Plant architecture comparison methods: A review of existing algorithms and examples of application

Aïda Ouangraoua[a], Vincent Segura[b], Evelyne Costes[b] and Pascal Ferraro[a]

[a]LaBRI – Université Bordeaux 1 - Talence, France {ouangrao|ferraro}@labri.fr
[b]INRA – UMR DAP - AFEF Team INRA, Montpellier, France {segura|costes}@supagro.inra.fr

Keywords: Plant comparison, edit distances, local similarities

Introduction

Among a number of generic tools that have been developed in the last decade to measure, explore, analyze, model and visualize plant architecture in 3-dimensions, this paper focuses on methods dedicated to the comparison between plant architectures. After a first step that has been reached with the comparison of branching sequences along axes (Guédon *et al.*, 2003), Ferraro and Godin (2000) have extended comparison methods to the whole branching systems. Their method was applied to the comparison of plants formalized as tree-graphs (Godin and Caraglio, 1998). In the last two years, new algorithms were explored and implemented in order to compare plant architectures with different relationships between components of the plant structure taken into account in the corresponding formal representations. The present paper provides a brief review of the comparison techniques now available to compare plant architectures depending on their formal representation. Their relative advantages / disadvantages are presented through an example of application on the comparison of two-year-old apple F1 hybrids.

Formal representation of plants

Plant architecture can be formally described as tree-graphs (Godin and Caraglio, 1998) by defining a set of vertices V that represents the plant components, and a list E of vertex pairs that describes the adjacency of these components. Modeling a plant topology by an *unordered tree-graph* (*ie.* no ordering is considered on the set of siblings of any vertex) corresponds to a general representation of plant architectures (Ferraro and Godin, 2000). However, in plants that give rise to one branch on each node, the set of plant components is totally ordered and the topological structure of the plant can be represented by an *ordered tree-graph* (Fig. 1a). For plants bearing more than one branch on each node, only a semi-order between siblings of a node can be defined. This phenomenon is observed for instance in whorl plants, or can result from sampling procedure when a large number of plants have to be described (Segura *et al.*, 2006). In this case, the plant topology is represented by a *semi-ordered tree-graph* (Fig. 1b) (Ouangraoua and Ferraro, 2007b).

Furthermore, to take into account the multiscale nature of plant structures, plants can also be represented by quotiented tree-graphs (Godin and Caraglio, 1998). A *quotiented tree-graph* is a tree-graph with an equivalence relation defined on the set of vertices such that the resulting quotient graph is also a tree-graph. So far, a plant can be represented by a tree-graph quotiented or not, and ordered, semi-ordered or unordered. Thus, the comparison of plant architectures consists in comparing tree-graphs of one of these six data structures types.

Plant comparison methods

The Tree-to-Tree Editing Problem (Selkow, 1977) consists in computing a distance between two tree-graphs as the minimum cost of a *sequence of elementary operations* that converts one tree-graph into the other. Three elementary operations called *edit operations* on tree-graphs are currently used: vertex *insertion*, vertex *deletion* and *substitution* of vertices. The core structure of the

algorithms computing an edit distance between tree-graphs is based on the dynamic programming principle and determines an optimal *mapping* between tree-graphs. Intuitively, a mapping between two tree-graphs is a description of which operation in a sequence of edit operations was applied to each vertex of both tree-graphs. Definitions of mappings between tree-graphs and algorithms determining the corresponding optimal mapping have been first proposed by Zhang and Shasha (1989) and Zhang (1996) for respectively ordered and unordered tree-graphs. To deal with other plant architecture representations, we recently introduced new constraints on the definition of mappings (and then new algorithms) adapted to semi-ordered (Ouangraoua and Ferraro, 2007a), quotiented unordered (Ferraro and Godin, 2003), quotiented ordered and quotiented semi-ordered (Ouangraoua and Ferraro, 2007b) tree-graphs.

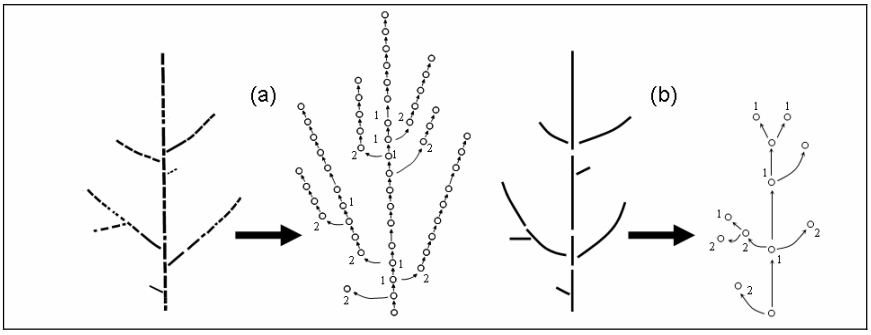

Fig. 1**: Schematic representations of a plant architecture (left) and corresponding tree-graph (right); (a) Ordered tree-graph representation: each node bears no more than two branches; (b) Same tree without information at node level and the corresponding semi-ordered tree-graph. The order relationship between children is indicated only when there is more than one child.**

Furthermore, in many cases plants share only a limited region of similarity. This may be a common domain or simply a short region of recognizable similarity. The *local similarity problem* (Smith and Waterman, 1981) aims at identifying the best pair of regions, one from each plant, such that the similarity of these two regions is the highest possible. In order to deal with this problem, we used two variants of the global edit distance algorithms to locally compare tree-graph representations of plant architecture: namely local similarity algorithm (Ouangraoua *et al.*, 2007) and *end-space free alignment* (Gusfield, 1997). The end-space-free alignment between two trees T_1 and T_2 is similar to the global edition. In this variant, any indel operations at the end or at the beginning of T_2 contribute to a weight zero. Since the indel operations have different cost depending if there are applied in one or the other tree, the corresponding dissimilarity measure $d(T_1,T_2)$ is not symmetric. The symmetry was reached by taking the minimum between $d(T_1,T_2)$ and the end-space-free alignment between T_2 and T_1 (ie. $d(T_2,T_1)$) : $D(T_1,T_2) = \min \{d(T_1,T_2), d(T_2,T_1)\}$. By contrast, the local similarity algorithm tends to maximize the similarity score between the compared plants and the resulting score is directly symmetric.

For each algorithm a distance matrix containing the distances between topological structure tree-graph representations of all pairs of plants in the database can be computed. Plants can thus be classified on the basis of the distance matrix, by classic clustering algorithms such as Ward's method (Gordon, 1999). However, in end-space-free comparison the relative dissimilarities are not conserved within the set of trees, and the triangular inequality is not preserved. This didn't allow us to biologically interpret the clustering method which was thus not applied in that particular case. By contrast, in the local similarity, the triangular inequality is preserved and allowed us to interpret the clustering which was performed after transformation of the scores into dissimilarities.

In all cases, 3D representations of the plants, obtained with PlantGL viewer (Boudon, 2001), provides a useful tool to interpret the results. Within-tree local similarities are then analyzed by identifying mapped entities on pairwise tree-graphs.

Example of applications to two-year-old apple hybrids

Global comparison algorithms presented in the previous section were performed to evaluate topological similarity of plants on a database of hybrids of apple trees (Segura *et al*., 2006). The results show that ordered, semi-ordered and unordered edit distance methods allow us to globally compare plant architectures (Segura *et al.*, 2007). Plants were grouped in three clusters according to their small, medium or large number of topological components (Fig. 2). However, the clustering remained unchanged when geometrical features or the class of entities were taken into account in the comparisons. An effect of the entity rank along ordered axes (mainly the trunk) was detected in the semi-ordered comparison, but only in the deepest steps of the clustering. When quotiented comparisons were performed, the results were still strongly correlated to the plant size, even though the dispersion of matching components was avoid by merging the conserved areas.

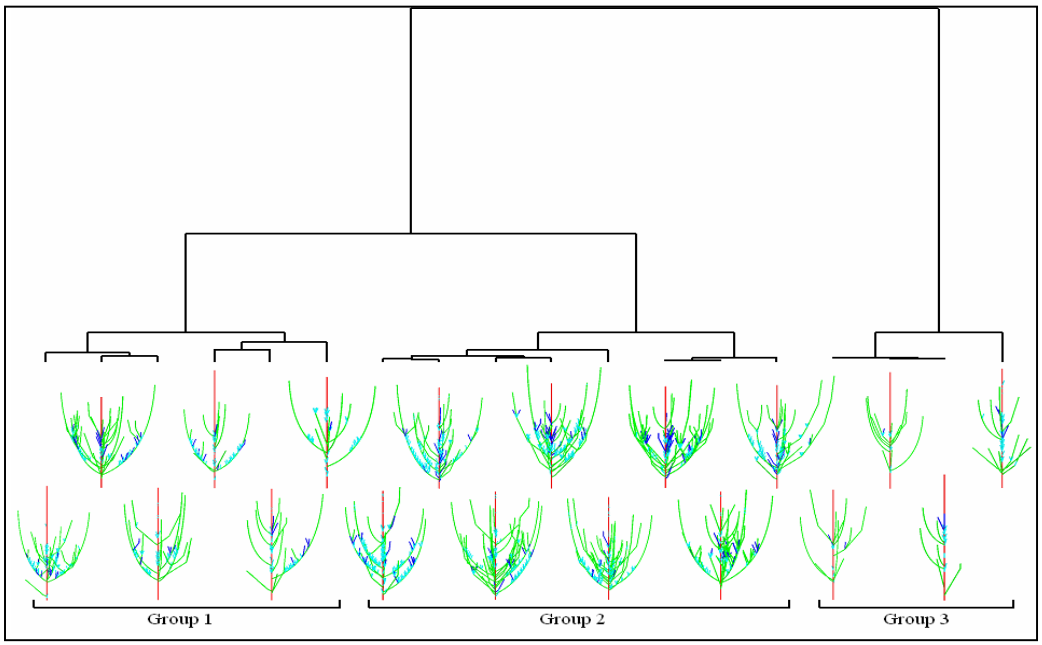

Fig. 2: **Cluster dendrogram from the distance matrix resulting from the unordered comparison of a sub-sample of 18 apple hybrids. Groups resulting from the selected separating level and used in the analysis are indicated.**

Local and end-space free comparisons were also applied to the database. Clustering methods show that the local comparison led to similar results as global comparison since the plants are still grouped according to their topological size. By contrast, different results were obtained with end-space free comparisons mainly because the non-matched extremities of trees were not taken into account during the distance computation. As previously mentioned, end-space free comparisons could not be interpreted by clustering techniques, and rather 3-D representations were used for interpretation purpose (Fig. 3). The distance between plants decreased spectacularly when compared plants had very different topological sizes. In fact, relatively small plants defined sub-parts that aligned in larger plants. By contrast, when plants with a close topological size were compared, the distance did not vary with the comparison method.

Conclusion

Finally, a set of edit distance algorithms is currently available to compare plants with different topological representations. For a given formal representation, both global and local comparison methods can be applied depending on the biological context and goal. However, each method is more or less appropriate depending on (*i*) the heterogeneity of the topological size of the compared

plants; (*ii*) the traits that must be taken into account. From a theoretical point of view, this analysis opens new perspectives to improve and extend the plant architecture comparison methods, especially the local comparisons or for dealing with more than two scales in the formal representation. Moreover, some general aspects concerning plant architecture, such as clustering problems, automatic labeling of plant structure and the evaluation of simulated plants arose from the definition of a distance between plants and will need further discussion.

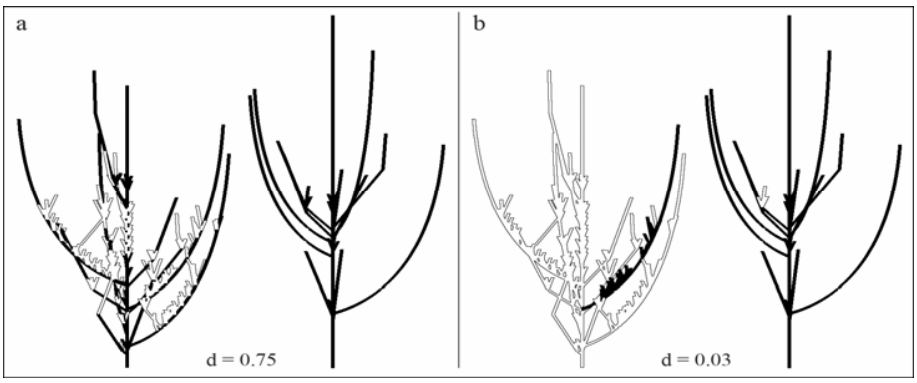

Fig. 3: **3-D Representation of the topological (a) and end-space free (b) comparisons between a pair of semi-ordered trees. Axes were colored depending on the comparison. Unmatched axes, resulting from insertion and deletion, were colored in white, while matched axes were colored in black. Distances between plants compared are indicated by the letter d.**

References

Boudon, F., Nouguier and C., Godin, C., GEOM module manual user guide, CIRAD, **2001**.

Ferraro, P. and Godin, C., A distance measure between plant architectures, *Annals of Forest Science*, **2000**, 57, 445-461.

Ferraro, P. and Godin, C., An edit distance between quotiented graphs, *Algorithmica*, **2003**, 36, 1-39.

Godin, C. and Caraglio, Y., A multiscale model of plant topological structures, *Journal of Theoretical Biology*, **1998**, 191, 1-46.

Gordon, A. D., Classification. 2nd edition. London: Chapman & Hall, **1999**.

Guédon, Y., Heuret, P. and Costes, E., Comparison methods for branching and axillary flowering sequences, *Journal of Theoretical Biology*, **2003**, 225, 301-325.

Gusfield, D., Algorithms on strings, trees and sequences - Computer Science and Computational Biology Cambridge University Press, **1997**.

Ouangraoua, A. and Ferraro, P., A new constrained edit distance between quotiented ordered trees, *submitted to Journal of Discrete algorithms*, **2007a**.

Ouangraoua, A. and Ferraro, P., An edit distance between partially ordered trees to evaluate similarity between plants, *submitted to Information Processing Letters*, **2007b**.

Ouangraoua, A., Ferraro, P., Dulucq, S. and Tichit, L., Local similarity between quotiented ordered trees, *Journal of Discrete Algorithms*, **2007**, 5, 23-35.

Segura, V., Cilas, C., Laurens, F. and Costes, E., Phenotyping progenies for complex architectural traits: a strategy for 1-year-old apple trees (Malus x domestica Borkh.),*Tree Genetics and Genomes*, **2006**, 2, 140-151.

Segura, V., Ouangraoua, A., Ferraro, P. and Costes E., Comparison of tree architecture using tree edit distances: application to two-year-old apple hybrids, *Euphytica*, **2007, doi:** 10.1007/s10681-007-9430-6.

Selkow, S. M., The tree-to-tree editing problem, *Information processing letters*, **1977**, *6*, 184-186.

Smith, T. and Waterman, M., Identification of common molecular subsequences, *Journal of Molecular Biology*, **1981**, 147, 195-197.

Zhang, K., A constrained edit distance between unordered labeled trees, *Algorithmica*, **1996**, 15, 205-222.

Zhang, K. and Shasha, D., Simple fast algorithms for the editing distance between trees and related problems, *SIAM Journal on Computing*, **1989**, 18, 1245-1262.