

Exploring and visualising spaces of tree reconciliations: Supplementary material II (Computational results)

K. T. Huber, e-mail: k.huber@uea.ac.uk and
V. Moulton, e-mail: v.moulton@uea.ac.uk and
Marie-France Sagot, e-mail: marie-france.sagot@inria.fr and
Blerina Sinaimeri, e-mail: blerina.sinaimeri@inria.fr

July 16, 2018

1 Datasets

As described in the paper, to evaluate how our metrics performed for different biological datasets, we extracted 8 pairs of host/parasite trees from the literature for the datasets in the following table. We include the host/parasite trees that we used for each dataset below in the form of a tanglegram [1].

Dataset name	Organisms involved	Reference	Type of dataset	$ L(H) $	$ L(P) $
FE	<i>Formicidae & Eucharitidae</i>	[3]	(a)	4	5
CP	<i>Cichlidae & Platyhelminthes</i>	[4]	(b)	6	29
PMP	<i>Pelican & Lice</i>	[5]	(c)	18	18
RH	<i>Rodents & Hantaviruses</i>	[6]	(d)	34	42
EC	<i>Encyrtidae & Coccidae</i>	[2]	(a)	7	10
GL	<i>Gopher & Lice</i>	[7]	(a)	8	10
SC	<i>Seabirds & Chewing Lice</i>	[8]	(c)	11	14
SFC	<i>Smut Fungi & Caryophyllaceus plants</i>	[9]	(c)	15	16

Table 1: For each dataset used we present its name and reference, its type, number of leaves in the host tree, number of leaves in the parasite tree and supplementary material if any. All datasets involve parasitism.

In what follows we also present the distributions, correlations and multidimensional scaling (MDS) plots for the normalised distances that we obtained for each of our datasets. In the case of correlations we present the Pearson Correlation coefficient at the top of each correlation panel. The cost of the optimal reconciliation for each dataset and each cost vector is given in the following table. As described in the paper, blue in an MDS plot represents low cost and red high cost.

Cost vector	FE	CP	PMP	RH	EC	GL	SC	SFC
$c_c = 0, c_d = c_l = c_s = 1$	4	21	8	26	6	4	6	11
$c_c = 0, c_d = c_l = 1, c_s = 2$	7	31	14	43	10	7	10	21

Table 2: For each dataset and given costs, the cost of the optimal reconciliation is presented.

2 Host and parasite trees

In this section, we present the host trees and parasite trees that we obtained for the datasets described in Section 1 in the form of a tanglegram. In all cases, the host tree is depicted in blue and the parasite tree in brown.

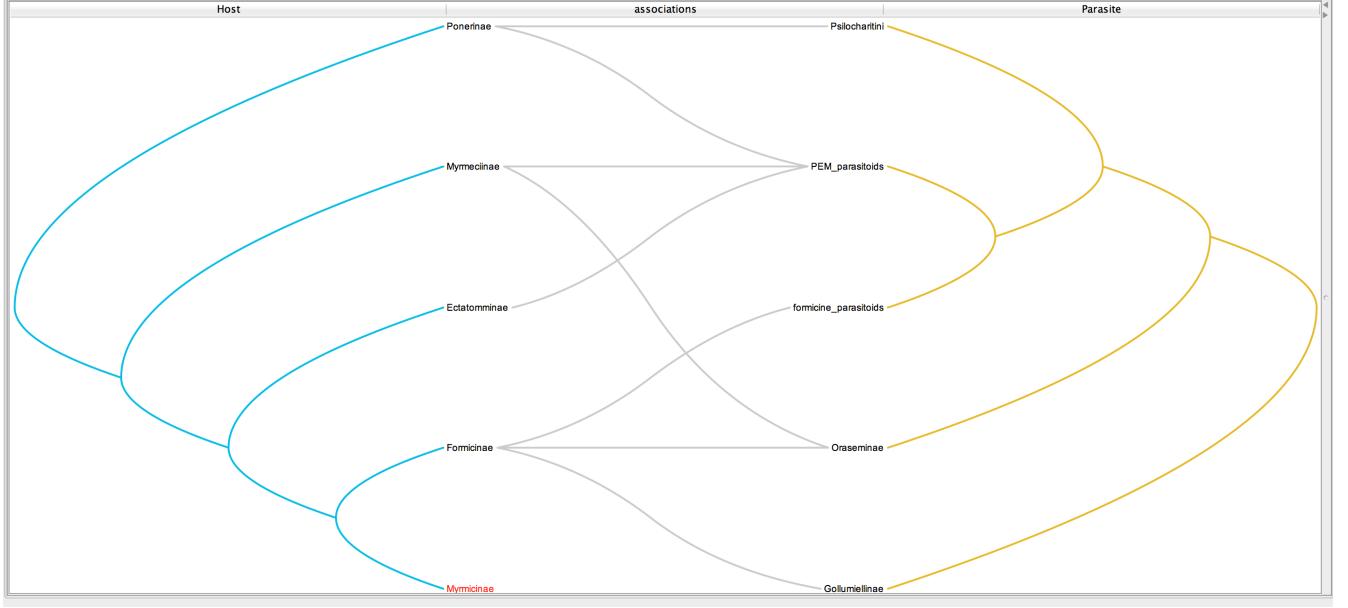


Figure 1: The tanglegram for the FE dataset.

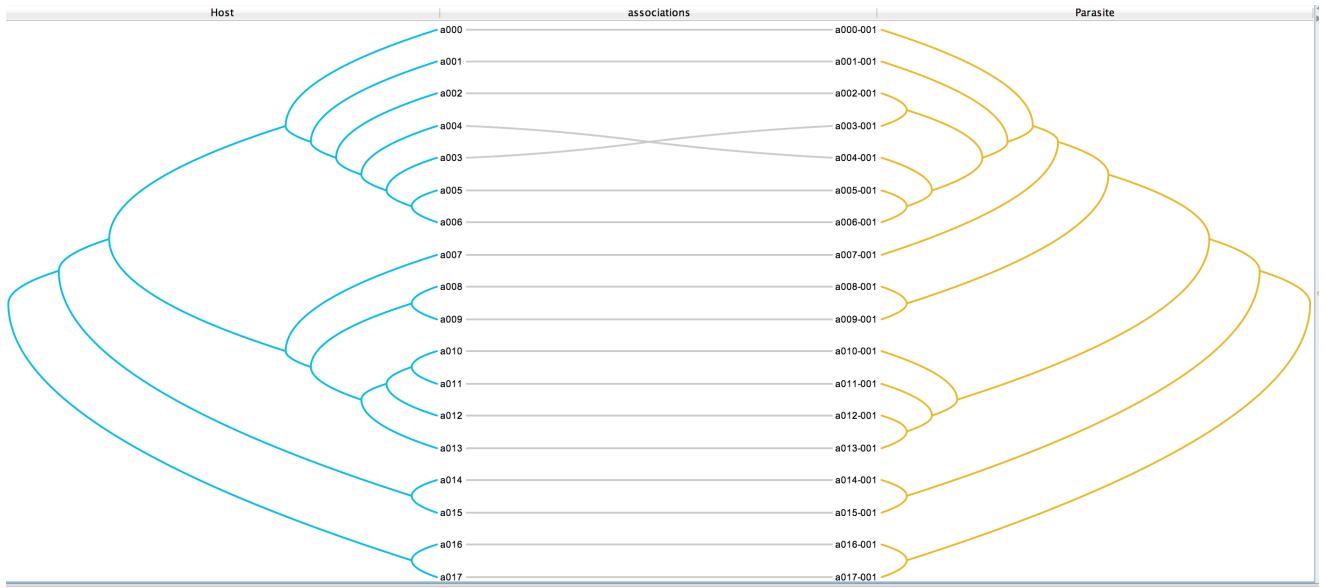


Figure 2: The tanglegram for the PMP dataset.

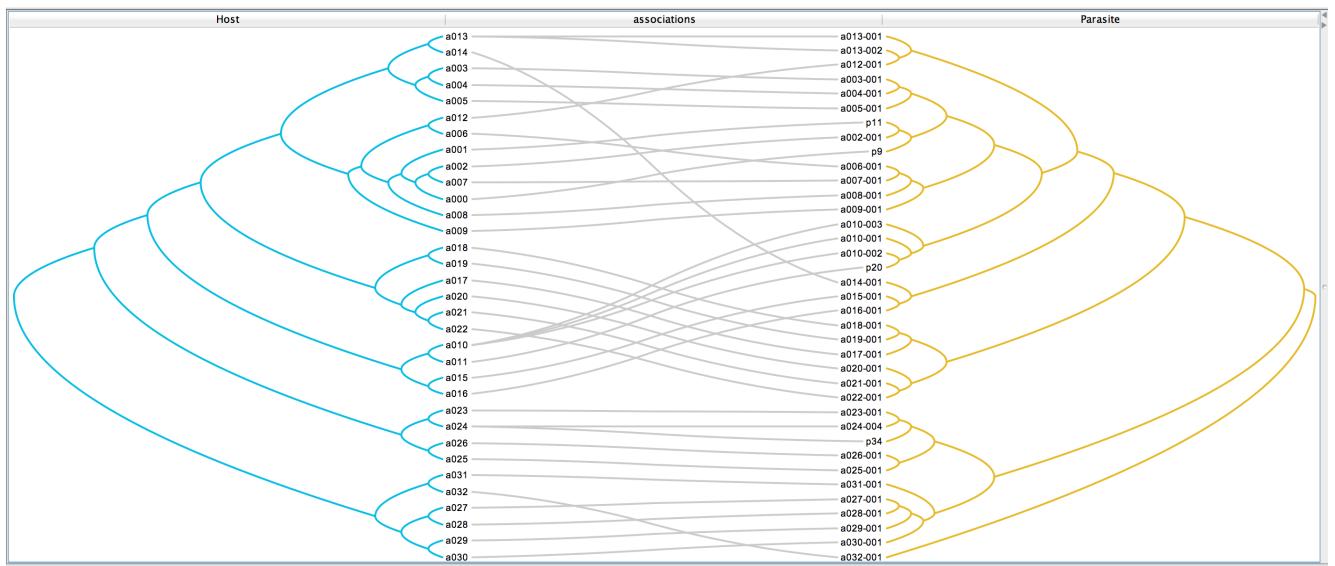


Figure 3: The tanglegram for the RH dataset.

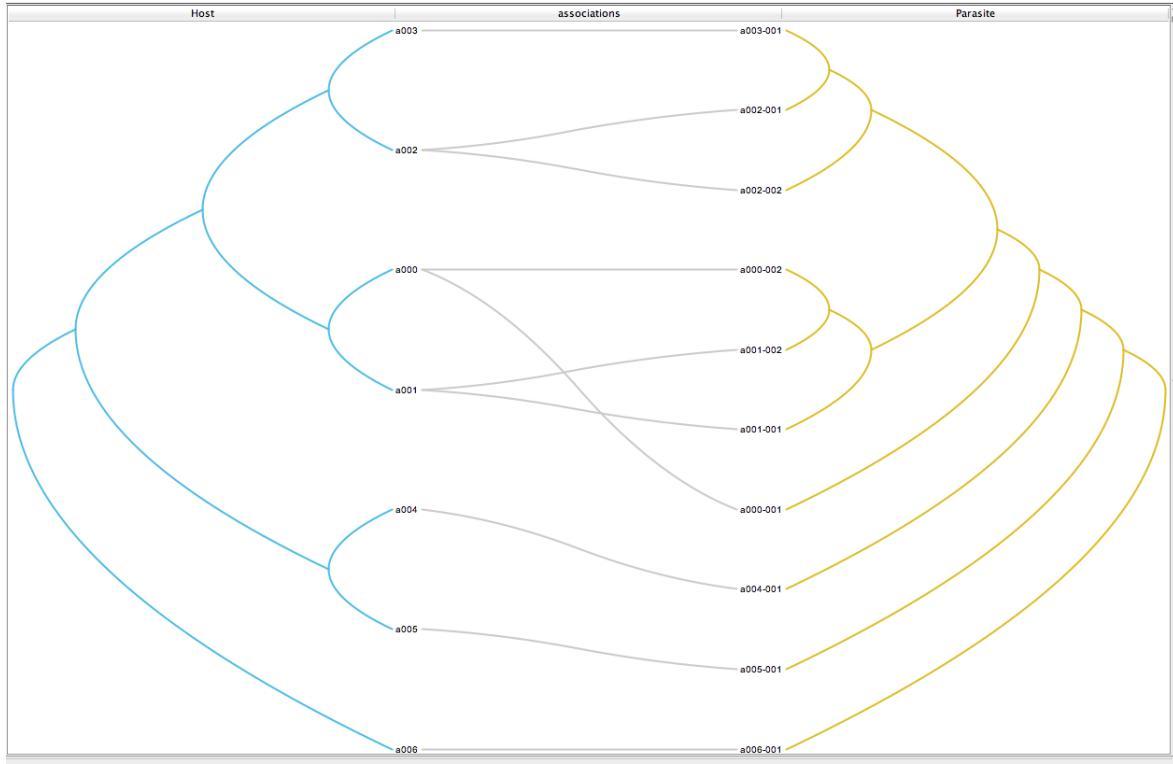


Figure 4: The tanglegram for the host tree (blue) and the parasite tree (brown) for the EC dataset.

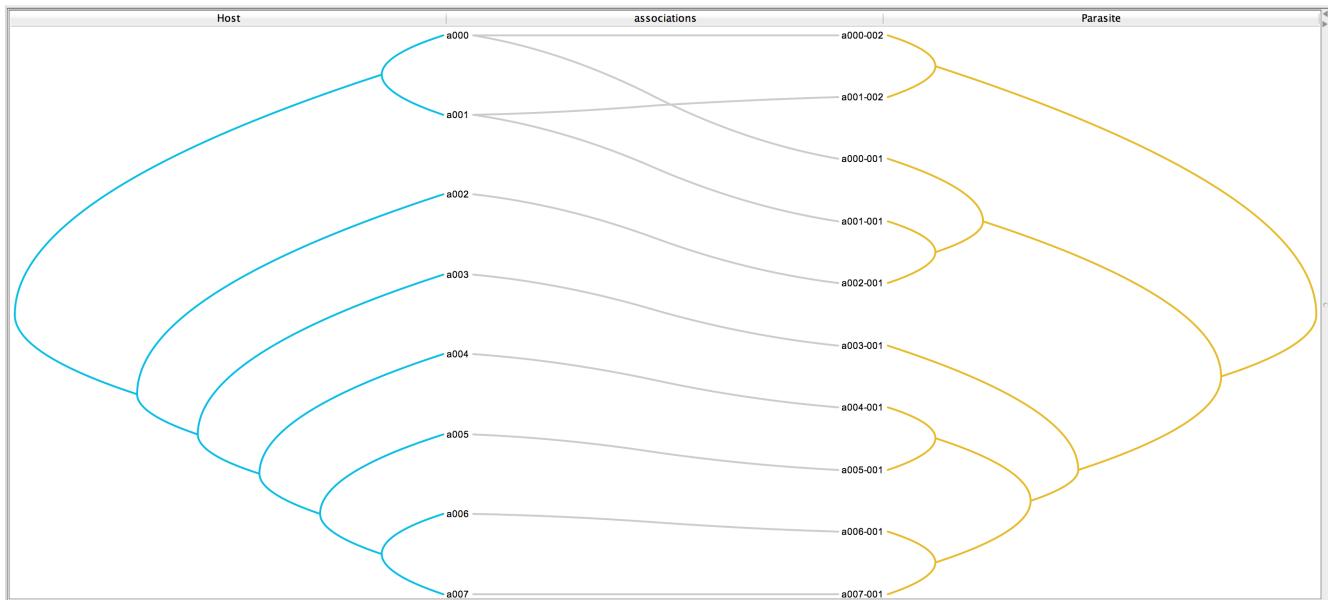


Figure 5: The tanglegram for the GL dataset.

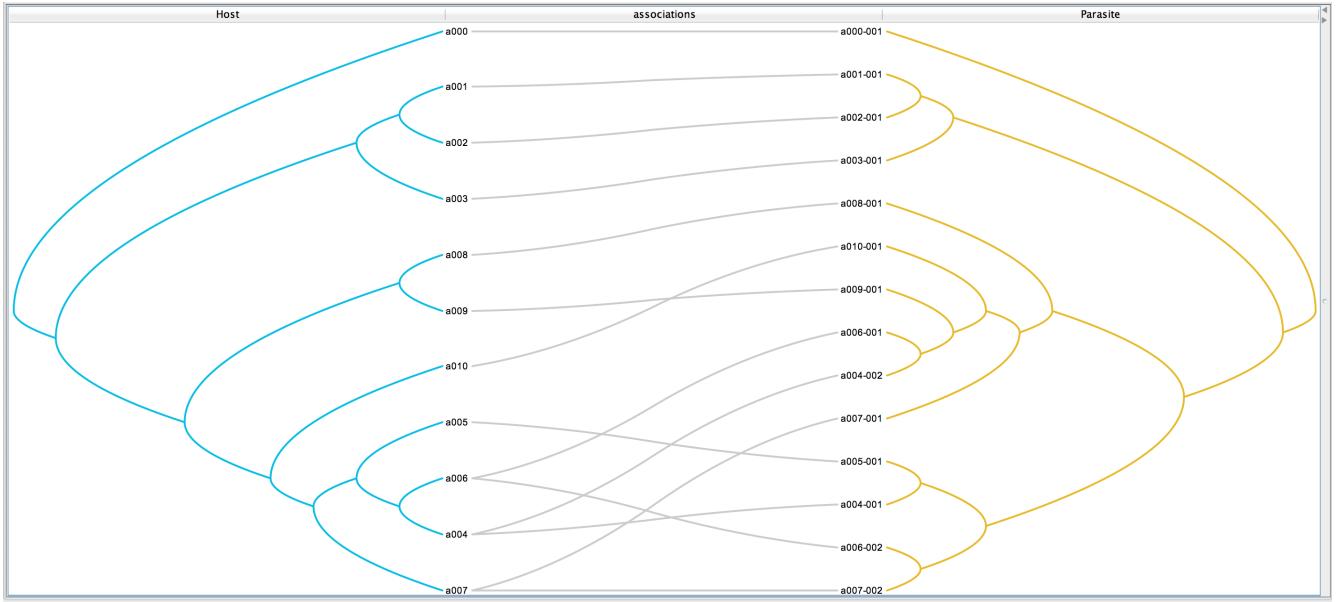


Figure 6: The tanglegram for the SC dataset.

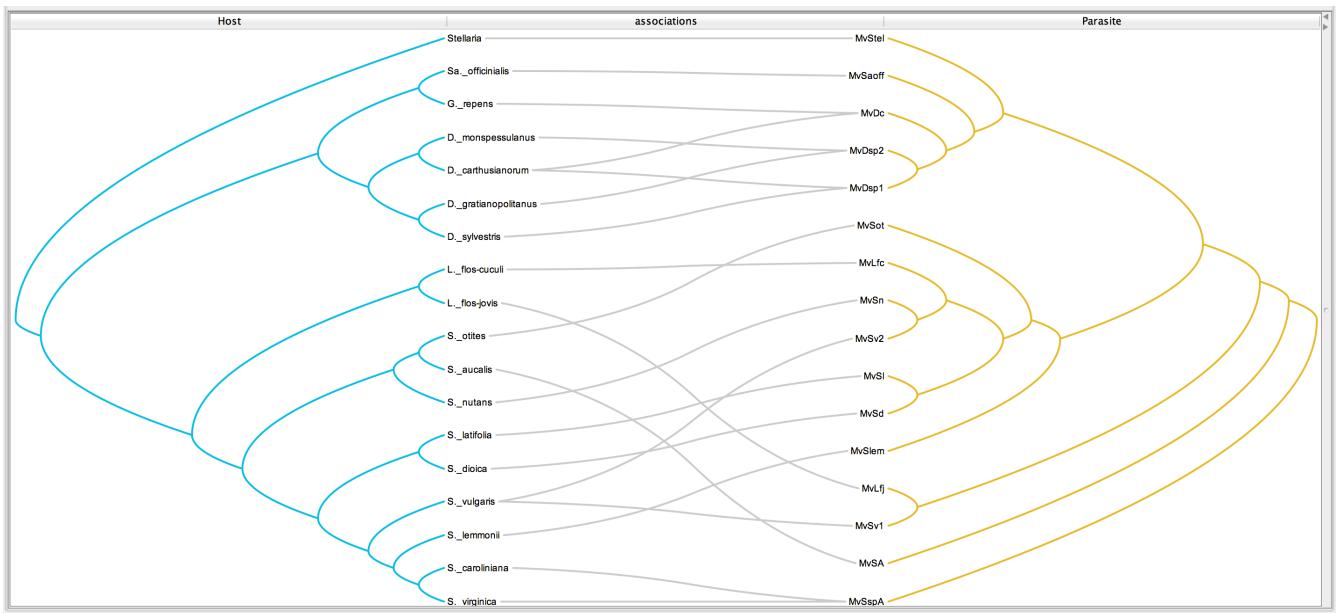


Figure 7: The tanglegram for the SFC dataset.

3 The FE dataset

The following results are based on all 215 reconciliations between the host tree and the parasite tree.

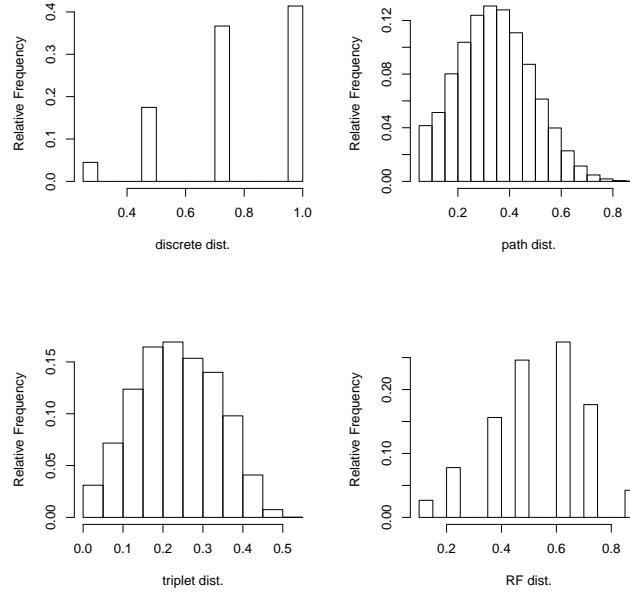


Figure 8: The relative frequency distributions of the (normalized) pairwise distance values between reconciliations.

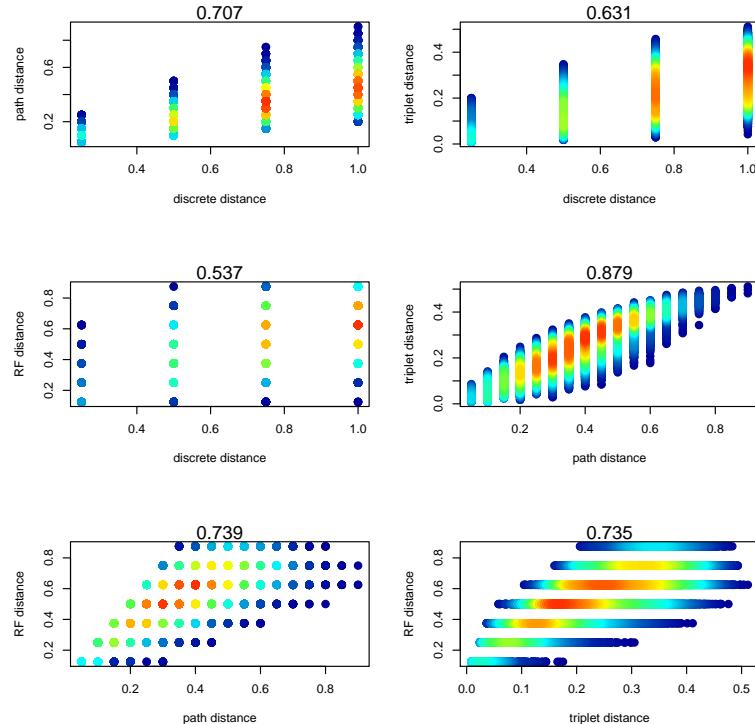


Figure 9: Correlations between the discrete/triplet/path/edit-distances. The Pearson correlation coefficient is above each panel.

The colour scale to the right of each panel corresponds to the reconciliation scores.

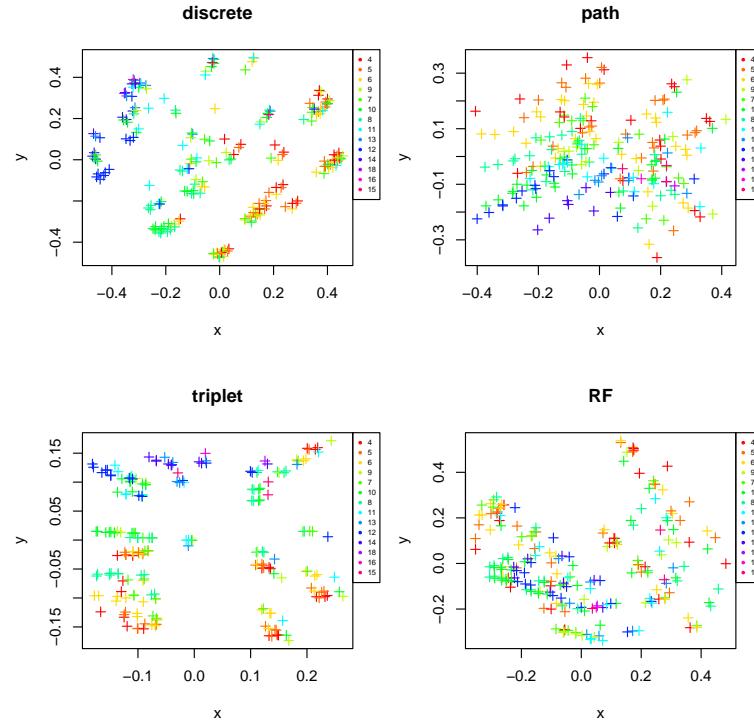


Figure 10: MDS plot with cost $c_c = 0, c_d = c_l = c_s = 1$ for the indicated distances.

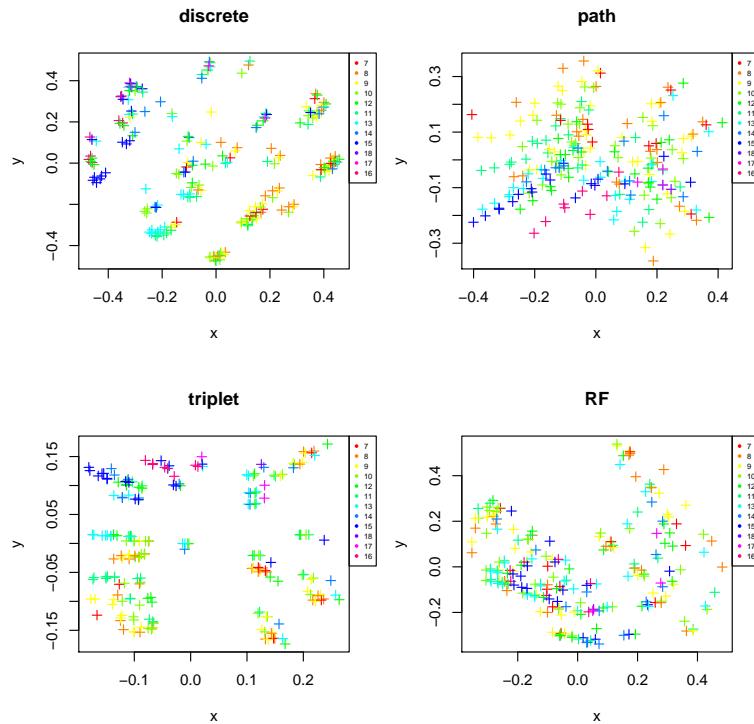


Figure 11: MDS plot with cost $c_c = 0, c_d = c_l = 1, c_s = 2$ for the indicated distances.

4 The CP dataset

Due to the large number of possible reconciliations, we uniformly sampled 200 reconciliations.

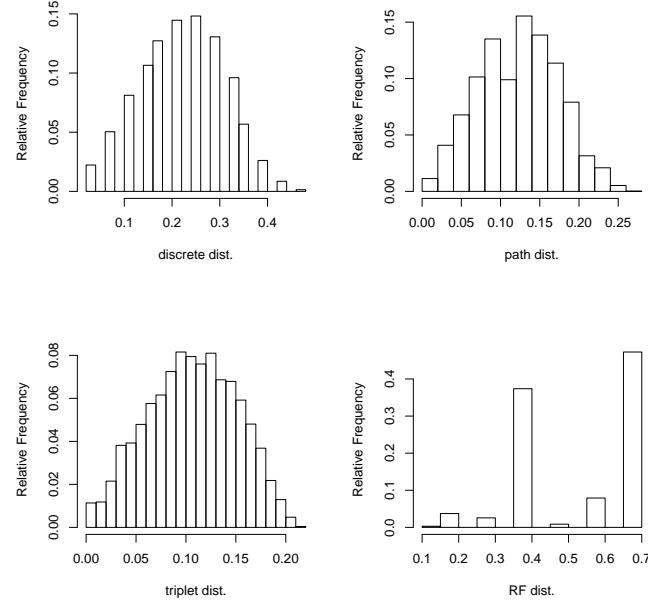


Figure 12: The relative frequency distributions of the (normalized) pairwise distance values between reconciliations,

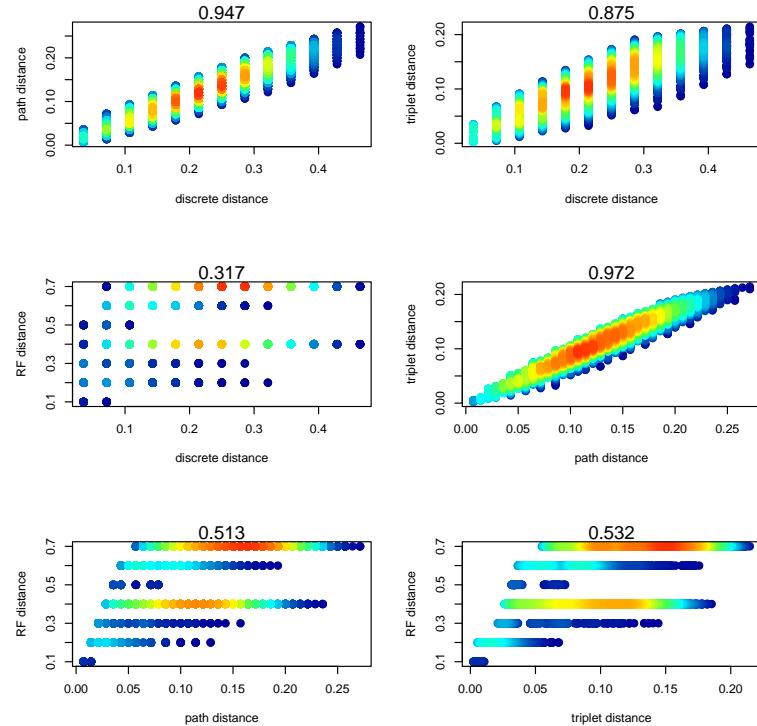


Figure 13: Correlation between the discrete/triplet/path/edit-distances. The Pearson correlation coefficient is above each panel.

The colour scale to the right of each panel corresponds to the reconciliation scores.

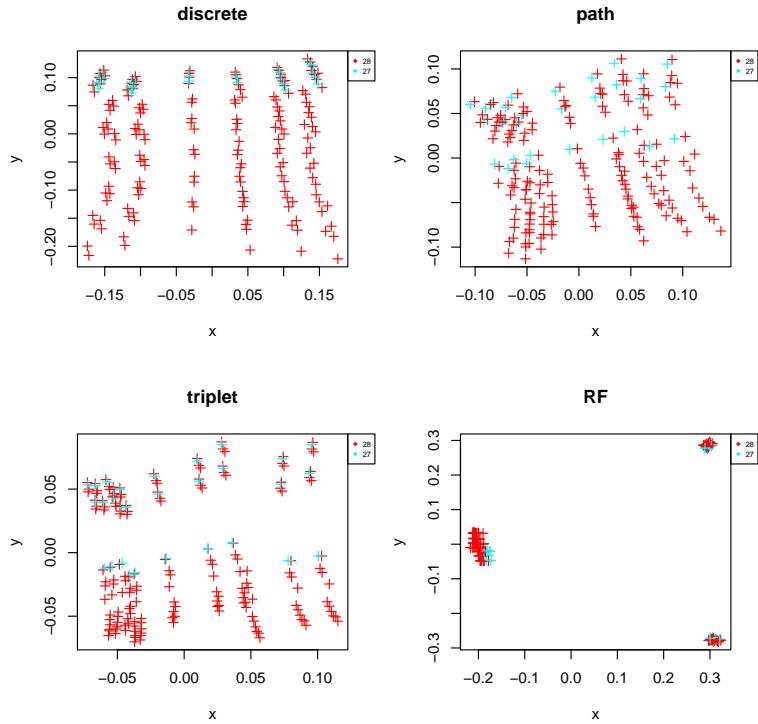


Figure 14: The MDS plot with cost $c_c = 0, c_d = c_l = c_s = 1$ for the indicated distances.

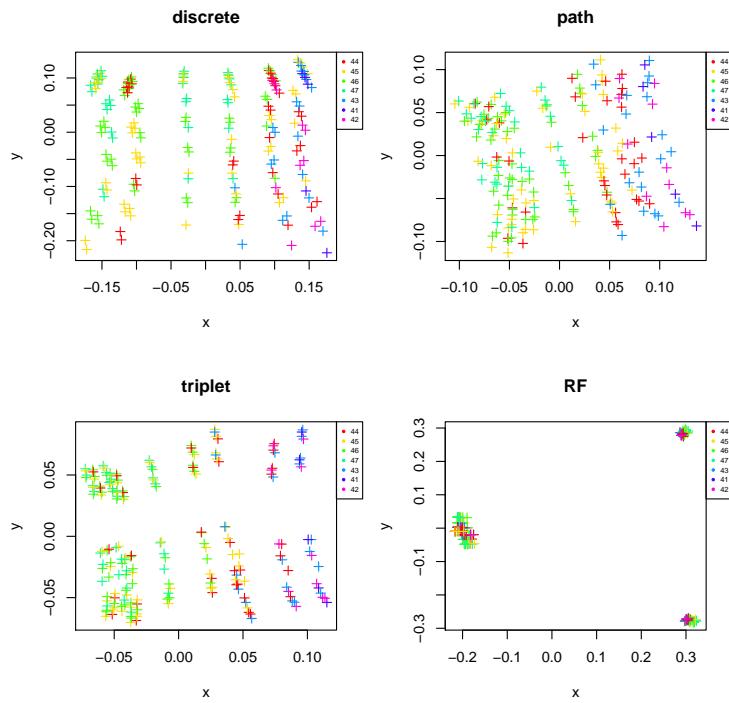


Figure 15: MDS plot with cost $c_c = 0, c_d = c_l = 1, c_s = 2$ for the indicated distances.

5 The PMP dataset

Due to the large number of possible reconciliations, we uniformly sampled 200 reconciliations.

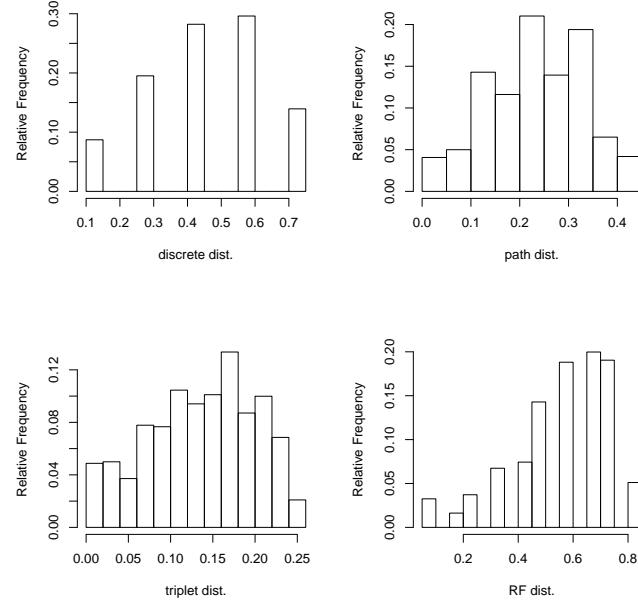


Figure 16: The relative frequencies distributions of the (normalized) pairwise distance values between reconciliations.

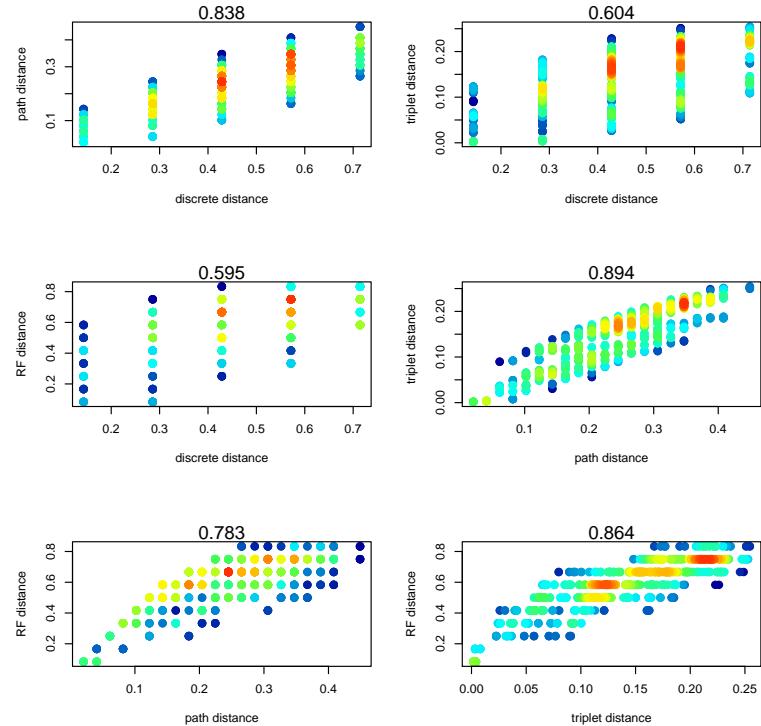


Figure 17: Correlations between the discrete/path/triplet/edit-distances. The Pearson correlation coefficient is above each panel.

The colour scale to the right of each panel corresponds to the reconciliation scores.

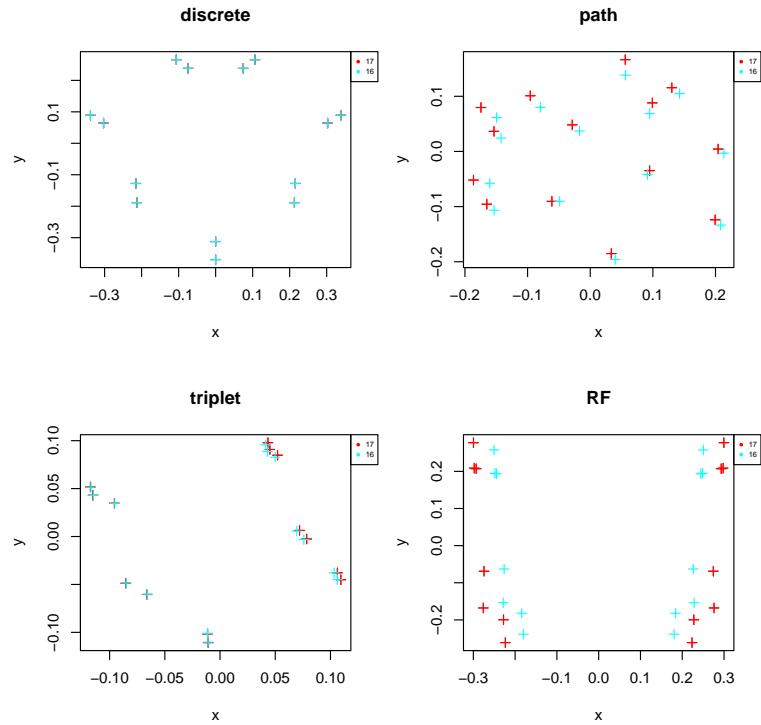


Figure 18: MDS plot with cost $c_c = 0$ and $c_d = c_s = c_l = 1$ for the indicated distances.

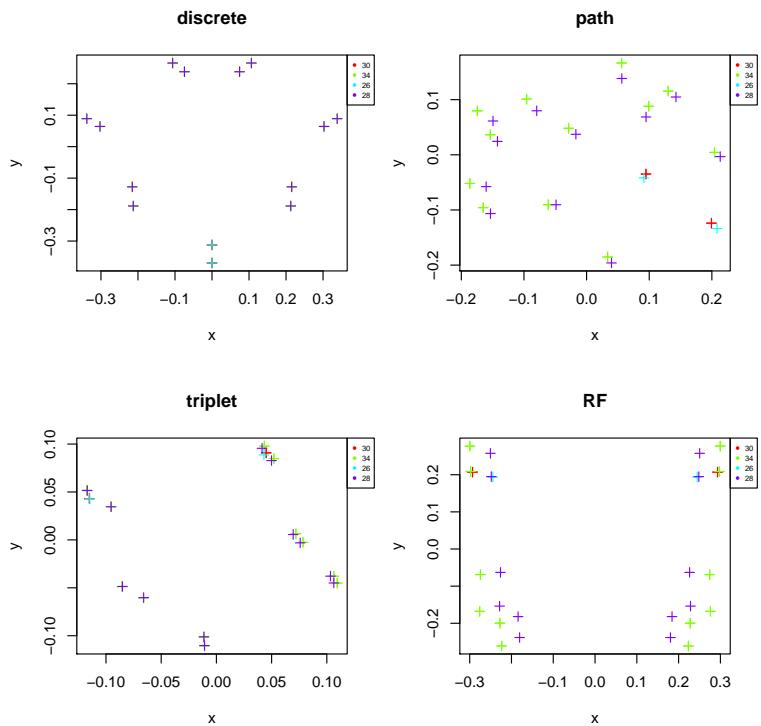


Figure 19: MDS plot with cost $c_c = 0$, $c_d = c_l = 1$, and $c_s = 2$ for the indicated distances.

6 The RH dataset

Due to the large number of possible reconciliations, we uniformly sampled 200 reconciliations.

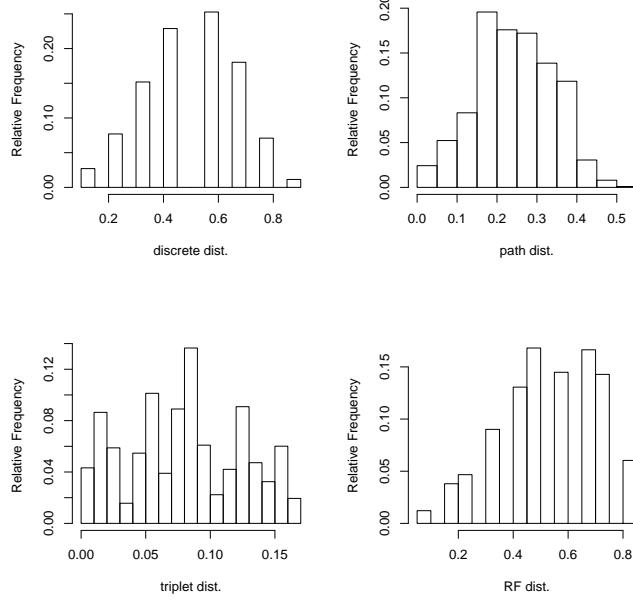


Figure 20: The relative frequency distributions of the (normalized) pairwise distance values between reconciliations.

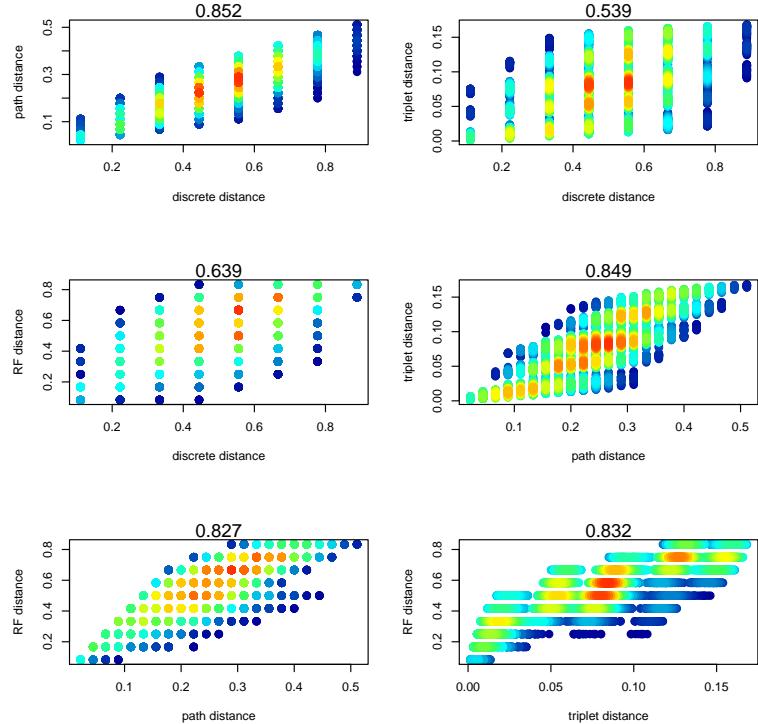


Figure 21: Correlations between the discrete/triplet/path/edit-distances. The Pearson correlation coefficient is above each panel.

The colour scale to the right of each panel corresponds to the reconciliation scores.

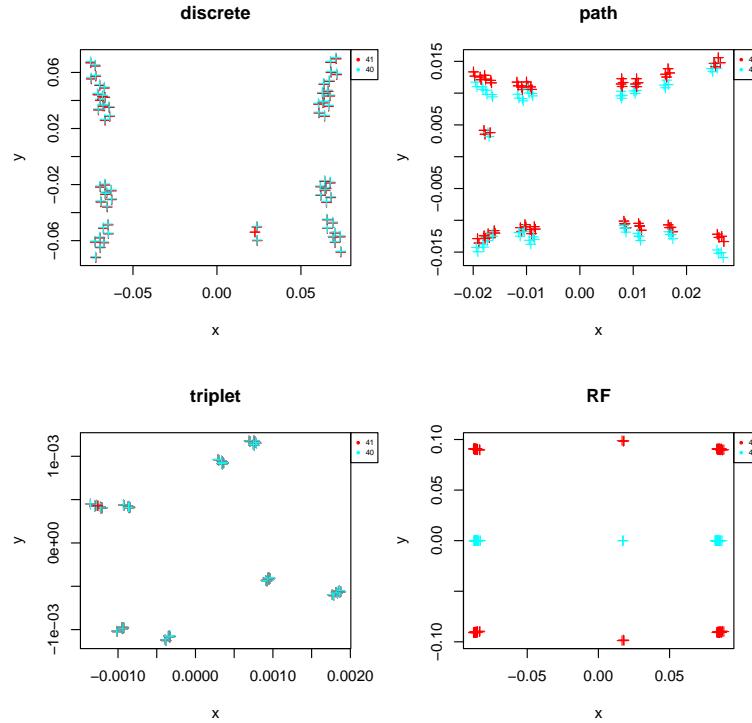


Figure 22: MDS plot with cost $c_c = 0, c_d = c_l = c_s = 1$ for the indicated distances.

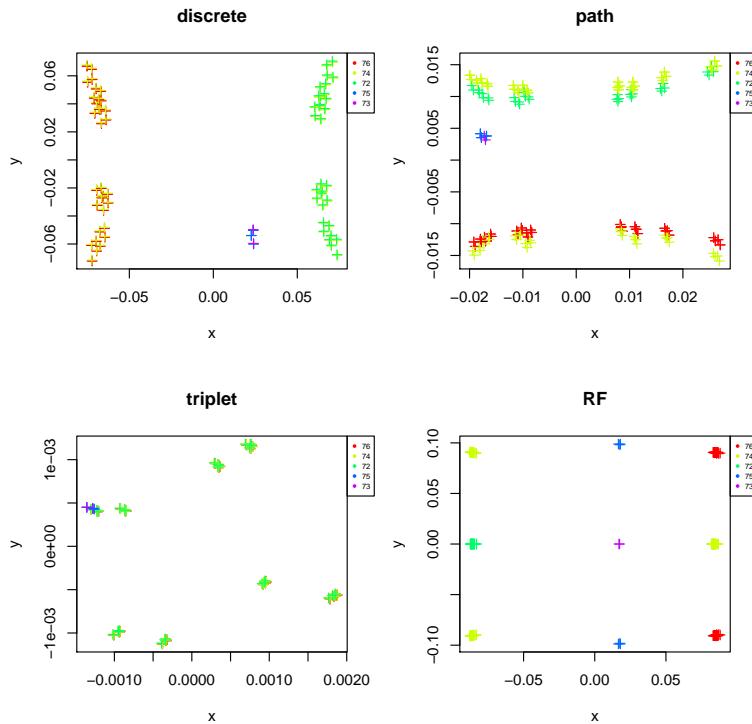


Figure 23: MDS plot with cost $c_c = 0, c_d = c_l = 1, c_s = 2$ for the indicated distances.

7 The EC dataset

Due to the large number of possible reconciliations, we uniformly sampled 200 reconciliations.

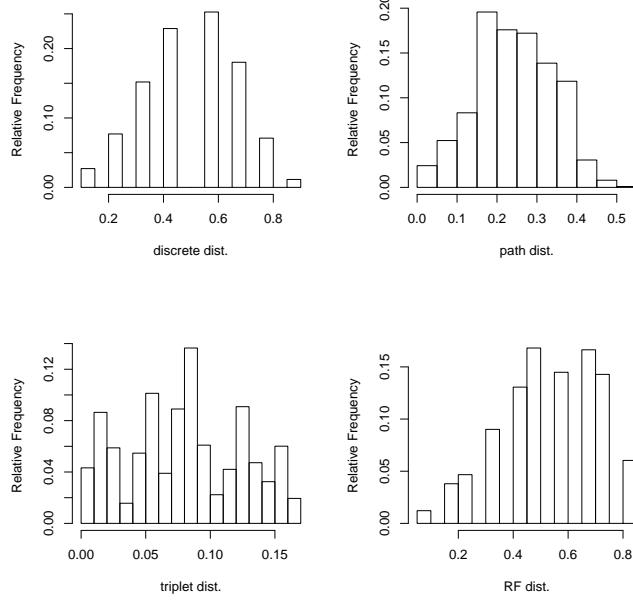


Figure 24: The relative frequency distributions of the (normalized) pairwise distance values between reconciliations.

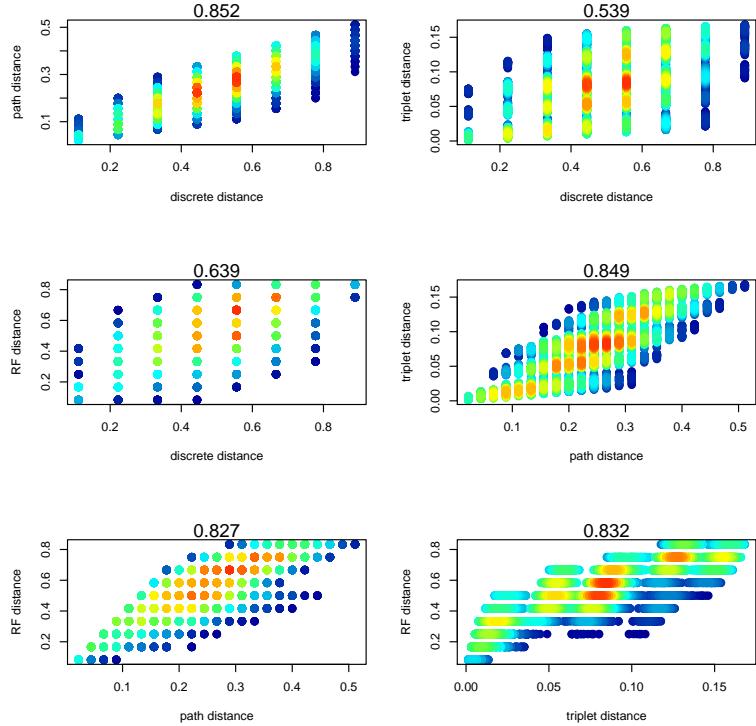


Figure 25: Correlations between the discrete/triplet/path/edit-distances. The Pearson correlation coefficient is above each panel.

The colour scale to the right of each panel corresponds to the reconciliation scores.

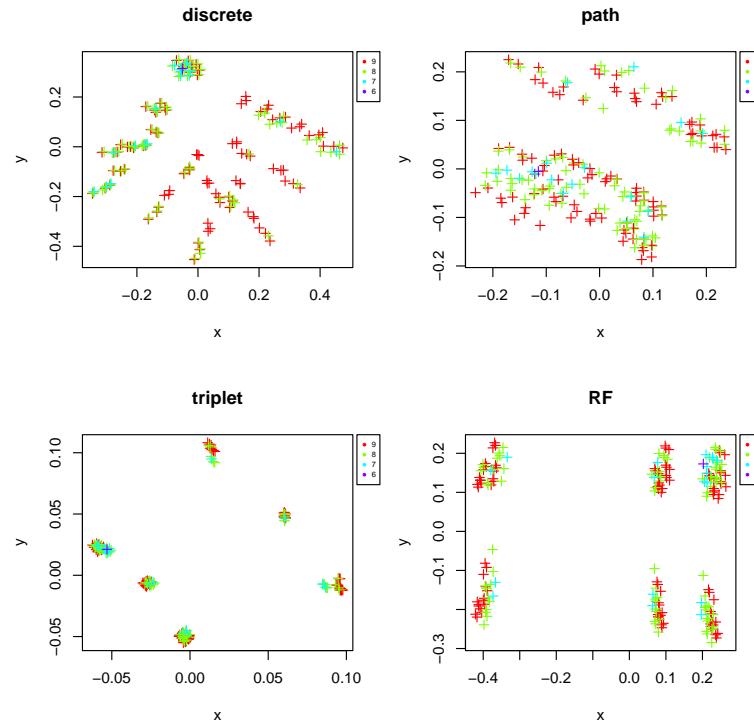


Figure 26: MDS plot with cost $c_c = 0, c_d = c_l = c_s = 1$ for the indicated distances.

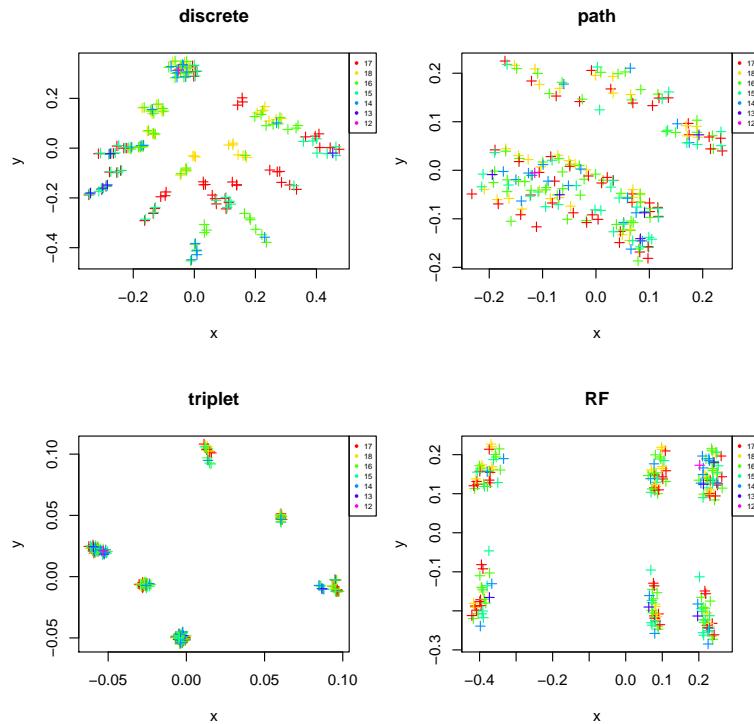


Figure 27: MDS plot with cost $c_c = 0, c_d = c_l = 1, c_s = 2$ for the indicated distances.

8 The GL dataset

Due to the large number of possible reconciliations, we uniformly sampled 200 reconciliations.

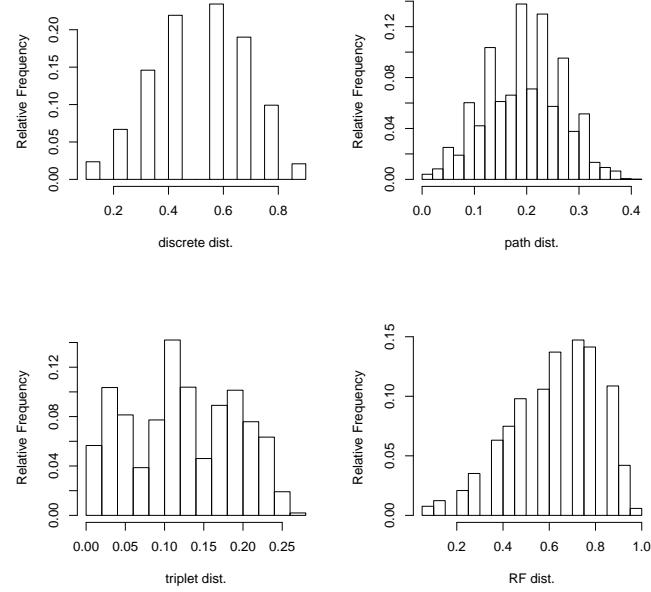


Figure 28: The relative frequency distributions of the (normalized) pairwise distance values between reconciliations.

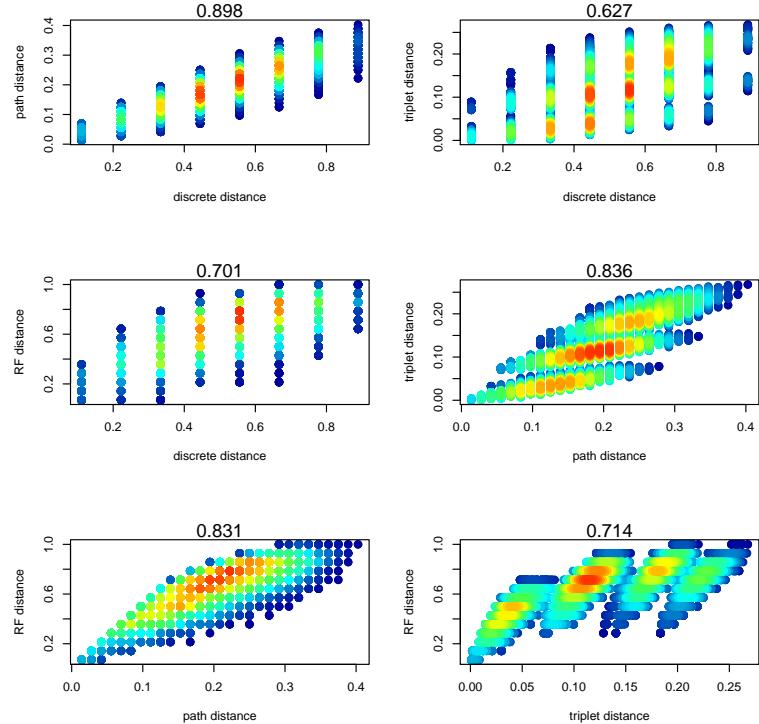


Figure 29: Correlations between the discrete/triplet/path/edit-distances. The Pearson correlation coefficient is above each panel.

The colour scale to the right of each panel corresponds to the reconciliation scores.

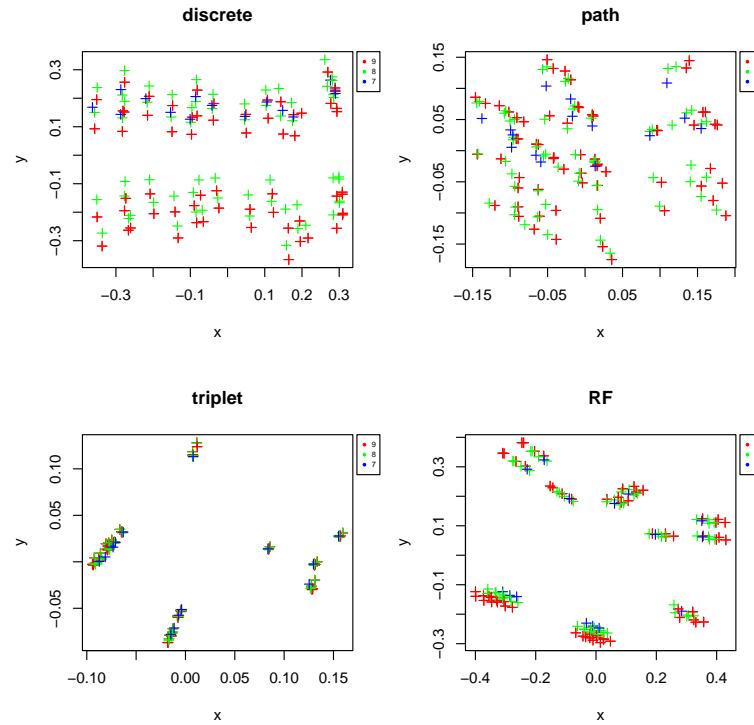


Figure 30: MDS plot with cost $c_c = 0, c_d = c_l = c_s = 1$ for the indicated distances. 0111

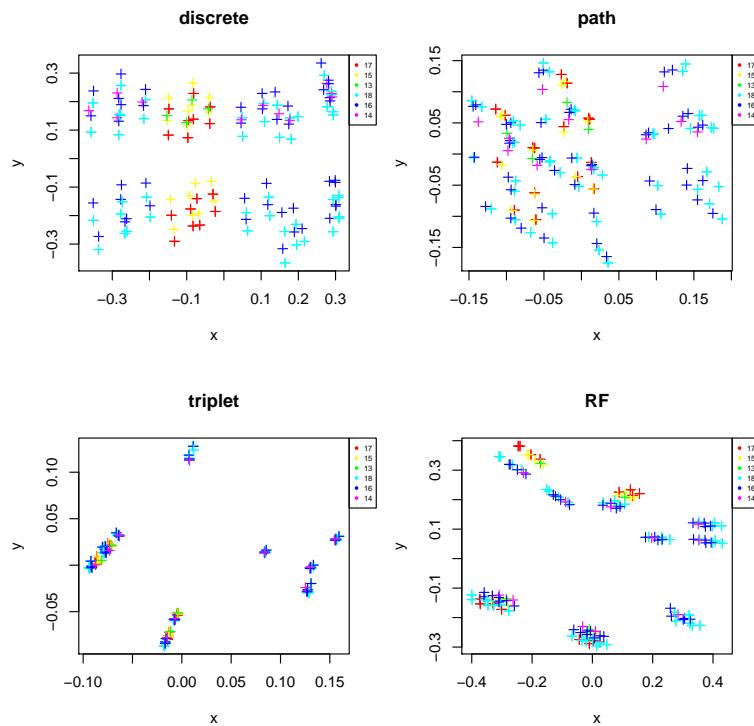


Figure 31: MDS plot with cost $c_c = 0, c_d = c_l = 1, c_s = 2$ for the indicated distances.

9 The SC dataset

Due to the large number of possible reconciliations, we uniformly sampled 200 reconciliations.

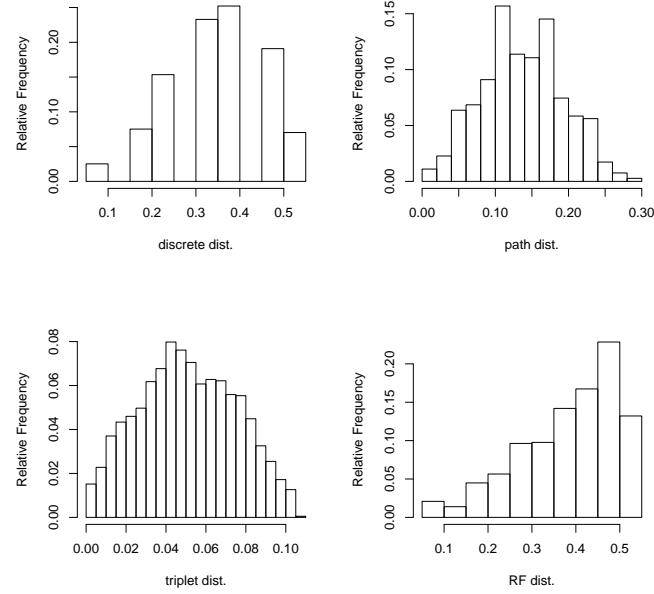


Figure 32: The relative frequency distributions of the (normalized) pairwise distance values between reconciliations.

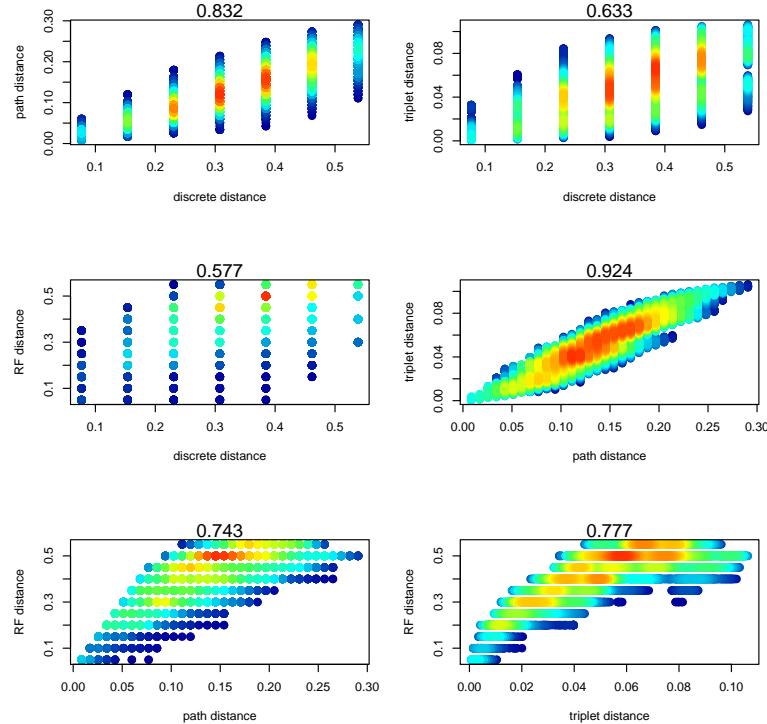


Figure 33: Correlations between the discrete/triplet/path/edit-distances. The Pearson correlation coefficient is above each panel.

The colour scale to the right of each panel corresponds to the reconciliation scores.

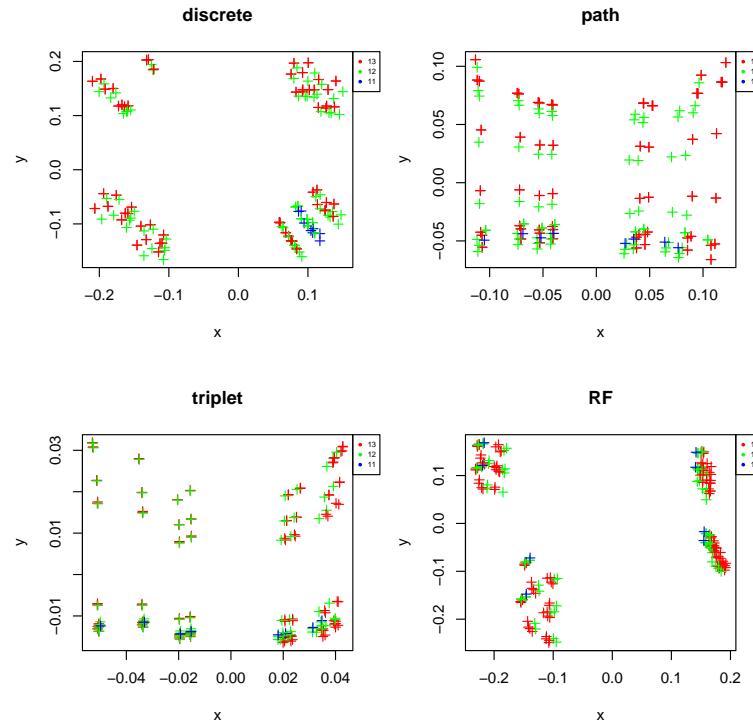


Figure 34: MDS plot with cost $c_c = 0, c_d = c_l = c_s = 1$ for the indicated distances.

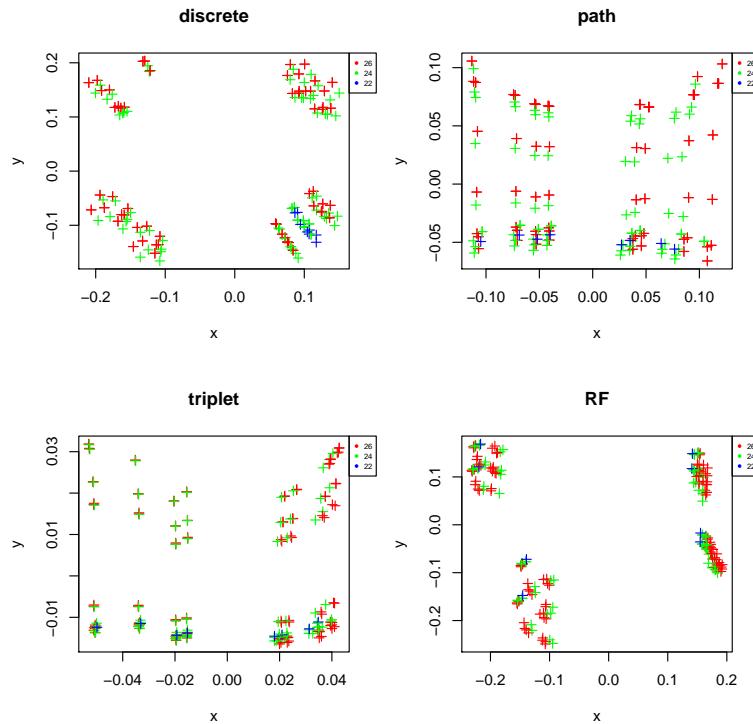


Figure 35: MDS plot with cost $c_c = 0, c_d = c_l = 1, c_s = 2$ for the indicated distances.

10 the SFC dataset

Due to the large number of possible reconciliations, we uniformly sampled 200 reconciliations.

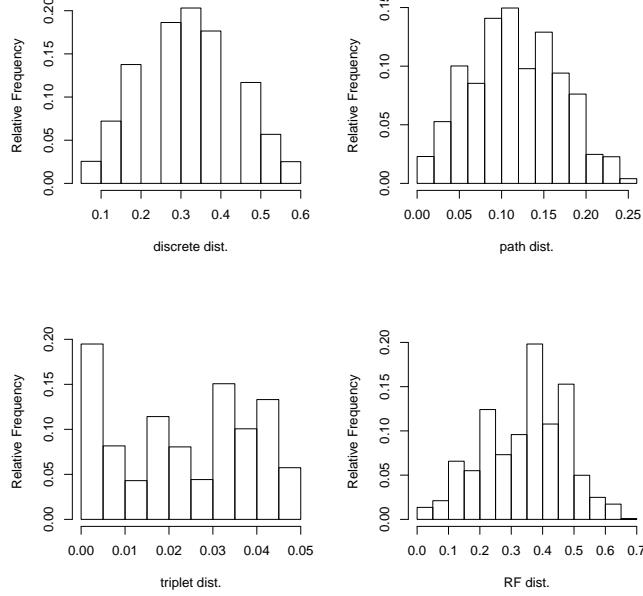


Figure 36: The relative frequency distributions of the (normalized) pairwise distance values between reconciliations.

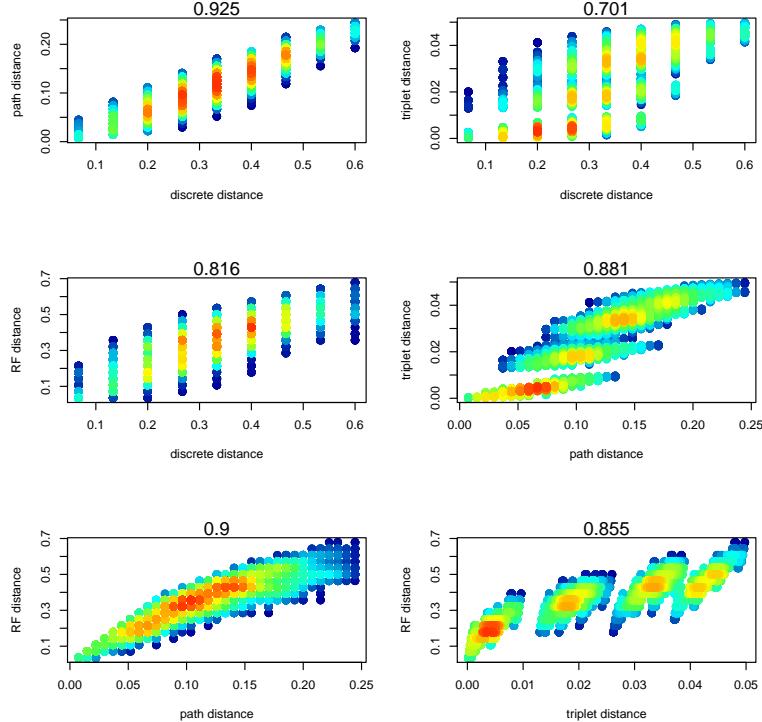


Figure 37: Correlations between the discrete/triplet/path/edit-distances. The Pearson correlation coefficient is above each panel.

The colour scale to the right of each panel corresponds to the reconciliation scores.

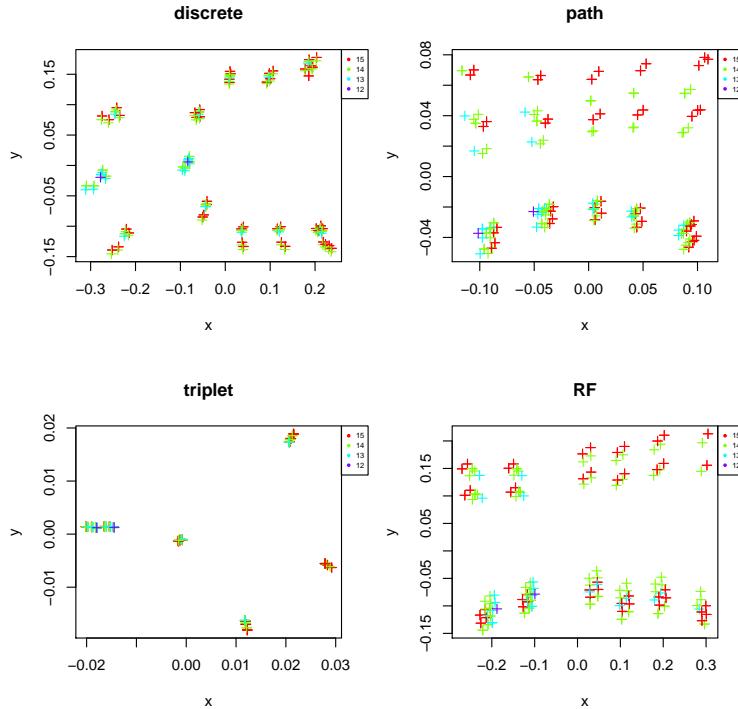


Figure 38: MDS plot with cost $c_c = 0, c_d = c_l = c_s = 1$ for the indicated distances.

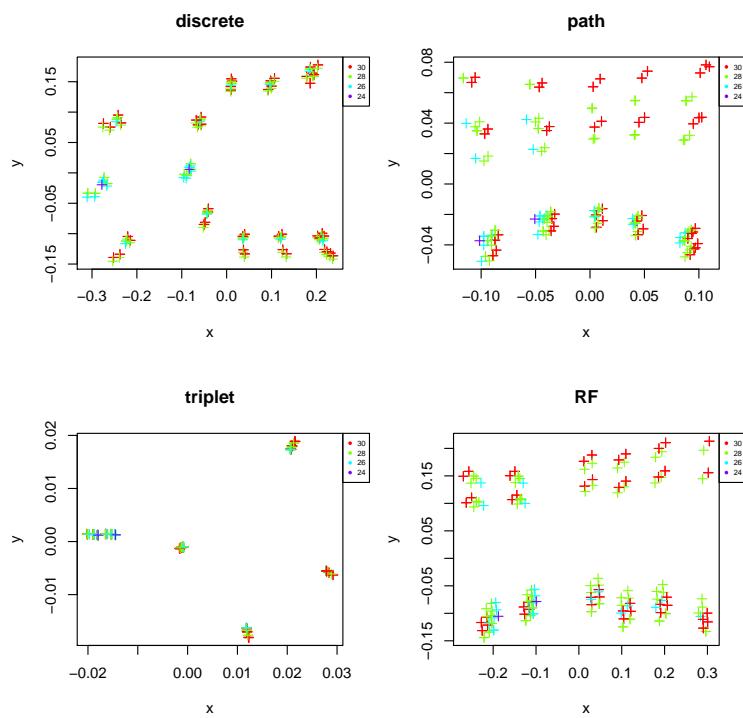


Figure 39: MDS plot with cost $c_c = 0, c_d = c_l = 1, c_s = 2$ for the indicated distances.

References

- [1] Charleston, M.A.: Jungles: a new solution to the host/parasite phylogeny reconciliation problem. *Math. Biosci.*, 149(2), 1998, 191–223.
- [2] Deng, J., Yu, F., Li, H.-B., Gebiola, M., Desdevises, Y., Wu, S.-A. and Zhang, Y.-Zh.: Cophylogenetic relationships between Anicetus parasitoids (Hymenoptera: Encyrtidae) and their scale insect hosts (Hemiptera: Coccidae), *BMC Evol. Biol.*, 13(1), 2013, 1–11.
- [3] Murray, E. A., Carmichael, A. E. and Heraty, J. M. : Ancient host shifts followed by host conservatism in a group of ant parasitoids, *Proceedings of the Royal Society of London B: Biological Sciences*, The Royal Society, 280(1759), 2013.
- [4] Mendlová, M., Desdevises, Y., Civáňová, K., Pariselle, A. and Šimková, A. : Monogeneans of West African cichlid fish: evolution and cophylogenetic interactions, *PLoS One*, Public Library of Science, 7 (5), 2012.
- [5] Hughes, J., Kennedy, M., Johnson, K. P., Palma, R. L. and Page, R. DM. : Multiple cophylogenetic analyses reveal frequent cospeciation between pelecaniform birds and Pectinopygus lice, *Syst. Biol.*, 56(2) , 2007, 232–251.
- [6] Ramsden, C., Holmes, E. C. and Charleston, M. A.: Hantavirus Evolution in Relation to Its Rodent and Insectivore Hosts: No Evidence for Codivergence, *Mol. Biol. Evol.*, 26(1), 2009, 143–153.
- [7] Hafner, M. and Nadler, S. : Phylogenetic trees support the coevolution of parasites and their hosts, *Nature*, 332, 1988, 258–259.
- [8] Paterson, A. M., Ricardo, L. P. and Gray, R. D. : Drowning on arrival, missing the boat, and x-events: how likely are sorting events?, *Tangled trees: Phylogeny, cospeciation, and coevolution*, Page,Roderic D. M., Univ. Chicago Press, 2003, 287–309.
- [9] Refregier, G., Le Gac, M., Jabbour, F., Widmer, A., Shykoff, J., Yockteng, R., Hood, M. and Giraud, T. : Cophylogeny of the anther smut fungi and their caryophyllaceous hosts: Prevalence of host shifts and importance of delimiting parasite species for inferring cospeciation, *BMC Evol. Biol.*, 8 (1), 2008, 100.