



HAL
open science

Tanglegrams are misleading for visual evaluation of tree congruence

Damien M. de Vienne

► **To cite this version:**

Damien M. de Vienne. Tanglegrams are misleading for visual evaluation of tree congruence. 2018. hal-01807411

HAL Id: hal-01807411

<https://hal.science/hal-01807411>

Preprint submitted on 4 Jun 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Tanglegrams are misleading for visual evaluation of tree congruence

Damien M. de Vienne^{1,*}

¹Univ Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR5558, F-69622 Villeurbanne, France

*Corresponding author: Damien M. de Vienne damien.de-vienne@univ-lyon1.fr

Abstract. Evolutionary Biologists are often faced with the need to compare phylogenetic trees. One popular method consists in visualizing the trees face to face with links connecting matching taxa. These *tanglegrams* are optimized beforehand so that the number of lines crossing (the entanglement) is minimal. This representation is implicitly justified by the expectation that the level of entanglement is correlated with the level of similarity (or congruence) between the trees compared. Using simulations, we show that this correlation is actually very weak, which should preclude the use of such technique for getting insight into the level of congruence between trees.

Tanglegram is a classical way of graphically representing correspondence between two labeled phylogenies, two hierarchical clusters, or any pair of tree-like structures. It is widely used in the host-parasite literature where host and parasite trees are represented face to face and links between leaves depict their interactions. The idea behind this visualization is straightforward: after having found the graphical layout that produces the minimum crossings of the lines linking both trees, the obtained tanglegram is expected to help in more clearly visualizing co-evolutionary relationships between species (Matsen et al. 2016).

Numerous studies focused on finding the best algorithm for reordering the leaves in order to produce the less entangled representation (Bansal et al. 2009; Böcker et al. 2009; Fernau, Kaufmann, Poths 2010; Venkatachalam et al. 2010 and references therein; Scornavacca, Zickmann, Huson 2011). But none ever questioned the real utility of such representation for getting insight into the level of topological similarity (or congruence) between the compared trees. There seems to be a common belief that the level of entanglement between two trees (the number of lines crossing) gives an approximation of their level of congruence, justifying the use of tanglegrams as a visual tool for “comparing” trees. However, this has never been explored. Planet (2006) clearly formulated this question: “It is not clear if more tangle actually translates to more incongruence.” Here we address this question on simulated data and evaluate the validity of tanglegrams for getting insight into the level of (dis)similarity between trees.

Pairs of random tree topologies with 20 leaves were generated in R using the *rtree* function of the *ape* package (Paradis, Claude, Strimmer 2004), and sorted according to their topological distance (using the commonly used RF distance, Robinson, Foulds 1981). 1000 pairs of trees were kept for each possible value of RF and an optimized tanglegram (minimizing the number of lines crossing) was computed for each pair with the R package *dendextend* (Galili 2015). The optimization used, described in Dwyer and Schreiber (2004),

consisted in solving repeatedly the One Tree Crossing Minimization problem (see Fernau, Kaufmann, Poths 2010), each time switching the roles of the two trees.

Figure 1A represents the observed level of entanglement between the trees (measured by the number of lines crossing in the tanglegram) as a function of their topological distance. While there is a tendency for increased entanglement at large topological distances, there is also a very strong overlap between the distributions of entanglement at different topological distances (Figure 1B). The same observations were made when using a different measure of congruence between the trees (the Icong index, de Vienne, Giraud, Martin 2007; data not shown). This shows that the level of entanglement is a poor estimate of the level of congruence between the trees compared. As an illustration, using the number of lines crossing in the tanglegram as a surrogate for testing if a pair of trees is more congruent than another one would give the wrong answer in almost one out of four cases (22.6%) when sampling from the population of tree pairs displayed in Figure 1.

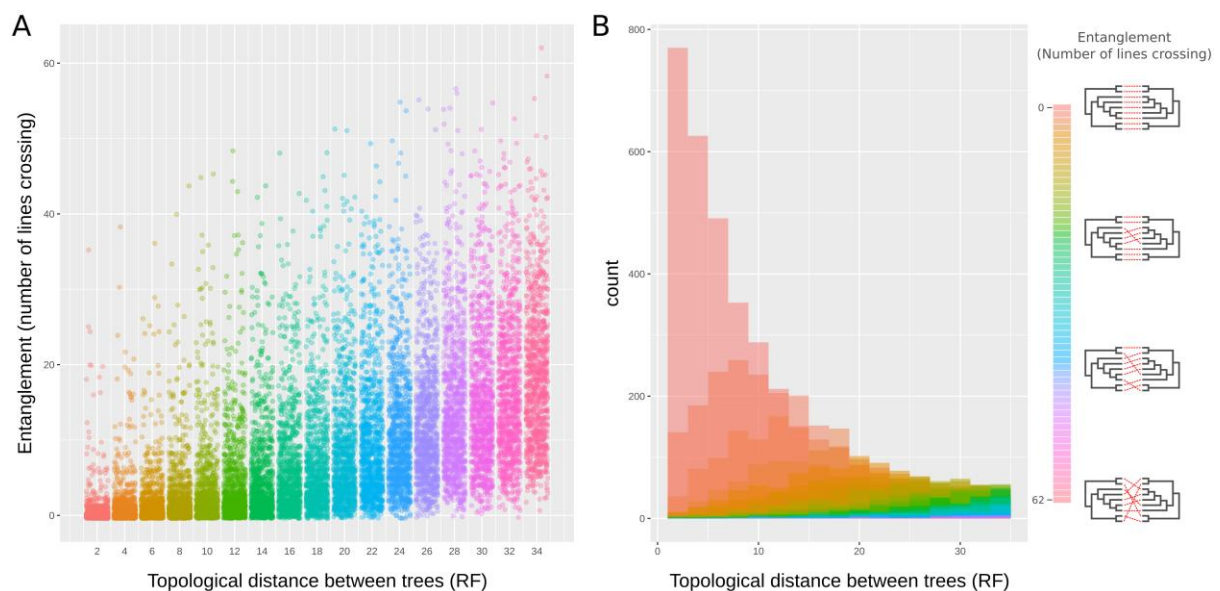


Figure 1. **A.** Relationship between the topological distance (RF) between trees (x-axis) and the level of entanglement obtained after optimization (y-axis). Each dot represents one pair of trees, and dots are grouped and colored by their RF distance. 1000 pairs were generated for each possible RF distance. **B.** Distributions of RF distances for increasing entanglement (from 0 — no tangle — to 62).

This is however not surprising *a posteriori*: two trees can have very different topologies but show no tangle at all in the tanglegram. In fact, the number of such cases for a tree with $N+1$ leaves is the number of unlabeled rooted binary trees with N internal nodes (which is the Catalan number C_N , minus 1 to remove self comparison). So for a 20-leaf tree, there are 1,767,263,189 topologically distinct trees for which the tanglegram would show no crossing!

As far as we know, this is the first analysis that explores the link between the level of entanglement in a tanglegram and the level of congruence between trees. Given the very weak predictive power of the former on the latter, we advocate that no conclusion should be drawn from the visual inspection of tanglegrams on the overall similarity between the trees analyzed. This could question the very use of tanglegrams, notably in the host-parasite literature.

Bibliography

- Bansal, MS, W Chang, O Eulenstein, D Fernández-Baca. 2009. Generalized Binary Tanglegrams: Algorithms and Applications. In: S Rajasekaran, editor. *Bioinformatics and Computational Biology*. Berlin, Heidelberg: Springer.
- Böcker, S, F Hüffner, A Truss, M Wahlström. 2009. A faster fixed-parameter approach to drawing binary tanglegrams. 4th International Workshop of Parameterized and Exact Computation: Springer. p. 38-49.
- de Vienne, DM, T Giraud, OC Martin. 2007. A congruence index for testing topological similarity between trees. *Bioinformatics* 23:3119-3124.
- Dwyer, T, F Schreiber. 2004. Optimal leaf ordering for two and a half dimensional phylogenetic tree visualisation. *Australasian Symposium on Information Visualisation*. Christchurch, New Zealand. p. 109-115.
- Fernau, H, M Kaufmann, M Poths. 2010. Comparing trees via crossing minimization. *Journal of Computer and System Sciences* 76:593-608.
- Galili, T. 2015. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* 31:3718-3720.
- Matsen, F, S Billey, A Kas, M Konvalinka. 2016. Tanglegrams: A Reduction Tool for Mathematical Phylogenetics. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.
- Paradis, E, J Claude, K Strimmer. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20:289-290.
- Planet, PJ. 2006. Tree disagreement: Measuring and testing incongruence in phylogenies. *Journal of Biomedical Informatics* 39:86-102.
- Robinson, DF, LR Foulds. 1981. Comparison of Phylogenetic Trees. *Mathematical Biosciences* 53:131-147.
- Scornavacca, C, F Zickmann, DH Huson. 2011. Tanglegrams for rooted phylogenetic trees and networks. *Bioinformatics* 27:i248-256.
- Venkatachalam, B, J Apple, K St John, D Gusfield. 2010. Untangling Tanglegrams: Comparing Trees by Their Drawings. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 7:588-597.