Analysis of the Plant Architecture via Tree-structured Statistical Models: the Hidden Markov Trees

J.-B. Durand*, Y. Guédon*, Y. Caraglio* and 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/Agro.M/CIRAD/IRD BDPPC, Equipe « Architecture et Fonctionnement des Espèces Fruitières », 2 place Pierre VIALA, 34060 Montpellier cedex 1, France

1. 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, axes and branching systems) (Barthélémy, 1991), through growth and branching processes. In addition, the plant components have been 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 axes for successive entities and according to their position for the lateral ones. These changes reflect the impact of the plant topology on the entities. They occur in various plant species in which, on the one hand, the nature of the botanical entities and that of their successors tends to be equivalent, and on the other hand, branching tends to induce marked qualitative changes between the bearing entity and the borne branching system. The differences between the entities reflect different stages of differentiation in the meristems, which are ordered in time and correspond to the notion of physiological age (Barthélémy *et al.*, 1997). In the present study, we assume that the physiological age of meristems can be assessed indirectly, *i.e.* deduced from the biological characteristics of the plant. 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 which have 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 differentiation leads to sympodial branching. The discrimination between dominant and dominated axes 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.

A statistical approach is relevant for the analysis of architectural data, both for the exploratory analysis and for inferring some embedded structures not directly apparent in the data. Statistical models are intended to make explicit some regularity, patterns or levels of organisations from the attributes, for instance tree-structured zones. The statistical analysis of sequential data from 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 hidden Markov trees (HMTs), introduced by Crouse *et al.* (1998) in the signal processing context to efficiently model homogeneous zones within a tree-structured process whose topology, fixed by the data, is thus non-random. These models are based on hidden states whose persistence, leading to homogeneous zones, is obtained by defining local dependencies between them. The HMT modelling is complementary with the plant

comparison method of Ferraro and Godin (2000, 2003), 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 in the AMAPmod software (Godin *et al.*, 1997). Following a presentation of tree-structured representations of plants, this paper develops the statistical modelling of architectural data by HMTs, relying 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 outlined.

2. Tree-structured representation of plants

As discussed in Godin *et al.* (1997) 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 a 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 short shoots borne by a given axe. Such balance between topological information within the tree-structured data and its representation at the attribute level results in 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 or axis).

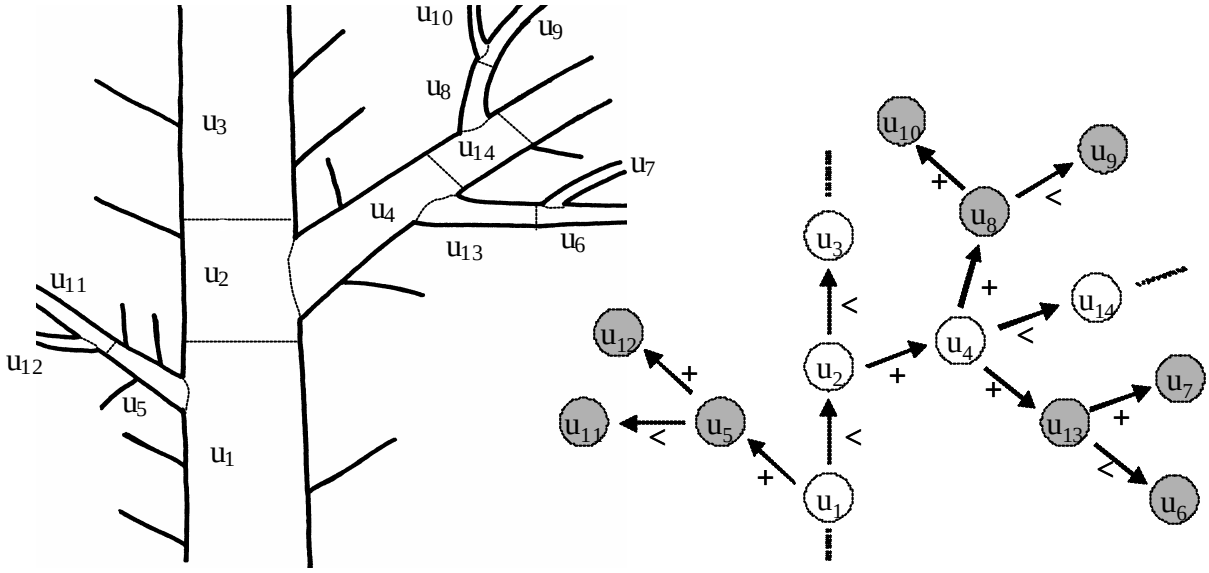
For each vertex u of the tree graph, the attribute vector is denoted by X_u and can mix qualitative and quantitative variables. The parent of u is denoted by $r(u)$ (except if u is the root vertex) and the set of children of u is denoted by $\boldsymbol{c(u)}$. If this set is empty, u is called a *leaf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

3. Modelling zones in plants with hidden Markov trees

The plant architecture organisation is modelled by affecting one state to each entity. The states are ordered and take a small number of values. Since each entity corresponds to a vertex u of the tree graph, this amounts to associate the tree representation of the plant with a state tree. The 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, such 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. The notion of order induced by the physiological age (see section 1) mostly applies at the state level. This is ensured by particular structures of the transition probability matrix $P = (p_{ij})_{i,j}$, where $p_{ij} = P(S_u = j \,|\, S_{r(u)} = i)$. The dependencies between hidden states are

essentially local; in the basic HMT model proposed by Crouse *et al.* (1998), each state in independent from all his non-descendant 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. These dependencies reproduce the structure of the observed process at the state level.



A plant represented at the growth unit scale.

Tree-structured formal representation of the plant. Part of the topological information is represented at the attribute level (*e.g.* the three short shoots borne by u_1).

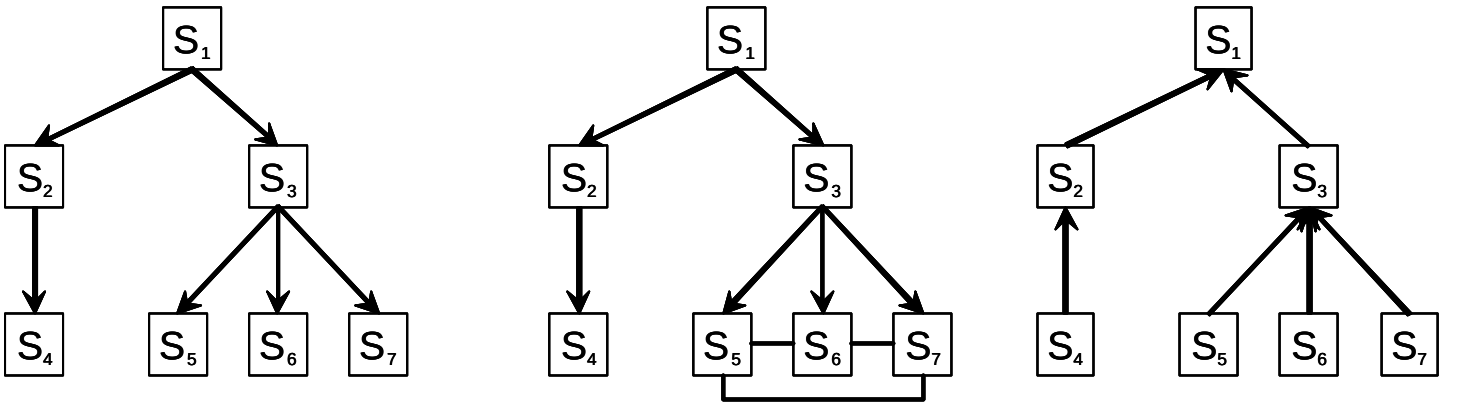
Figure 1. *Tree-structured representation of a plant.*

4. 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 state tree corresponding to the observed tree. The restoration makes the underlying 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 entity length and number of internodes. 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 where the ruptures are.

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 probability matrix *P*.



independent HMOT *general HMOT* *HMIT*

$$P(S_3 = k, S_2 = j \mid S_1 = i) = p_{ik} p_{ij} \qquad P(S_3 = k, S_2 = j \mid S_1 = i) \qquad P(S_1 = i \mid S_2 = j, S_3 = k)$$

Figure 2. *Classes of HMTs and their parameters.*

We call this model the *independent hidden Markov out-tree (independent HMOT)*. As a consequence from the above properties, the children conditional distribution has the following permutation property: let assume that vertex 1 has children set $\{2;3\}$. Then $P(S_3 = k, S_2 = j \mid S_1 = i) = p_{ik} p_{ij} = P(S_3 = j, S_2 = k \mid S_1 = i)$, 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 children states given the parent state. This model is parameterised by transition probabilities from the parent state to the set of the children states $P(S_3 = k, S_3 = j \mid S_1 = i)$, different from $P(S_3 = j, S_3 = k \mid S_1 = i)$.

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 cambial growth and self-pruning) on the most inner 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 are illustrated in 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 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.

5. Perspectives of application

The following main applications are considered, both in a forestry/ecological and an agronomic/genetic context. In fruit trees, one objective is to describe the intra-species diversity of tree forms which interacts with both its productivity and regularity, and its training easiness 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 one-year-old trunk (Costes and Guédon, 2002). Further exploration was carried 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 flowering occurrence could be analysed at a local scale as well as globally, using the tree segmentation obtained by state restoration.

A widely applicable use of HMTs is zone identification within a plant for its automatic segmentation into several parts of similar nature or for the extraction of remarkable paths. This is especially useful when the amount of available entities per plant is large, which is typically the case in the context of forest trees, which crown is usually huge and complex. The homogeneous datasets obtained by such a sampling method could be analysed independently using other statistical models. A second general application of our approach is the determination of reiterated complexes by identifying branching locations where the bearing and the borne entities are affected to the same state. This is expected to occur when the borne axis is qualitatively similar to the bearing one.

More generally, the segmentation of the plants into a small number of states expectedly offers new possibilities for quantifying the physiological age. As a perspective, this should lead to a methodology for the validation of this notion.

References

Barlow, P.W., From cell to system: repetitive units of growth in the development of roots and shoots. *Growth Patterns in Vascular Plants.* M. Iqbal Ed. (1994), 19-58.

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.

Barthélémy, D., Levels of organization and repetition phenomena in seed plants. *Acta Biotheoritica*, **39** (1991), 309-323.

Costes E. and Guédon Y., Modelling Branching Patterns on 1-year-old Trunks of Six Apple Cultivars. *Annals of Botany*, **89** (2002), 513-524.

Costes E., Sinoquet H., Kelner J.J. and Godin C., Exploring Within-tree Architectural Development of Two Apple Tree Cultivars Over 6 Years, *Annals of Botany*, **91** (2003), 91-104.

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., Gonçalvès 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.

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

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.