Journal of Mathematical Biology manuscript No.
(will be inserted by the editor)

Gabriel Cardona · Mercè Llabrés · Francesc Rosselló · Gabriel Valiente

Nodal distances for rooted phylogenetic trees (Supplementary material)

Received: / Revised version:

1. Distributions

Figure 1 shows the distributions of d_1^s and $(d_2^s)^2$ (that is, of d_2^s squared) on $\mathcal{NT}_n$ for $n = 3, 4, 5, 6$; with more than two million trees in $\mathcal{NT}_7$, the computation of these distributions for values of n greater than 6 is beyond our computational power. In each such graphic, the columns represent the percentage of pairs of different trees having the distance corresponding to the column. We use the square of d_2^s instead of the metric without squaring in order to obtain nonnegative integer distance values. The files `{3,4,5,6}-tree-nt-1-stat.dat` and `{3,4,5,6}-tree-nt-2-stat.dat`, available at the Supplementary Material web page, contain, respectively, the numbers of pairs of trees in $\mathcal{NT}_n$, for $n = 3, 4, 5, 6$, at each of the possible distances d_1^s and d_2^s (squared).

We have also studied the distribution of the metrics d_1^s and d_2^s (squared) on TreeBASE version 6.29, which contains 2,592 phylogenies with over 36,000 taxa among them. In this study, we first discarded the 18 entries that are not phylogenetic trees (because they contain repeated node labels) and then, applied the metrics to the topological restriction of each pair of the remaining 2,574 trees to their common node labels.

Phylogenetic trees with nested taxa can share node labels in both leaves and internal nodes and thus, the topological restriction of two trees to their common node labels was obtained by first removing any node labels not

Gabriel Cardona: Department of Mathematics and Computer Science, University of the Balearic Islands, E-07122 Palma de Mallorca, Spain `gabriel.cardona@uib.es`

Mercè Llabrés, Francesc Rosselló: Department of Mathematics and Computer Science, University of the Balearic Islands, E-07122 Palma de Mallorca, Spain and Research Institute of Health Science (IUNICS), E-07122 Palma de Mallorca, Spain, {`merce.llabres,cesc.rossello`}`@uib.es`

Gabriel Valiente: Algorithms, Bioinformatics, Complexity and Formal Methods Research Group, Technical University of Catalonia, E-08034 Barcelona, Spain and Research Institute of Health Science (IUNICS), E-07122 Palma de Mallorca, Spain, `valiente@lsi.upc.edu`

Key words: Key words should be given

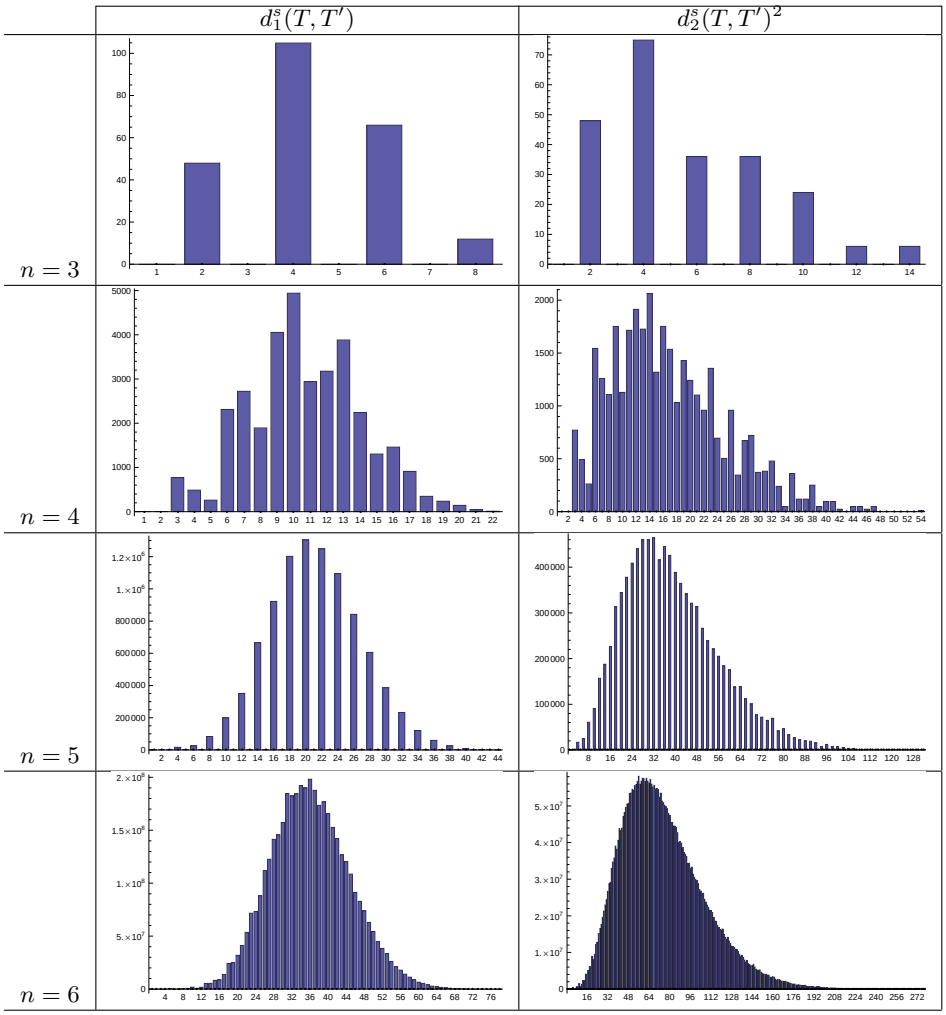

Fig. 1. Distributions of d_1^s and $(d_2^s)^2$ on $\mathcal{NT}_n$ for $n = 3, 4, 5, 6$.

shared by the two trees, then removing all unlabeled leaves in postorder and finally, contracting to an arc any elementary path with unlabeled intermediate nodes. Fig. 2 displays the distributions of the values of d_1^s and $(d_2^s)^2$ applied to pairs of trees in TreeBASE sharing $n = 2$ to 6 labels. The files {2,3,4,5,6}`-tree-nt-1-tb-stat.dat` and {2,3,4,5,6}`-tree-nt-2-tb-stat.dat`, available at the Supplementary Material web page, contain, respectively, the numbers of pairs of trees in TreeBASE sharing $n = 2, 3, 4, 5, 6$, at each of the possible distances d_1^s and d_2^s (squared).

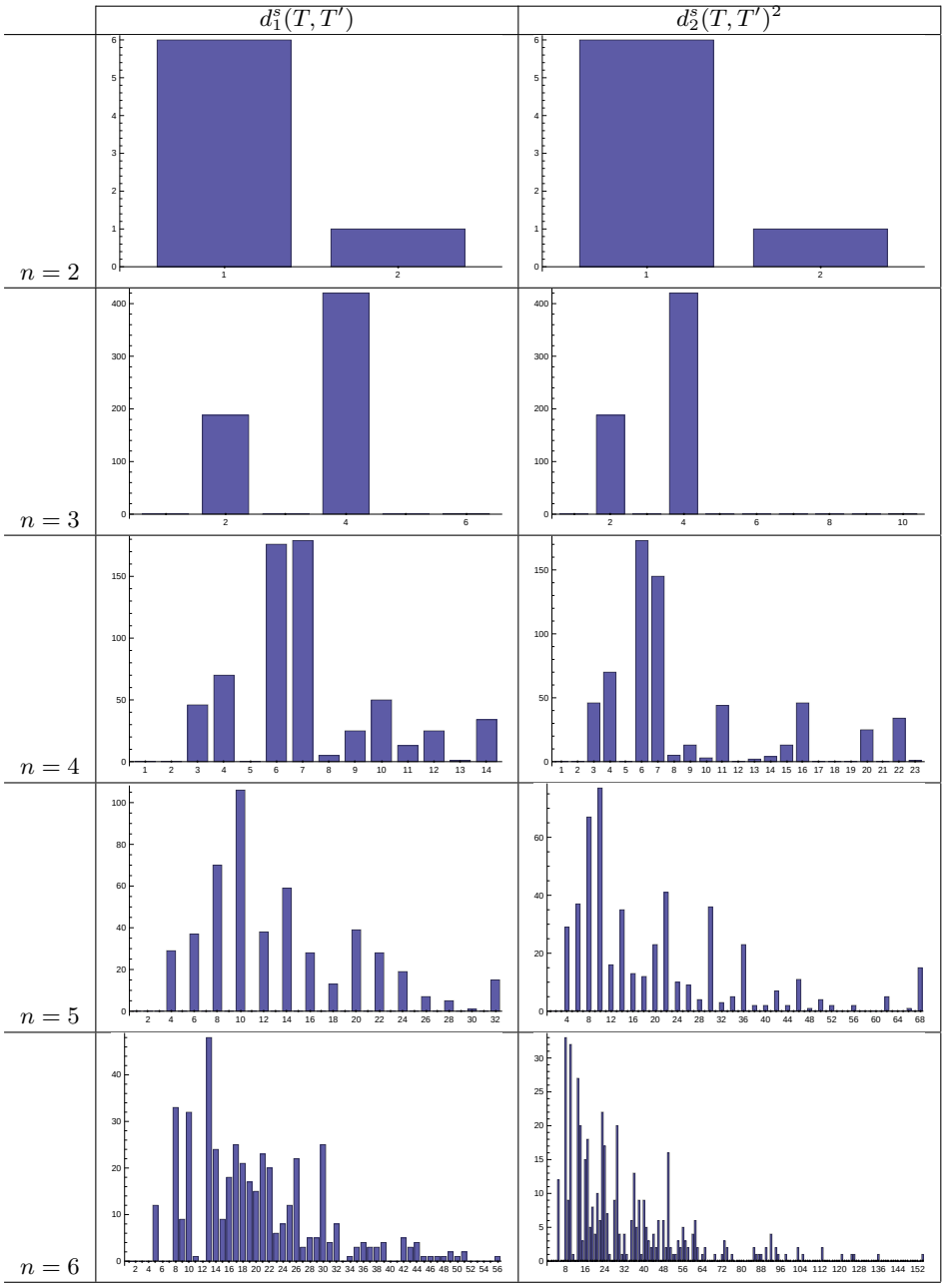

Fig. 2. Distributions of d_1^s and $(d_2^s)^2$ for pairs of trees in TreeBASE sharing $n = 2, \ldots, 6$ labels.