

---

## Analysis of the plant architecture via tree-structured statistical models: the hidden Markov trees

J.-B. Durand\*, Y. Guédon\*, Y. Caraglio\*, E. Costes\*\*

\*Unité Mixte de Recherche CIRAD/INRA/CNRS/Université de Montpellier II, Botanique et Bioinformatique de l'Architecture des Plantes TA40/PS2, 34398 Montpellier cedex 5, France

\*\* Unité Mixte de Recherche INRA/AgroM/CIRAD/IRD BDPPC, Equipe « Architecture et Fonctionnement des Espèces Fruitières », 2 place Pierre VIALA, 34060 Montpellier cedex 1, France

---

### Introduction

For many years, plant architecture has been viewed as the result of repetitions (Barlow, 1994), which occur at different levels of organisation (metamers, growth units, axis and branching systems) (Barthélémy, 1991). In addition, these plant components were shown to be distributed within individuals according to precise gradients (Barthélémy *et al.*, 1997). The changes which occur during plant ontogeny have been described along axis for successive entities and according to their position for the lateral ones. These changes reflect the impact of the plant topology, seen as a tree structure, on the plant entities. They occur in various plant species, for which the nature of the botanical entities and that of their successors tend to be equivalent, whereas branching tends to induce marked qualitative changes between the bearing entity and the borne branching system. However, the intensity of these changes has not yet been quantified, especially for comparing successive entities with lateral ones. We aim at characterising these changes by diverse quantitative or qualitative variables attached to a given entity such as the number of nodes, the length, the diameter and the presence/absence of flowering. These variables are called the entity *attributes*. Connected entities having similar attributes can be interpreted as homogeneous zones, as opposed to ruptures or transitions between zones. For example flowering is a factor of rupture in the plant architecture when the meristem death leads to sympodial branching. The discrimination between dominating and dominated axis in plants with different degrees of hierarchy can be formulated as the research of ruptures and continuities. More generally it makes sense to identify zones when the entities at a given scale can be clearly classified into a small number of classes defined by different morphological and functional characters. This is the case for various plant species where such a categorisation holds for the meristem functioning modes, in which only definite plant entities can be synthesised. For a given meristem, these modes are chronologically ordered and the order does not depend on the meristem. These ordered modes correspond to the notion of physiological age (Barthélémy *et al.*, 1997) describing the stage of differentiation of meristems. The physiological age of meristems can be assessed only indirectly; it is deduced from some of the biological characteristics of the plant, which are supposed to have an impact on the measured attributes.

A statistical approach is relevant for the analysis of architectural data, both for the exploratory analysis and for inferring some regularities or structures not directly apparent in the data. In our case, the aim of statistical models is to characterise such latent structures. These models are intended to make explicit some regularity, patterns or levels of organisation from the attributes, tree-structured zones for instance. The statistical analysis of sequential data of plant architecture, illustrated in (Guédon *et al.*, 2001), is mainly based on Markovian models, for instance hidden semi-Markov chains for modelling homogeneous zones. These models, though accurately accounting for the structure contained along remarkable paths in the plant (*e.g.* a tree trunk), are not relevant for identifying tree-structured zones, since the dependencies between entities of disjoint sequences are eluded. The complete topology has to be somehow included into the model for the existence of multiple dependent successors (or *descendants*) to be considered in the zone distribution.

We propose to use the statistical framework of the hidden Markov trees (HMTs) introduced by (Crouse *et al.*, 1998) in the context of signal processing to efficiently model homogeneous zones within a tree-structured process whose topology, fixed by the data, is thus non-random. The distribution of the vertex attributes is determined in HMTs by the value of a discrete hidden state. The persistence of these hidden states, leading to homogenous zones, is obtained by defining local dependencies between them. The HMT modelling is complementary with the plant comparison

method of (Ferraro and Godin, 2000), another approach for the analysis of tree-structured data based on tree alignment. This alignment relies on a distance between trees integrating the comparison of topology and a distance on the attributes. Instead, our method determines zones with common attribute distribution, the plant topology being locally taken into account by the dependencies between one entity and the adjacent ones. The Markovian models for sequences and trees and the tree alignment have been integrated to the AMAPmod software (Godin *et al.*, 1997) After a presentation of tree-structured representations of plants, the modelling of architectural data by HMTs is developed in this paper. This model relies on the above botanical concepts and hypotheses. Then some practical aspects and variants of the HMTs are presented, leading to refined botanical hypotheses and analysis methods, more relevant for given applications. Finally, several perspectives of concrete applications in agronomy and ecology are given.

### ***Tree-structured graphical representation of plants***

As discussed in (Godin *et al.*, 2000), plants can be formally described through rooted multiscale tree graphs whose vertices correspond to their constituting botanical entities and whose edges represent the physical connections between them. Each scale corresponds to one more or less macroscopic viewpoint on the plant. Since only single-scaled tree graphs can be analysed by HMTs, it is necessary to choose a scale for the plant description. Some topological information at a higher scale can nevertheless be taken into account in attributes, for example by counting the number of small shoots borne by a given axis. Such balance between topological information within the tree-structured data and its representation at the attribute level leads to modelling choices. For example, the representation of topology as an attribute may be required when analysing the development of some subparts of the plant. This is why the plant is typically represented at a rather macroscopic level, lower than the internode scale (growth unit, annual shoot and axis).

For each vertex  $u$  of the tree graph, the attribute vector is denoted by  $X_u$  and can mix qualitative and quantitative variables. The parent of  $u$  is denoted by  $\rho(u)$  (except if  $u$  is the root vertex) and the set of children of  $u$  is denoted by  $\mathbf{c}(u)$ . If this set is empty,  $u$  is called a *leaf vertex*. The different connection modes of plant entities are represented by typed edges: “<” for succession and “+” for branching. These notations are illustrated in figure 1.

### **Modelling zones in plants with hidden Markov trees**

We model the plant architecture’s organisation through a small number of ordered states associated with each entity, whence with each vertex  $u$  of the tree graph. These states determine the distribution of the morphological and functional characteristics of the entities measured by the attributes  $X_u$ . A set of connected vertices affected to a given state defines a homogeneous zone, whereas connected vertices affected to different states induce ruptures in the plant architecture. The propagation of the states within the plant is related to its topological organisation. This can be modelled in a probabilistic framework by the HMT models.

The HMTs have been introduced by (Crouse *et al.*, 1998) for modelling the dependencies in wavelet coefficient trees in the context of signal processing. The principle is to associate each vertex  $u$  with a hidden state  $S_u$  taking values in a finite set, so that the distribution of the attributes  $X_u$  depends on the value of  $S_u$  only. The dependencies between the states  $(S_u)_u$  ensure their propagation from one vertex to its children. They determine how the states, hence the zones are distributed. Since from section 1 some zone changes are expected to be irreversible, the states are frequently ordered. This is ensured by particular structures of the transition matrix  $P = (p_{ij})_{i,j}$ , where

$p_{ij} = P(S_u = j | S_{\rho(u)} = i)$  are model parameters. The dependencies between hidden states are essentially local. In the basic HMT model proposed by (Crouse *et al.*, 1998), each state is independent from all its non-descendants given the parent state. This local dependency assumption gives its name to the Markov property for trees. The HMT model is quite close to the hidden Markov chains: both have the same parameter set and are based on local dependency assumptions between hidden states, which imitate the observed process structure.

### Practical issues with hidden Markov trees and variants of this model

The interpretation of the hidden states relies before all on the state tree restoration, as discussed in (Durand *et al.*, 2003). This method consists in finding the most likely tree state corresponding to the observed tree. The restoration makes the zones directly apparent. Their actual meaning depends on the application and particularly on the nature of the attributes. For example, when searching for dominating paths in plants, these can be identified by extracting the sequences of consecutive states in the tree associated with large values of the diameter. Generally, different zones in a same state have equivalent attribute distributions, by definition of the HMT model. Thus, the plant is automatically segmented into comparable parts, whereas state changes highlight some ruptures.

The HMT model of (Crouse *et al.*, 1998) has also the following remarkable properties, deduced from the assumptions above:

- The privileged orientation is from the root to the leaf vertices. Thus the propagation of one hidden state  $S_u$  to its children  $c(u)$  can be seen as state splitting.
- The children states are independent given  $S_u$ . Consequently their conditional distribution is deduced from the transition matrix  $P$ .

We call this model the *independent hidden Markov out-tree (independent HMOT)*. As a consequence of the above properties, the children's conditional distribution has a property of symmetry, which seems unrealistic when ruptures caused by branching (opposed to succession) are expected.

To overcome this drawback, we propose a model where the conditional independence assumption is relaxed. We obtain the *general hidden Markov out-tree*, also oriented from the root to the leaf vertices but with dependent child states given the parent state. This model is parameterised by transition probabilities from the parent state to the set of the child states, which cannot be deduced from the individual transition probabilities between each child state and its parent.

For some applications it seems more relevant to orient the tree from the leaf vertices to the root, particularly when the attributes cannot be observed (due to diameter/secondary growth and self-pruning) on the innermost part of the plant. This leads to the *hidden Markov in-tree (HMIT)* model, parameterised by transition probabilities from the children states to the parent state. Thus, the propagation of the children states to the parent state can be seen as state merging. The three HMT models above are illustrated by figure 2.

The transition matrices of the general HMOT and HMIT are quite similar to that of the high order Markov chains. A way to reduce their high number of parameters is to use the information on the edge types (“<” and “+”) to partially order the children. In this case the transition probabilities are assumed invariant by permutation of the non-successor entities. Other practical aspects of the HMT methodology include selection of the number of hidden states and exploratory analysis. The latter requires specific methods compared to the sequence framework, due to the combinatorial issues occurring when the notion of unique descendant is lost.

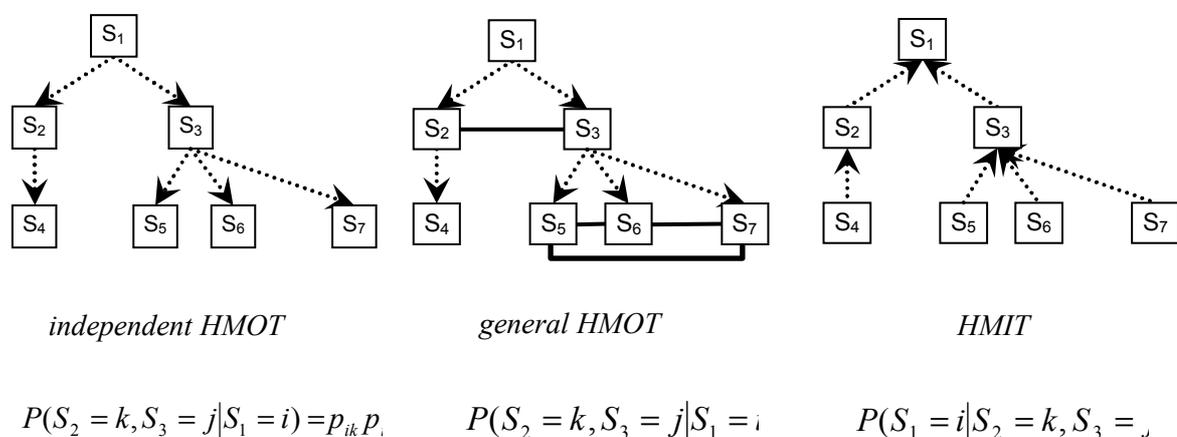


Figure 2. Variants of HMTs and their parameters.

### Perspectives of application

Two main applications are considered in a forestry/ecological and an agronomic/genetic context respectively. In fruit trees, the main aim is to describe the intra-species diversity of tree forms which interacts with both its productivity and regularity, and the easiness of its training in the orchard. Previous studies described the early stages of development of a set of cultivars of apple tree (*Malus domestica* Borkh, *Rosaceae*), exploring branching pattern along a one-year old trunk (Costes and Guédon, 2001). Further exploration was carried out on the architectural development over six years, for two genotypes, using basic statistical methods (Costes et al., 2003). The method presented herein could thus improve the modelling and characterisation of the plant structure. Furthermore, the periodicity of the flowering occurrences could be analysed at a local scale as well as globally, using the tree segmentation obtained by state restoration.

A first other potential application of zone identification within a plant is its automatic segmentation into several parts of similar nature or the extraction of remarkable paths. The homogeneous datasets obtained by such a sampling method could be analysed independently using other models. A second application is the determination of reiterations by identifying branching locations where the bearing and the borne entities are contained in a single zone. This is expected to occur when the borne axis is qualitatively similar to the bearing one.

### References

- Barthélémy D., Caraglio Y. and Costes E., Architecture, gradients morphogénétiques et âge physiologique chez les végétaux. In: *Modélisation et simulation de l'architecture des végétaux*. Bouchon J., de Reffye P. et Barthélémy D. Edt., INRA Editions (1997), 89-136.
- Crouse M.S., Nowak R.D. and Baraniuk R.G., Wavelet-Based Signal Processing Using Hidden Markov Models, *IEEE Trans. Sign. Proc.*, **46**(4) (1998), 886-902.
- Durand J.-B., Gonçalves P. and Guédon Y., Computational Methods for Hidden Markov Trees - An Application to Wavelet Trees, accepted in *IEEE Trans. Sign. Proc.* (2003).
- Ferraro P. and Godin C., A distance measure between plant architectures, *Annals of Forest Sciences*, **57** (2000), 445-461.
- Godin C., Costes E. and Caraglio Y., Exploring plant topological structure with the AMAPmod Software: an outline, *Silva Fenn.* **31** (1997), 355-366.
- Godin C., Guédon Y., Costes E. and Caraglio Y., Measuring and analysing plants with the AMAPmod software. In: *Plants to Ecosystems - Advances in Computational Life Sciences* (Michalewicz, M. T. Ed.), Vol. I, Melbourne, Australia: CSIRO Publishing (1997), 53-84.
- Guédon Y., Barthélémy D., Caraglio Y. and Costes E., Pattern Analysis in Branching and Axillary Flowering Sequences, *J. theor. Biol.*, **212** (2001), 481-520.