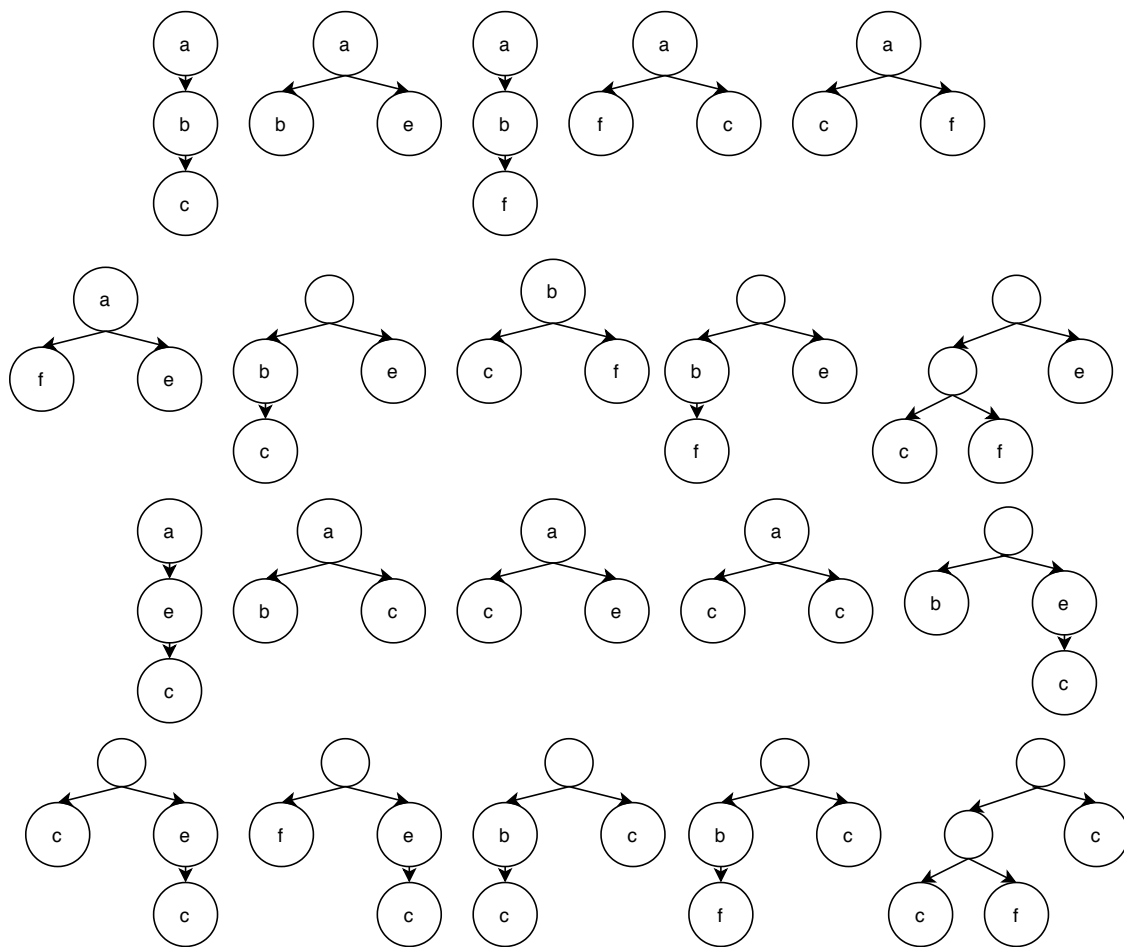


Figure S11: Minimal tree topology of each triplet of nodes in Tree B.

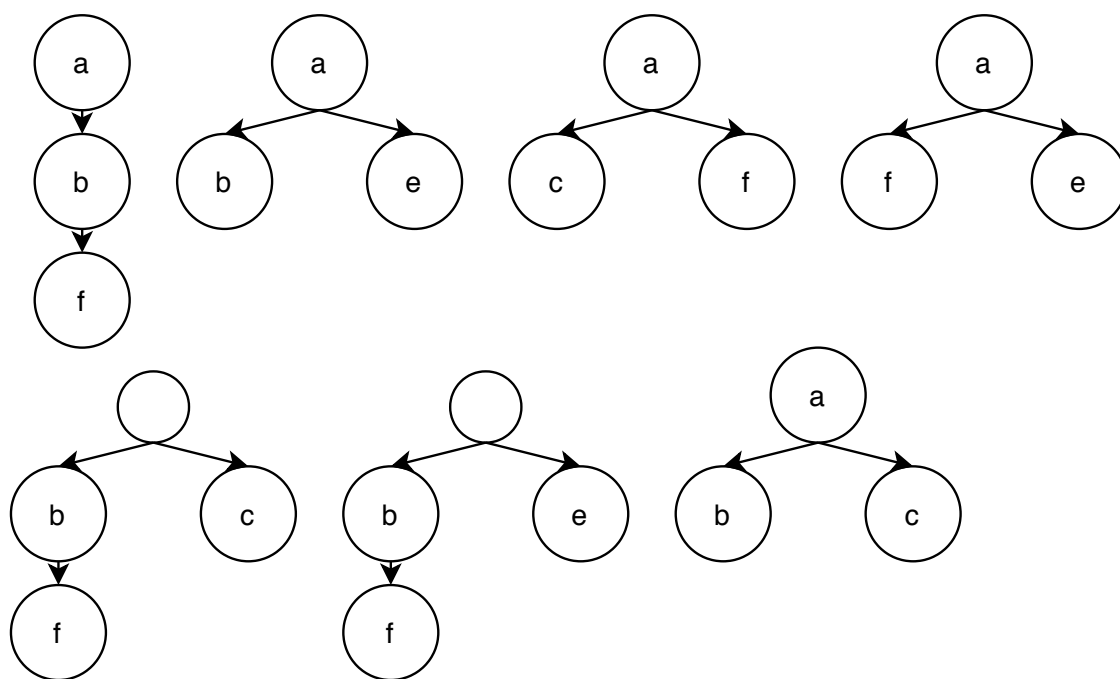


Figure S12: Shared Minimal tree topologies shared between Tree A and Tree B. Note that the MTT of the triplet  $(a, c, f)$  appears only once.