Extracting the speciation signal from multigene families

J.F. Dufayard[1], C. Scornavacca[2], V. Ranwez[3], V. Berry[1] and V. Daubin[4]

[1] LIRMM, CNRS - Univ. Montpellier 2, 161 rue Ada, 34392 Montpellier Cedex 5, France
[2] Center for Bioinformatics ZBIT, Tübingen University, Sand 14, 72076 Tübingen, Germany
[3] ISEM, CNRS - Univ. Montpellier 2, Place E. Bataillon - CC 064 - 34095 Montpellier, France
[4] LBBE, CNRS - Univ. Lyon 1, 43 bd du 11 novembre 1918, 69622 Villeurbanne Cedex, France

Introduction

Due to reasonable running times, **supertree methods** are often employed to build species trees when dealing with very large sets of gene trees, such as those stored in phylogenomic databases. This approach first infers individual gene trees from sets of orthologous sequences, and then combines the gene trees in a comprehensive tree on the set of all species. One serious limitation of current supertree methods is handling only **mono-copy gene trees**, i.e. that built from genes appearing in at most one copy in each studied species. In contrast, gene trees are usually **multi-copy** (**MUL trees**), i.e. several leaves being labeled by the same species whose genome contributed several sequences due to duplication or transfer events. These gene families are currently discarded when building supertrees for the species concerned. As more genomes become available, the percentage of gene families with paralogous sequences can only increase, which emphasizes the fact that MUL trees definitely need to be integrated in phylogenomic analysis.

Summarizing MUL trees into mono-copy gene trees

We propose to tackle this problem by extracting a largest possible amount of speciation information from MUL trees [7, 8]. This speciation signal can then be turned into mono-copy gene trees to feed supertree methods. First of all, **observable duplication nodes** are identified by a linear time algorithm. For gene trees containing duplications, we propose to separately preprocess them in order to remove their redundant parts with respect to speciation events. This is achieved by resorting to a linear time isomorphism algorithm that accepts MUL trees as input. This algorithm is applied to the pairs of subtrees hanging from duplication nodes in a MUL tree to reduce the number of duplication nodes. Note that this process generalizes the natural handling of MUL trees whose paralogy is limited to **in-paralogy**. For gene trees that still have duplication nodes, we define a set of triplets (binary rooted trees on three leaves) containing the topological information of a MUL tree that can be considered to be related to speciation events. When this set is compatible, the MUL tree contributes a coherent speciation signal to build the species tree. In such a case, we propose to extract a set of subtrees, both coherent and free of duplication events (see Figure 1). All algorithms were implemented in C++ using Bio++ [2] and they are available at `http://www.atgc-montpellier.fr/ssimul/`.

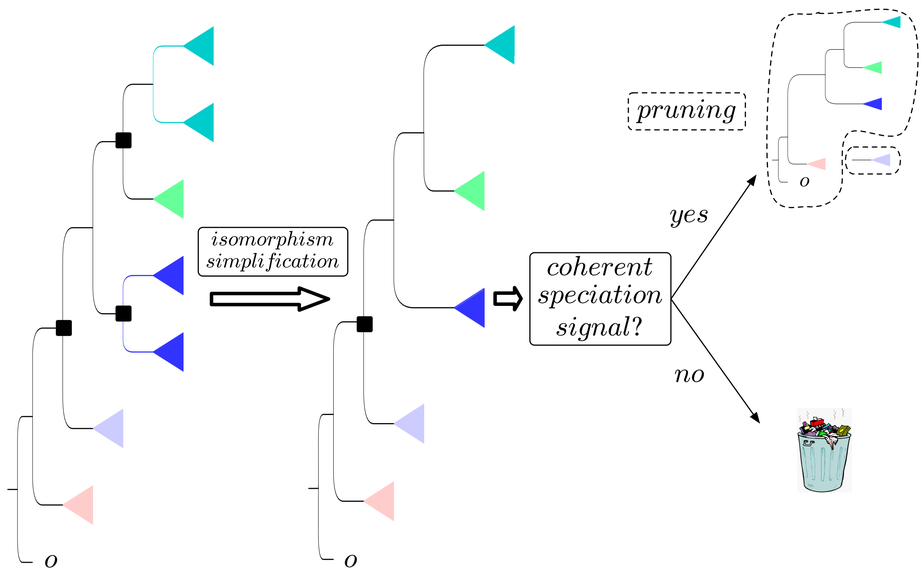


Figure 1: Procedure used to obtain mono-copy trees from MUL trees. Observable duplication nodes are indicated by a black square.

An application to phylogenomic databases: the Eukaryotes

We applied the above-described approach to the analysis of the HOGENOM database – release 5 [4], a database of homologous genes from fully sequenced genomes. We restrict our attention to the gene families containing at least three species belonging to the Eukaryotes **and** at least one species belonging to the Archaea. The latter restriction is a prerequisite to use the Archaea as outgroup for the Eukaryotes.
The sequenced were aligned with Muscle (default settings) and a step of alignment curation was performed with Gblocks (Allowed Gap Positions: With Half, default settings for other options). Gene trees were reconstructed with PhyML (WAG model, gamma law with 4 categories and shape parameter estimated, SPR, 5 random starting points + BioNJ starting point).
Eukaryotes trees were rooted using Archaea sequences as outgroup. If the two groups were not monophyletic, two more refined methods were applied: the first one tries to collapse some weakly supported branches at the root, the second one tries to detect misplaced subtrees of Archaea in Eukaryotes for various reasons (methodological artifacts or ancient horizontal transfers). Finally, the set of rooted gene trees was restricted to the Eukaryote species.
At the end of this procedure, we obtained a forest F composed of **826 rooted trees**. From F we computed:

- F_1, the forest of mono-copy gene trees of F;
- F_2, the forest of trees of F that are multi-copy and can be turned into mono-copy gene trees by removing a copy of each pair of isomorphic sibling subtrees;
- F_3, the forest of trees of F obtained by pruning the gene trees that are still multi-copy after applying the isomorphic simplification, but are auto-coherent.

We denote by F_{all} the forest obtained by the union of F_1, F_2 and F_3. The main characteristics of these forests are given in Table 1. Running times were computed on a 2.53-GHz-Intel Core 2 Duo and 4 GB RAM machine.

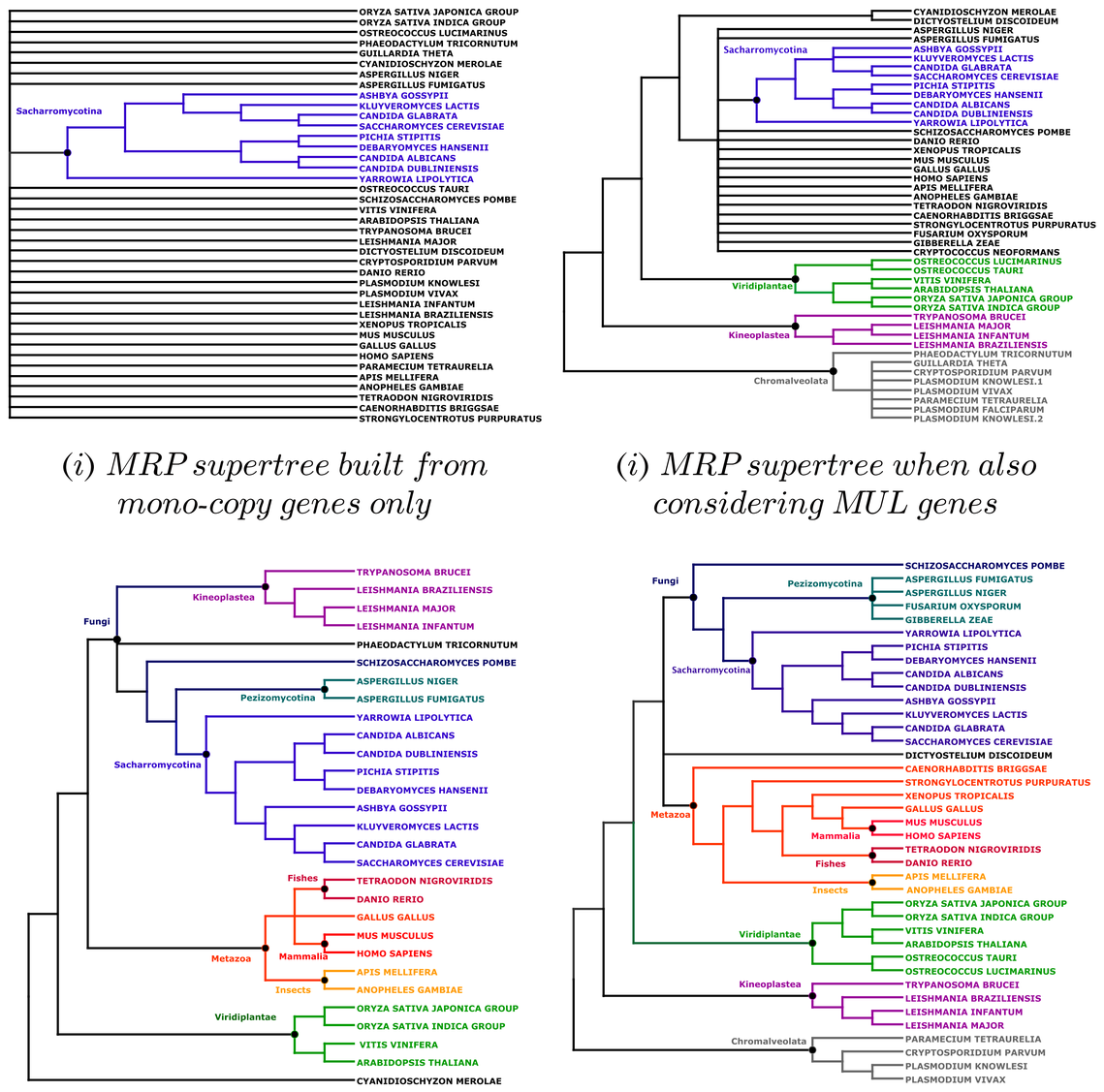
	F_1	F_2	F_3	F_{all}
#trees	117	142	336+224	595+224
#triplets	15,487	173,516	613,090	802,093
#distinct triplets	5,296	24,702	27,376	28,544
#species	40	45	44	45
% of resolved **triples**	44.75	71.4	75.35	74.83
running time	ca 4 s		ca 7s	ca 11 s

Table 1: Information contained in the four forests considered to build the species tree for 45 eukaryotic species.

From the number of trees in the different forests displayed in Table 1, it can be observed that the methods proposed in this work allows the eukaryotic species tree to be built on the basis of up to **595 gene families**, where only 117 mono-copy gene trees can be used by the traditional approach. Even more impressive is the amount of topological information gained to build the species tree, e.g. the number of triplets contained in F_{all} is **ca 50 times** that contained in F_1.

Building the species tree

We now examine whether the increase in the amount of available information benefits the species tree construction step, i.e. whether the information extracted from MUL trees is of good quality. To build supertrees, we used the F_1 and F_{all} forests. Two supertree methods were used: the well-known MRP method [1] and the more mathematically founded PhySIC_IST method [6]. PhySIC_IST was used with a correction threshold of 0.5, having previously collapsed the branches with an aLRT support less than 0.8.
A first general observation is that the supertrees proposed by both methods are much more resolved when using the forest F_{all} rather than only the set of mono-copy gene trees F_1. The most resolved tree, depicted in Figure 2(iv), is obtained by PhySIC_IST on the forest F_{all} and it is in agreement with our current knowledge of the eukaryotic phylogenies (e.g. [5]), based on the concatenation of a few universal, single copy genes. Notably, animals and fungi form a monophyletic group. The inclusion of Dictyostelium in this clade is a particularly interesting result. However, this tree may also exhibit some well known systematical artifacts present in gene trees such as the early branching of Caenorhabditis in the animal clade. Such artifacts may disappear when more complete eukaryotic genomes are added to the tree.



(i) *MRP supertree built from mono-copy genes only*

(i) *MRP supertree when also considering MUL genes*

(iii) *PhySIC_IST supertree built from mono-copy genes only*

(iv) *PhySIC_IST supertree when also considering MUL genes*

Figure 2: Supertrees built by MRP (i, ii) and PhySIC_IST (iii, iv), using as input the forest F_1 (i, iii) or the forest F_{all} (ii, iv). The supertrees have been drawn using Dendroscope [3].

Conclusions

We proposed several algorithms to transform multi-copy gene trees into mono-copy ones, so that they can be used by supertree methods. Results on an eukaryotic dataset showed that not only do these algorithms allow more information to be extracted than with traditional approaches, but that supertrees inferred from this extra information are much more resolved and globally in accordance with phylogenetic knowledge. Moreover, the effort required to obtain efficient algorithms results in very reasonable running times.

Acknowledgments

This work was funded by the ANR-08-EMER-011 project (`http://www.lirmm.fr/phylariane/`).

References

[1] B. R. Baum and M. A. Ragan. The MRP method. In O.R.P. Bininda-Emonds, editor, *Phylogenetic supertrees: combining information to reveal the Tree of Life*, pages 17–34. Kluwer, 2004.

[2] J. Dutheil, S. Gaillard, E. Bazin, S. Glemin, V. Ranwez, N. Galtier, and K. Belkhir. Bio++: a set of c++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics. *BMC Bioinformatics.*, 7(1):188, 2006.

[3] D. H. Huson, D. C. Richter, C. Rausch, T. Dezulian, M. Franz, and R. Rupp. Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics*, 8(1):460, 2007.

[4] S. Penel, A. M. Arigon, J. F. Dufayard, A. S. Sertier, V. Daubin, L. Duret, M. Gouy, and G. Perrière. Databases of homologous gene families for comparative genomics. *BMC Bioinformatics*, 10 Suppl 6:S3, 2009 (`http://pbil.univ-lyon1.fr/`).

[5] N. Rodriguez-Ezpeleta, H. Brinkmann, G. Burger, A.J. Roger, M.W. Gray, H. Philippe, and B.F. Lang. Toward resolving the eukaryotic tree: The phylogenetic positions of jakobids and cercozoans. *Current Biology*, 17:1420–1425, 2007.

[6] C. Scornavacca, V. Berry, V. Lefort, E. J. P. Douzery, and V. Ranwez. Physic_ist: cleaning source trees to infer more informative supertrees. *BMC Bioinformatics*, 9(8):413, 2008 (`http://www.atgc-montpellier.fr/physic_ist/`).

[7] C. Scornavacca, V. Berry, and V. Ranwez. From gene trees to species trees through a supertree approach. In Adrian Horia Dediu, Armand-Mihai Ionescu, and Carlos Martín-Vide, editors, *LATA*, volume 5457 of *Lecture Notes in Computer Science*, pages 702–714. Springer, 2009.

[8] C. Scornavacca, V. Berry, and V. Ranwez. Building species trees from larger parts of phylogenomic databases. *Information and Computation* (accepted), 2010.