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

J.-B. Durand[1], Y. Guédon[2], Y. Caraglio[2] and E. Costes[3]

[1]Institut National Polytechnique de Grenoble, Laboratoire de Modélisation et Calcul/IMAG, BP 53, 38041 Grenoble cedex 9, France; [2]Unité Mixte de Recherche CIRAD/CNRS/INRA/IRD/Université de Montpellier II, Botanique et Bioinformatique de l'Architecture des Plantes TA40/PS2, 34398 Montpellier cedex 5, France; [3]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

Author for correspondence:
Jean-Baptiste Durand
Tel: +33 4 76 63 57 09
Fax: +33 4 76 63 12 63
Email: Jean-Baptiste.Durand@imag.fr

Received: *30 September 2004*
Accepted: *14 December 2004*

Summary

• Plant architecture is the result of repetitions that occur through growth and branching processes. During plant ontogeny, changes in the morphological characteristics of plant entities are interpreted as the indirect translation of different physiological states of the meristems. Thus connected entities can exhibit either similar or very contrasted characteristics.

• We propose a statistical model to reveal and characterize homogeneous zones and transitions between zones within tree-structured data: the hidden Markov tree (HMT) model. This model leads to a clustering of the entities into classes sharing the same 'hidden state'.

• The application of the HMT model to two plant sets (apple trees and bush willows), measured at annual shoot scale, highlights ordered states defined by different morphological characteristics. The model provides a synthetic overview of state locations, pointing out homogeneous zones or ruptures. It also illustrates where within branching structures, and when during plant ontogeny, morphological changes occur.

• However, the labelling exhibits some patterns that cannot be described by the model parameters. Some of these limitations are addressed by two alternative HMT families.

Key words: apple tree, bush willow, clustering of tree-structured entities, differentiation of meristems, hidden Markov tree models, plant architecture modelling, tree segmentation.

New Phytologist (2005) **166**: 813–825

© *New Phytologist* (2005) **doi**: 10.1111/j.1469-8137.2005.01405.x

Introduction

For many years, plant architecture has been viewed as the result of repetitions (Barlow, 1994), which occur at different levels of organization (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 rules (Barthélémy *et al.*, 1997). The changes that occur during plant ontogeny have been described along axes for successive annual shoots, and according to their position for lateral shoots. These changes reflect the impact of plant topology on the potentiality of annual growth. The differences between entities were interpreted as different stages of differentiation of the meristems, which are ordered in time and correspond to the notion of physiological age (Nozeran *et al.*, 1971; Gatsuk *et al.*, 1980; Barthélémy *et al.*, 1997). In the present study, we assume that the physiological age of meristems can be assessed retrospectively, that is, deduced from the morphological characteristics of the plant entities within the tree structure. We aim to characterize these changes by diverse morphological variables attached to a given entity, such as number of nodes, length, and presence/absence of flowering. These variables are called entity attributes. Connected entities

that have similar attributes can be interpreted as homogeneous zones, as opposed to transitions between zones, corresponding to abrupt changes in the values of the attributes. For example, flowering is a factor of rupture in the plant architecture when meristem differentiation leads to sympodial branching. The discrimination between dominating and dominated axes, in sympodial plants with different degrees of hierarchy, can also be formulated as the search for ruptures and continuities. Indeed, the architectural concept of reiteration (Barthélémy, 1991) corresponds to axes or branching systems with a same degree of hierarchy. More generally, it makes sense to identify zones where the entities, at a given scale, can be classified clearly into a small number of classes defined by different morphological and functional characteristics. Such classification can lead biologists to identify when during plant ontogeny, or where in branching systems, morphological changes are significant enough to be discriminated by statistical studies. This can provide objective criteria for the design of sampling procedures within tree crowns, especially for physiological investigations that need to target tissues or organs in specific states (e.g. flowered vs not flowered).

A statistical approach is relevant for the analysis of architectural data, both for exploratory analysis, and for inferring some embedded structures that are not directly apparent in the data. Statistical models are intended to make explicit some regularity, patterns or levels of organization from attributes, for instance tree-structured zones. The statistical analysis of sequential data from plant architecture, illustrated by Guédon *et al.* (2001), is mainly based on Markovian models, for instance hidden semi-Markov chains for modelling homogeneous zones. These models, although 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, as the dependencies between entities of disjoint sequences are eluded. The complete topology has somehow to be included in the model for the existence of multiple dependent successors (or descendants) to be considered in the distribution of zones.

We propose to use the statistical framework of the hidden Markov tree (HMT) model, introduced by Crouse *et al.* (1998) in the signal-processing context, to model homogeneous zones efficiently within a tree structure whose topology is fixed in the data. These models are based on hidden states whose persistence, which leads to homogeneous zones, is obtained by defining local dependencies between the states attached to adjacent entities. The HMT modelling is complementary to the plant-comparison method of Ferraro & Godin (2000, 2003), based on an edit distance between tree-structured data. This edit distance integrates the comparison of topology and that of the attributes. Instead, our method determines zones with common attribute distribution, the plant topology being taken into account locally by the dependencies between one entity and the adjacent ones. Markovian models for tree-structured data, as well as the edit distance between tree-structured data, have been integrated in the AMAPmod software (Godin *et al.*, 1997).

The labelling of the tree entities using the model states, the 'state tree restoration', provides a synthetic overview of the state locations. The plant is automatically segmented into comparable parts, whereas state changes highlight where the ruptures are. Moreover, the labelling procedure visually reveals some features that cannot be described explicitly by the model parameters (macroscopic tree-structured patterns, influence of branching on state succession, etc.). These features can be analysed, *a posteriori*, from the restored state tree (extraction of counts, frequencies of occurrences, etc.).

Following the presentation of tree-structured representations of plants, the statistical modelling of architectural data by HMT models is developed in this paper, relying on the aforementioned botanical concepts and hypotheses. Some practical aspects of the application of the HMT model to botanical data are addressed. The importance of HMT modelling is illustrated through applications in agronomy and ecology. Finally, different families of HMT models and other potential extensions of the model are presented, and perspectives of other applications are outlined.

Materials and Methods

Tree-structured representation of plants

As discussed by Godin & Caraglio (1998), plant topology can be described formally through rooted multiscale tree graphs, the vertices of which correspond to their constituent botanical entities, and the edges of which represent the physical connections between them. Each scale corresponds to a more-or-less macroscopic viewpoint on the plant. As only single-scaled tree graphs can be analysed using HMT models, it is necessary to choose a scale for the plant description. Part of the topological information contained 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 axis. Such a balance between topological information within the tree-structured data and its representation at the attribute level relies on modelling choices. This is why the plant is typically represented at a macroscopic level, lower than the internode scale (growth unit, annual shoot or axis). Moreover, the changes that occur during plant ontogeny, which are those of interest in the present study, should be (or are assumed to be) more easily revealed at macroscopic scales, such as the annual shoot scale, than at finer scales. More generally, these changes are directly related to the level of organization at which growth periodicity is expressed.

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 $c(u)$. If this set is empty, u is called a leaf vertex. The different types

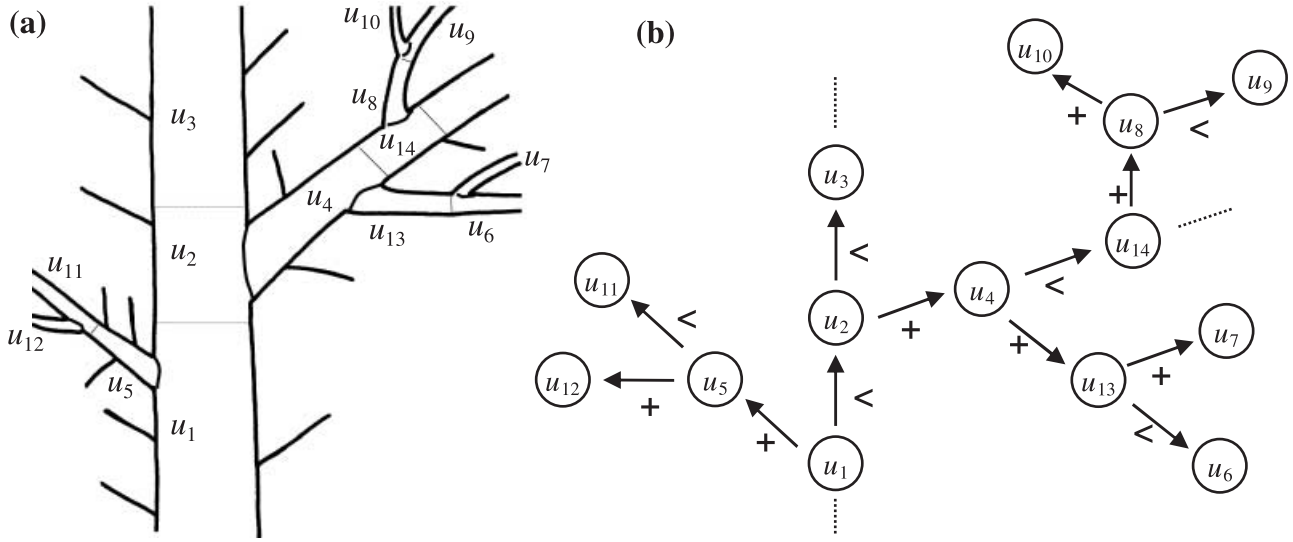

Fig. 1 Tree-structured representation of a plant. (a) Plant represented at growth unit scale. (b) Tree-structured formal representation of the plant. Part of the topological information is represented at the attribute level only (e.g. the three shoots borne by u_1).

of connection of the plant entities are represented by typed edges: < for succession; + for branching. These notations are illustrated in Fig. 1.

Modelling homogeneous zones in plants with HMT models

The plant architecture is modelled by assigning one state to each entity. This state represents the class of the entity. Each class contains entities that have similar attributes. The states are ordered (at least partially), and take a small number of values. As each entity corresponds to a vertex u of the tree graph, this is equivalent to associating 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 assigned to a given state defines a homogeneous zone, whereas connected vertices assigned to different states induce ruptures in the plant architecture. The propagation of the states within the plant is related to its topological organization. This can be modelled in a probabilistic framework by the HMT models.

The HMT models were introduced by Crouse *et al.* (1998) for modelling the dependencies and heterogeneity into a tree-structured process. 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, and hence the zones, are distributed. The notion of order induced by physiological age mostly applies at the state level. This is ensured by particular structures of the transition probability matrix $P = (p_{ij})_{i,j}$, where $p_{ij} = P(S_u = j \mid S_{\rho}(u) = i)$ represents the probability of switching from state i in the parent vertex $\rho(u)$ to state j in vertex u. The dependencies between hidden states are essentially local; in the basic HMT model proposed by Crouse *et al.* (1998), the state of vertex u depends on the state of its parent vertex only. This local dependency assumption gives its name to the Markov property for trees. The HMT model is quite close to the hidden Markov chains used for sequence or time-series analysis (Ephraim & Merhav, 2002). 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.

Practical issues with HMT models

The application of the HMT model to botanical data is broken down into two successive steps, as discussed by Durand *et al.* (2004). The first is parameter estimation from the measured entities; the second is state tree restoration. Let n denote the number of entities and K the number of states. Parameter estimation is based on an iterative method (instance of the EM algorithm). Following parameter estimation, state tree restoration is used for model interpretation and validation. Its purpose is the search for the most likely state tree corresponding to the entities. The restoration procedure takes into account the dependencies between connected vertices, as the model does. During this procedure, one of the K classes is assigned to each entity, based on the estimated model and the measurements for the entity.

The state tree 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 from the tree sequences of consecutive states associated with large values of the entity length and of the 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 (e.g. see Fig. 3a,c)

Moreover, the model can be assessed and interpreted by considering the fit between the marginal distribution of one variable and the corresponding empirical distribution (histogram of the data for this variable), as illustrated in Fig. 5. The restored state tree can be used for visualization of the fit between the observation distribution for each state j, and the histogram of all entities in state j (see Fig. 2).

The HMT model of Crouse *et al.* (1998) also has the following remarkable properties, deduced from the assumptions above.
(1) 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.
(2) The children states are independent, given S_u.
We call this model the independent-children hidden Markov out-tree (HMOT) model.

Because of the strong assumptions above, complex tree-structured patterns and long-range dependencies cannot be captured explicitly by this HMT model. In this case, the restoration procedure may reveal such features, which can be characterized quantitatively by the analysis, *a posteriori*, of the restored state tree (computation of counts, frequencies of occurrences, etc.).

Other practical aspects of the HMT methodology include selecting the number of hidden states. The actual number of states is determined using statistical criteria. The chosen criterion depends on the aim of the analysis. The Bayesian information criterion (BIC) is used frequently to determine the number of hidden states (Geiger *et al.*, 2001), although its properties have not been established in this context. This criterion is intended to assess the compromise between model fit to the data and parsimony. Alternatively, one may favour an easy interpretation of the model in the selection stage. This leads us to select the model that maximizes a compromise between state separation and model fit. This is the purpose of the integrated classification likelihood criterion proposed by Biernacki *et al.* (2000).

Applications to botanical data

The following applications are considered in both an agronomic/genetic and a forestry/ecological context. Both species studied (apple tree and bush willow) exhibit a sympodial branching which occurs, respectively, after terminal meristem flowering and death. Thus the main biological question addressed here concerns the impact of sympodial development on subsequent growth, and its role in the organization of the whole tree. A secondary aim is to address issues of synchronism within entire branching systems. In apple tree, there is a particular focus on flowering occurrences and their alternation with vegetative growth. In bush willow, the aim is to test the assumption of ordered morphological changes in a whole sympodial branching system, and to analyse the impact of climatic conditions on this order.

Apple tree – Context and aim of the analysis In fruit trees, one objective is to describe the intraspecies diversity of tree forms and branching patterns, which interact with productivity, regularity and ease of training in the orchard (Lauri *et al.*, 1997). Previous modelling approaches have described the early stages of development of a set of cultivars of apple tree (*Malus domestica* Borkh., *Rosaceae*), exploring branching patterns along 1-yr-old trunks (Costes & Guédon, 2002). Further exploration into the architectural development over 6 yr was carried out for two genotypes, using basic statistical methods (Costes *et al.*, 2003). Although similar growth and flowering behaviours were demonstrated for all the branches within the tree crown, whatever their location, these studies did not consider the patterns of flowering occurrence at the tree scale. The method presented here aims to improve the approach to modelling plant structure by extending the previous models, which were carried out at local scales and focused on the branching process, towards characterization of the whole plant structure.

Plant material Two trees per scion cv. Fuji, grafted on Lancep Pajam 1 (type M9), were described in 1999 at Melgueil INRA experimental station (south-east France), when the trees were 6 yr old. In short, each tree was broken down into three scales of organization corresponding to the axes; growth units (GU); and metamers (as defined by White, 1979). Four GU types were considered: long GU (labelled U) >20 cm long; medium GU (W) ≤20 cm but >5 cm long; short GU (D) ≤5 cm long; and a fourth GU type corresponding to the floral GU or 'bourse' (I), which results from floral differentiation of the apical meristem. Bourse shoots can develop into short, medium or long GU, and were categorized in the same manner as the other vegetative GUs. Thus in springtime the three vegetative GUs (long, medium and short) can develop either from the terminal bud, if this has not differentiated into a bourse, or as a bourse shoot. Metamers were counted on the long and medium GUs only, while the short GUs were not broken down at metamer scale in order to simplify the observations.

Spatial coordinates and diameters were collected at the metamer scale, each five leaves along the long GU, and at the top of the axis for short axes. Coordinates were collected using 3SPACEFASTRACK (Polhemus Inc.) and 3A software (Adam *et al.*, 1999). From the database, which combined both topological and geometrical observations, 3D reconstructions of the trees were obtained using the AMAPmod software (see Fig. 3).

Choice of scale and extraction of attributes In the following analysis the apple tree was considered at the GU scale in order to investigate the alternation between flowering and vegetative GU. The GUs whose growth stopped in 1998 (or before) were removed from the analysis. The selected attributes, at this scale, were the number of metamers per GU, and the presence or absence of a flower on the GU. This choice of GU scale was also motivated by a compromise between the number of vertices, the complexity of the tree topology, and the diversity of the attributes.

Statistical modelling We considered bivariate HMT models with parametric observation distributions for the number of metamers. These are discrete distributions, appropriate for count data, chosen among the binomial, negative binomial and Poisson distributions. Bernoulli distributions were chosen for the presence/absence of flowers. Both variables were assumed to be independent, given the hidden state.

Bush willows – Aims of the analysis This study aimed to demonstrate structured growth expression within bush willows submitted to various growth conditions. In order to reach this goal, we wanted to reveal dominating paths or branching systems, as opposed to equivalent ones.

Bush willow grows according to sympodial branching: terminal flowering is expressed on growth units (referred to as modules; Bell, 1991). Lateral axes (one or more) relay the previous module, and each one develops a new growth unit and flowers terminally. This phenomenon is repeated each growth season, thus the extension of plant structure is essentially due to the branching process (flowering axes or succession of modules). Some of these modules become stronger than others, and as time goes by a 'trunk' can be identified (dominating path). At any moment of the tree's life, the upper part of the tree (crown) is constituted of many modules which are more-or-less equivalent. In this part of the tree, entities were sampled and our modelling approach applied in order to obtain automatic labelling of the entities. Based on this labelling, we aimed to compare different populations, characterized by their growth conditions, and to propose a classification of entire individuals.

Plant material The data were collected in Mali by Bonnet (2002), in the Bamako area, on 111 individuals of *Combretum adenogonium* Steud. ex. A. Rich. (*Combretaceae*). The trees were sampled under various natural conditions and in different places. Each tree was characterized by diameter breast height; total height; type of stress or injury (fire, cutting, pruning); and some ecological descriptors (competition index, number of lianas).

On each tree, one 4-yr-old crownlet was sampled, and after cutting all the modules of the branching system were described. For this description, two levels of organization were considered: module (labelled A) and metamer (E). As a consequence of the modular development, only the link + was used at the module scale to describe the branching system. On each module, quite simple and directly available parameters were measured: total length, basal diameter, number of internodes, number of photosynthetic leaves, phyllotaxis, number of flowers, and bark colour.

Choice of scale and extraction of attributes The module, which is the level of expression of growth periodicity, was chosen in order to investigate the changes in meristematic activity according to successive branching. The attributes considered were number of leaves and length of the module. Diameter was not considered, as this attribute reflects both the age of the entity and the global functioning of the branching system rooted at this entity. For some modules the number of leaves, or the length, was not available: in this case all the branching system that originated from the module was omitted from the analysis. For 12 individuals of our data set, such missing measurements occurred at the base of the trunk. Consequently, these individuals were removed from the data set, and only 99 individuals were considered for the analysis.

Statistical modelling The maximal number of states considered in the model building was fixed at six. The observation distributions were assumed to belong to the same parametric families as for apple tree. The actual number of states was determined using the BIC, which focuses on the model fit.

Results

Apple tree

The model selected by the integrated classification likelihood criterion has four states, while the BIC selected a five-state model, interpretation of which is far more difficult. Both criteria show that a model with more than six states is not relevant, and the interpretation of the six-state model is rather tedious. Thus the results of the five-state model are not detailed here. State 0 (denoted L) is characterized by a high value of the number of metamers per GU (Fig. 2; observation distribution for state L is represented by a solid line). State 1 (M, dashed line) is characterized by a medium value; state 2 (S, dotted and dashed line) corresponds to GUs with a single metamer. The probability of flowering, given these states, is very close to zero. Hence these first three states, which correspond to vegetative GUs, are clearly ordered by the mean (and, incidentally, the variance) of their observation distributions. In contrast, state 3 (F, dotted line) is characterized by the systematic presence of flowers, and by a very low number of metamers per GU (fewer than six and generally fewer than three metamers; Fig. 2).

The state tree restoration algorithm was applied to obtain an automatic segmentation of the two apple trees, and also to

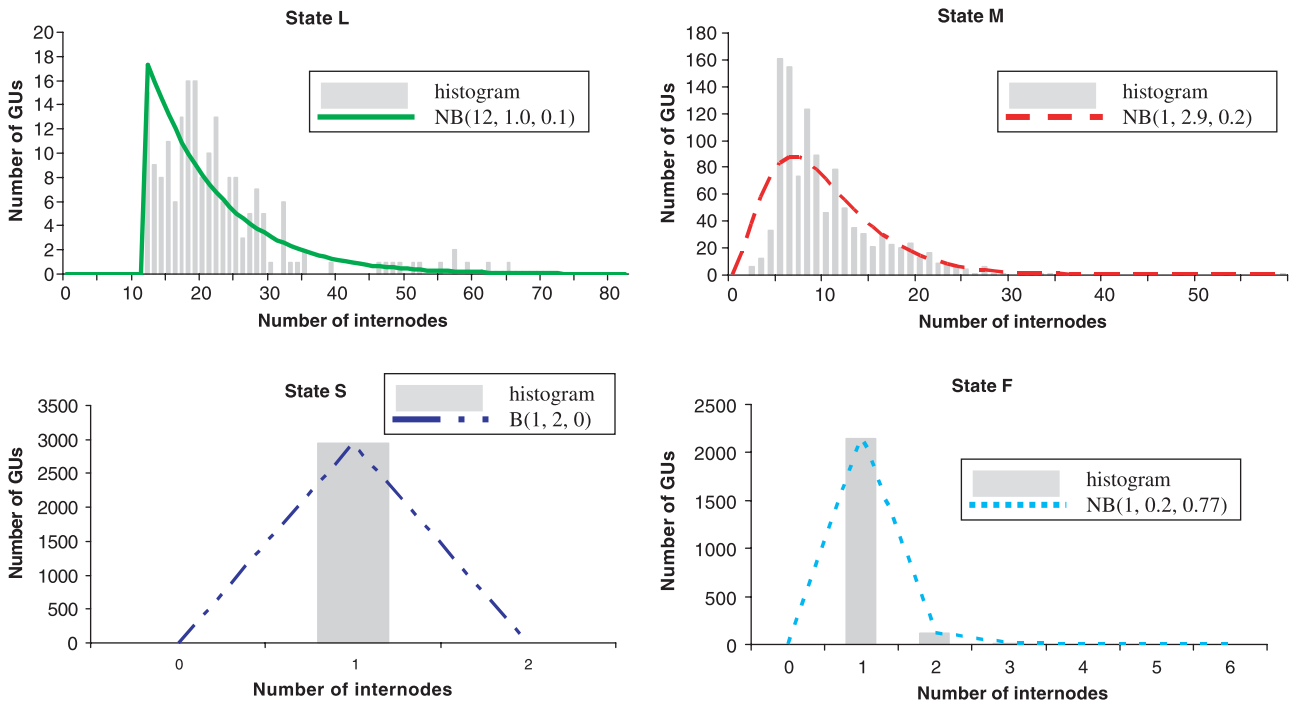

Fig. 2 Definition of the four states in the hidden Markov out-tree model for apple tree cv. Fuji. Each state is characterized by one observation distribution for the number of internodes per growth unit (GU). Binomial distributions, B; negative binomial distributions, NB. The type of distribution is followed by the values of the estimated parameters. Each observation distribution for state *j* is compared with the histogram extracted from the data assigned to state *j* (by the state tree restoration).

assess and interpret the model by matching, for each state *j*, the theoretical observation distribution for state *j* (deduced from model parameters) with the empirical one (Fig. 2). The empirical distribution is extracted from the data assigned to state *j* by the state tree restoration stage. In the restoration step, one of the four classes is assigned to each GU, based on the estimated model and the measurements for the GU (number of metamers and flowering). The restoration procedure, as well as the model, takes into account the dependencies between connected GUs.

Moreover, the state tree restoration provides a synthetic view of the states' locations. This is shown in Fig. 3a. For the sake of clarity, only the trunk and one of the branches are represented. State L (green) is mostly located at the base of the trunk and of the main branches. State M (red) typically corresponds to the distal part of the trunk and of the main branches. States S (dark blue) and F (light blue) follow states L and M, respectively, and then tend to alternate. The empirical and predicted distributions of the number of metamers per GU for each state are represented in Fig. 2. State L is represented by a solid line in Fig. 2 and is green in Fig. 3a; state M by a dashed line in Fig. 2 and red Fig. 3a, etc.

Information concerning the succession of states within the tree is summed up quantitatively in the transition probability matrix (Table 1). The initial state is state L. The selected model has a single recurrent class, although the transition probability matrix has a particular structure (Fig. 4).

Table 1 Transition probability matrix

	u			
ρ(*u*)	L	M	S	F
L	0.05	0.15	0.63	0.16
M	0.02	0.06	0.30	0.62
S	0.01	0.05	0.27	0.66
F	0.04	0.35	0.60	0.00

The value at line *i* and column *j* represents the probability of a transition from state *i* – 1 to state *j* – 1. States L, M and S are characterized by a high, medium and low number of metamers per GU, respectively; F by presence of flowers.

The transition probabilities do not consider both types of edge (succession and branching) separately in the independent-children HMOT model. However, the restoration of the state trees offers the possibility of estimating the transition probabilities, given the type of edge. Subsequent to the state restoration, the frequency of each possible transition from a vertex to a successor vertex (<) has been computed, as well as the number of parents (of successor descendants) whose restored state is *j* (for each state *j*). The estimated transition probability matrix is represented in Table 2. Then the frequency of each possible transition from a vertex to a branching vertex (+) has been computed. The estimated transition probability matrix

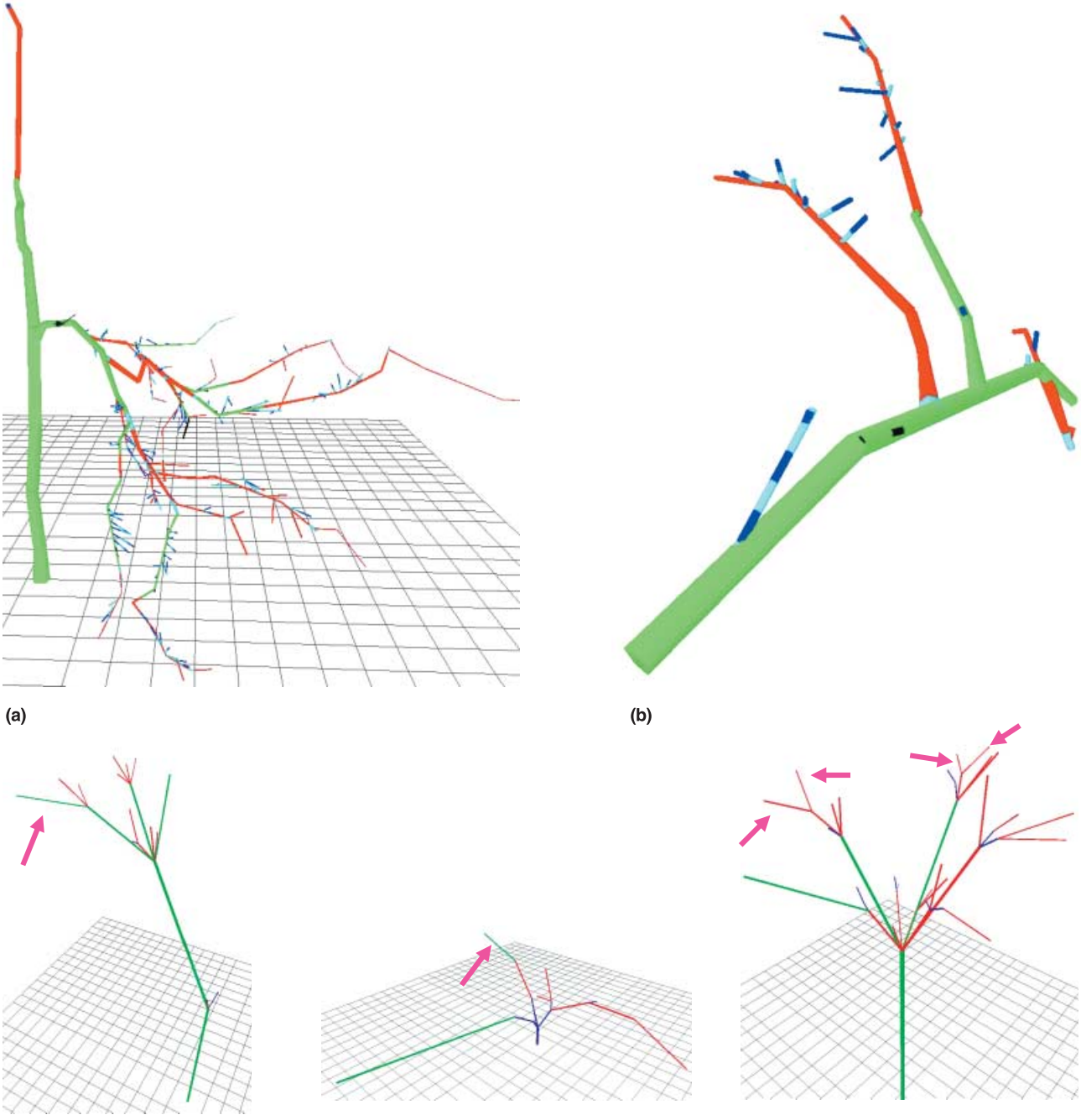

Fig. 3 (a) Restored state tree for the apple tree data set. For the sake of clarity, only the trunk and one of the branches are represented. Each growth unit (GU) is coloured according to its state. The green state (L) is characterized by a high number of metamers per GU; the red state (M) by an intermediate number of metamers; the dark blue state (S) corresponds to GUs with a single metamer. Flowering does not occur in any of these three states. In contrast, the light blue state (F) is characterized by the systematic presence of flowers, and by a very low number of metamers. (b) Apple tree data set: alternation between short flowering shoots (light blue) and short vegetative shoots (dark blue). (c) Restored state trees for the bush willow data set. Individuals from classes 1–3 (from left to right, respectively). State L, green (high length and number of leaves); state M, red (intermediate length and number of leaves); state S, blue (low length and number of leaves). Arrows, extremities of dominating paths starting from the base of the tree.

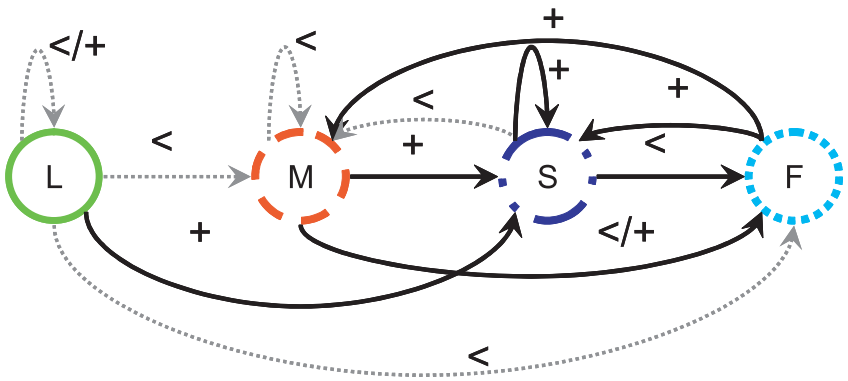

Fig. 4 Transition graph with information on succession or branching between successive states. The dotted arrows correspond to transition probabilities <0.3. Only the transitions with probability >0.05 are represented. Over-represented transitions are denoted by < for succession; + for branching; </+ denotes the absence of over-representation. Codes for states are as in Figs 2, 3a: state L, solid line, green; state M, dashed line, red, etc.

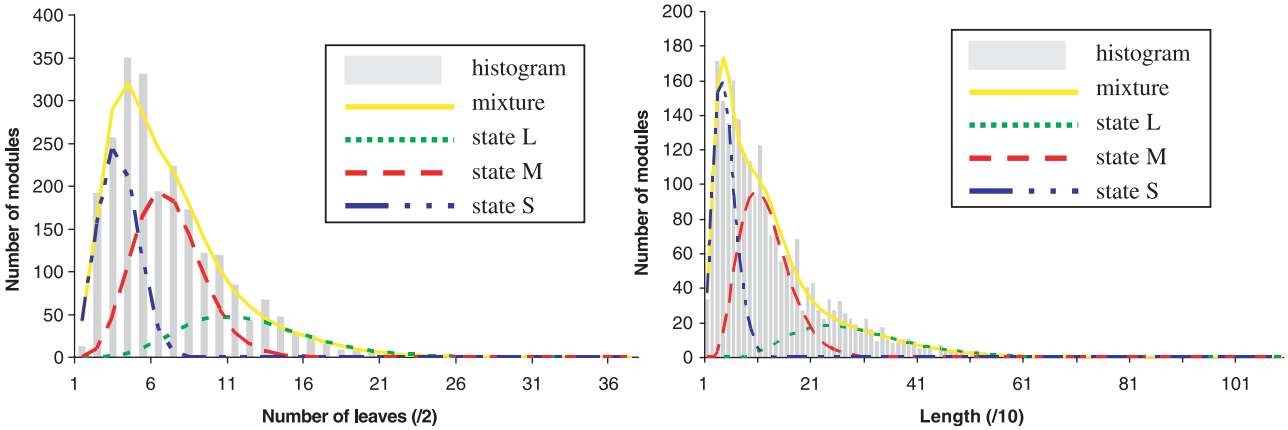

Fig. 5 Definition of the three states for the *Combretum* HMOT. State L, dotted line; state M, dashed line; and state S, dotted-dashed line are characterized by their observation distributions. Marginal distributions of the number of leaves and module length are a mixture of the three observation distributions deduced from the model parameters (solid line). For both variables the marginal distribution is compared with the empirical distribution (histogram).

Table 2 Transition probability matrix, conditional on the child entity being a successor

	L	M	S	F	Count
L	0.09	0.44	0.01	0.46	167
M	0.03	0.12	0.02	0.84	554
S	0.01	0.05	0.27	0.67	1337
F					0

The last column represents the total number of parent entities < (succession) in each state.

Table 3 Transition probability matrix, conditional on the child entity being a branching offspring

	L	M	S	F	Count
L	0.05	0.11	0.75	0.09	1128
M	0.01	0.03	0.40	0.56	1365
S	0.27	0.00	0.63	0.09	11
F	0.04	0.35	0.61	0.00	2045

The last column represents the total number of parent entities + (branching) in each state.

is represented in Table 3, together with the number of parents (of descendants borne) in each state.

When both matrices are compared, the results (Fig. 4) highlight certain under/over-represented transitions according to edge type. Homogeneous zones of GUs in state L or M tend to follow each other, while a transition to state S is related to branching. The separation between states L and M is not so clear from this viewpoint. Furthermore, state F tends to follow state S by succession, whereas state S systematically follows state F by branching. This is related to the sympodial development of the apple trees. Some extensions (see Discussion) allow the distinction between < and + to be voluntarily included in the model.

Bush willow

The best compromise between fit and parsimony is achieved by an HMT with three recurrent states. State 0 (L) is characterized by a high value of the length and number of leaves per module (Fig. 5, state L represented by a dotted line). State 1 (M, dashed line) is characterized by intermediate values, and state 2 (S, dotted and dashed line) by low values for both

Table 4 Transition probability matrix for the bush willow data set

	L	M	S
L	0.31	0.45	0.25
M	0.08	0.47	0.45
S	0.09	0.46	0.45

variables. Thus the three states can be ordered from the mean of their observation distributions.

One way of assessing the model is to consider the fit between the marginal distribution of the variables, computed from model parameters, and the corresponding empirical distribution, extracted from the data. The marginal distribution is a mixture of the three observation distributions. The weights of the mixture are estimated through the empirical frequency of each state, based on the state restoration. This is illustrated in Fig. 5, which shows the fit between the mixture (distribution predicted from the model) and the histogram (empirical distribution). It appears that, from a state-based viewpoint, the information contained in both variables is redundant. This is shown by the empirical correlation coefficient between number of leaves and length, which is 0.819. Moreover, the states appear to be quite well separated.

The state tree restoration algorithm has been applied to obtain an automatic segmentation of each bush willow (Fig. 3c). Examination of the restoration shows that the three states are roughly ordered spatially, as state M tends to follow state L, and state S to follow state M, when progressing toward the extremities. However, this remarkable succession, although roughly valid at the scale of the whole plant, is not so clear at a local scale. As 99 individuals have been used to estimate the model parameters (the 12 remaining individuals had missing data at the base of the trunk and were omitted), and as the HMT model handles only local dependencies, the transition matrix (Table 4) does not exhibit any left–right structure. However, any return from state M or S to state L is quite rare.

The selected model is made up of a single recurrent class, and state L is the initial state. Given the interindividual heterogeneity, which hides the expected succession of the states in the trees, we performed a classification of the individuals based on deterministic criteria of the transitions between the restored states. The five following classes were considered:

(1) no transition to a lower state occurs;

(2) at least one S→M transition occurs, followed by an M→L transition;

(3) at least one S→M transition and an M→L transition occur, at no particular position;

(4) only S→M transitions or transitions to a higher state occur;

(5) only M→L transitions or transitions to a higher state occur.

Typical individuals belonging to the first three classes are represented in Fig. 3c, jointly with the state tree restoration.

The analysis of dominating paths or branching systems, as opposed to equivalent ones, is based on the state tree restoration. As the states are ordered by means of conditional distributions, we used a heuristic function to compute the cost of each possible path in the plant, where the small values of the states and the basal locations of these values are favoured. The result of this algorithm is illustrated in Fig. 3c. If several dominating paths exist, we obtain equivalent paths as a by-product, as in the right-hand sample in Fig. 3c. For the determination of partially equivalent or subdominating paths, the same process can be iterated after deleting the extremities of the optimal paths.

In both applications the parameter estimation does not take longer than 5 min on a computer with a Pentium IV processor, and the state tree restoration is immediate.

Discussion

Interpretation of the model

In both applications considered, HMT modelling provides a clustering of the plant entities that can easily be interpreted as an index of vigour of the entities as the states are ordered: by the means of distributions of length; number of leaves; or number of internodes. In the case of apple trees, an *a priori* classification of the type of growth unit was available with the data, according to the length of the GUs. The aim is to validate this *a priori* classification of GUs by comparing it with the hidden states (which are based on the number of internodes instead of length), as shown in Table 5. Class U is expected to match with state L; class W with state M; class D with state S; and class I with state F. A close match between classes and states is achieved for classes D and I. However, classes U and W, although in agreement with the states in most cases, tend to be poorly separated; especially class U, which is distributed equally between states L and M. This mixing character of the model concerning the medium and long shoots only is in accordance with the properties of the HMT model illustrated in Fig. 2. On the one hand, the variable 'absence/presence of flower' (vegetative/flowering) allows the identification of flowering shoots; on the other hand, the short shoots can be identified without ambiguity with reference to medium or long shoots.

Table 5 Number of growth units of each class, given the hidden state

	Class			
State	U	W	D	I
L	136	55	0	0
M	132	939	14	1
S	4	35	2902	0
F	0	0	0	2291

States L, M, S: high, medium and low number of metamers per GU, respectively; state F: floral GU; classes U > 20 cm; 5 cm < W ≤ 20 cm; D ≤ 5 cm; class I: bourse.

The succession of the hidden states can be studied through the state tree restoration. A quantitative viewpoint on the succession of states is achieved by the analysis of the transition matrix. In the case of apple trees, the short flowering shoots and the short vegetative shoots clearly tend to alternate (Table 1). This is also striking through the state restoration (Fig. 3b). This behaviour is typical of the development of apple trees, where flowering occurs on terminal positions along axes. The flowers are followed by a vegetative shoot, which is sympodial, develops immediately, and is called a 'bourse shoot' (Crabbé & Escobedo, 1991). The following year flowering is located on this bourse shoot, hence the alternation of flowering and vegetative shoots.

However, the aspects of the hidden state succession that are specific to a given type of edge are not correctly accounted for in the HMT model. This is illustrated by Fig. 4: the edge type associated with each possible transition (if any) is obtained only indirectly from the model, through the state restoration. This result is partly specific to the sympodial development of apple trees after flowering. However, this is in accordance with the modelling proposed in the Introduction: branching corresponds to specific transitions (from any state to state S, or from state F to any state), while succession is associated with the repetition of entities or with transitions to a close state, which creates homogeneous zones.

Furthermore, the underlying Markovian model assumes that the child state is independent of its grandparents, given the father state. From this viewpoint, this model is similar to a first-order Markov chain. As a consequence, the remarkable tree-structured patterns visible in Fig. 3b (the transition to a 'short shoot' state followed by the alternation of several vegetative or flowering shoots) cannot be modelled precisely. The existence of such patterns could be accounted for only by dependencies between consecutive ancestors within a fixed range.

In both applications the HMT modelling emphasizes the existence of ordered states, as each following state represents a distribution with either lower vigour (number of leaves or internodes) or an ultimate stage of development (flowering). As stated in the Introduction, the different distributions can be interpreted as an underlying stage of differentiation: the physiological age of the meristems. Hence the states are a natural way to quantify physiological age, and this quantification now remains to be validated. In the bush willow data set the states clearly represent ordered levels of meristem potentiality. These levels still need to be crossed with the growth conditions, and particularly the type of stress or injury, before validating the notion of physiological age.

Given the above interpretation of the hidden states, we would expect the presence of absorbing states within the HMT models in both applications. These states would correspond to an ultimate state in which the local characteristics of the plant would not change any more. Models that include such absorbing states have been estimated from the data sets, but no statistical evidence in favour of these models has been revealed from the model selection criteria. However, in the case of apple trees the alternation between short vegetative and flowering GUs, highlighted by the state tree restoration, can be interpreted as an absorbing set of states (in a figurative sense). This set is quite stable and is reached after only a few transitions. The absence of any absorbing individual state could then be related to the scale considered (GUs).

In the case of bush willow, a first important result is that only three individuals among 99 do not exhibit any particular arrangement of the three modelled states, despite the very varied growth conditions of each tree. This reinforces the assumption of ordered stages of meristem differentiation. Moreover, some individuals that were struck by fire exhibit a succession of states in reverse order (middle part of Fig. 3c). This succession forbids the presence of an absorbing state, and can be interpreted as an intermediate phase before the resumption of plant development. In this study we can also reveal and characterize equivalent paths in branching systems, which correspond to the architectural concept of reiteration (Barthélémy, 1991). Finally, this modelling approach could provide a new framework for biological analysis of the climatic conditions and/or events that lead to classification of individuals on the basis of transitions between restored states.

Extensions to the proposed HMT model

Here we show that the HMT models allow homogeneous zones and ruptures to be identified (based on measured morphological characteristics of the entities), through hidden states that represent different and ordered distributions of these local characteristics. Although the efficiency of the modelling approach is demonstrated for two independent data sets, further investigations of different species and/or conditions are required to assess the general nature of the present results, and their interpretation via the concept of physiological age of the meristems. The succession of the hidden states is roughly modelled by a hypothesis of local dependency between parent and child states (Markov property). This hypothesis captures the basic characteristics of the hidden states' succession.

Limitations of the basic HMT model and robustness of the state tree restoration Considering the potential complexity of dependencies within a tree structure, the HMT model described above can be regarded as a rough model. This is the simplest hidden Markovian model that can take the tree structure into account. The main advantage of this rough statistical model is its parsimony – for a given number of states and a given modelling of the observation processes, its parameterization is that of a simple hidden first-order Markov chain. As a counterpart of such parsimony, some dependencies cannot be modelled, particularly the dependencies between child vertices of a given parent vertex, and those between a vertex and its nonparent ancestors.

In the examples studied, the states are markedly differentiated by the attached observation distributions. For instance, in the apple tree case, the potential ambiguity between states that explains the 'hidden' nature of the model applies to the medium and long shoots only. Hence the restoration of the hidden state tree, which can be considered as robust relative to a model misspecification, enables the structures that are improperly represented in the model parameters to be highlighted (see apple tree results).

As a consequence, in both applications considered in this paper, the features that could not be captured explicitly by the HMT model have been handled by the analysis of the restored state tree (e.g. the frequencies of state transitions for each type of edge, for apple tree). In the same direction, if several scales are relevant for the analysis, an HMT model can be identified for each scale separately and the restored state trees can be computed for each scale. Then the interscale dependencies can be analysed *a posteriori*.

Alternative families of HMT models However, some of the aforementioned limitations can be addressed by specific refinements of the model. The first is related to the absence of distinction between succession and branching. This is a consequence of the property of conditional independence of the children states, given the parent state. It follows from this property that the children's conditional distribution is invariant under any permutation of the children: assume that vertex 1 has children set {2,3}. Then:

$$P(S_3 = k, S_2 = j \mid S_1 = i) = P(S_3 = k \mid S_1 = i)P(S_2 = j \mid S_1 = i)$$
$$= P(S_2 = k \mid S_1 = i)P(S_3 = j \mid S_1 = i) = P(S_3 = j, S_2 = k \mid S_1 = i).$$

Consequently, the model cannot make the distinction between succession and branching (see apple tree results; Fig. 4). As we generally expect ruptures to be caused by branching (as opposed to succession), the conditional independence assumption is rather crippling.

Dependent-children hidden Markov out-tree model

To overcome this drawback, we propose a model where the conditional independence assumption concerning the child vertices is relaxed. We obtain the dependent-children HMOT model, also oriented from the root to the leaf vertices, but with dependent children states, given the parent state. This model is parameterized by transition probabilities from the parent state to the set of children states.

In the case of ordered children, this transition probability corresponds to $P(S_3 = k, S_2 = j \mid S_1 = i) = p_{i,jk}$. The transition probability $P(S_3 = j, S_2 = k \mid S_1 = i) = p_{i,jk}$ is another parameter of the model that is *a priori* different from $p_{i,j}$. The number of parameters of a dependent-children HMOT model is quite similar to that of a high-order or variable-order Markov chain.

A way to reduce the number of parameters is to use the information on the edge types (< and +) to partially order the children. Practically, this means that the successor $n(u)$ of the parent entity is distinguished from the other offspring entities $b(u)$ deriving from the parent entity by branching. The set of nonsuccessor entities is assumed to be unordered. In this case, the transition probabilities at vertex u are $P(S_{n(u)} = j, S_{b(u)} = \{k,m\} \mid S_u = i) = p_{i,j,km}$. They are assumed to be invariant under any permutation of the nonsuccessor entities. If the whole set of children is assumed unordered, then the transition probabilities at vertex u are $P(S_{c(u)} = \{k,m\} \mid S_u = i) = p_{i,km}$.

Hidden Markov in-tree model For some biological phenomena (e.g. apical dominance, delayed branching), and for some applications (when the attributes cannot be observed on the innermost part of the plant due to cambial growth and self-pruning), it seems more relevant to orient the tree from the leaf vertices to the root. This leads to the hidden Markov in-tree (HMIT) model, which is parameterized 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. As a consequence, the children states are dependent given the parent state. Thus the transition probability matrix is similar to that of a high-order or variable-order Markov chain where, in the context of tree structures, the variable number of child vertices plays the role of the order. As for the dependent-children HMOT model, potential assumptions on the order, partial order or absence of order on the set of children determine the transition probabilities. In the case of ordered children, the transition probability from the children states of vertex 1 to the state at vertex 1 would be $P(S_1 = i \mid S_2 = j, S_3 = k) = p_{jk,i}$.

The three HMT models are illustrated in Fig. 6. These graphs represent the conditional independence properties between the random variables in the model, as described by Smyth *et al.* (1997).

As a consequence, the next step towards modelling conditional dependencies between children states, given the parent, would be the implementation of the aforementioned families of HMT models that are synthetically represented in Fig. 6. Such dependencies would take into account the distinction between succession and branching through the transition matrix. They require additional assumptions of order, partial order, or absence of order on the set of children.

Modelling the dependencies between ancestors A second limitation of the independent-children HMOT is related to the direct dependencies between a state and several of its ancestors. For example, this would account for the existence of tree-structured patterns. Such issues rely on building variable-order HMT models, inspired by the variable-order Markov chains in the case of sequence analysis (Bühlmann & Wyner, 1999). However, the simultaneous modelling of dependencies between children and ancestors does not seem

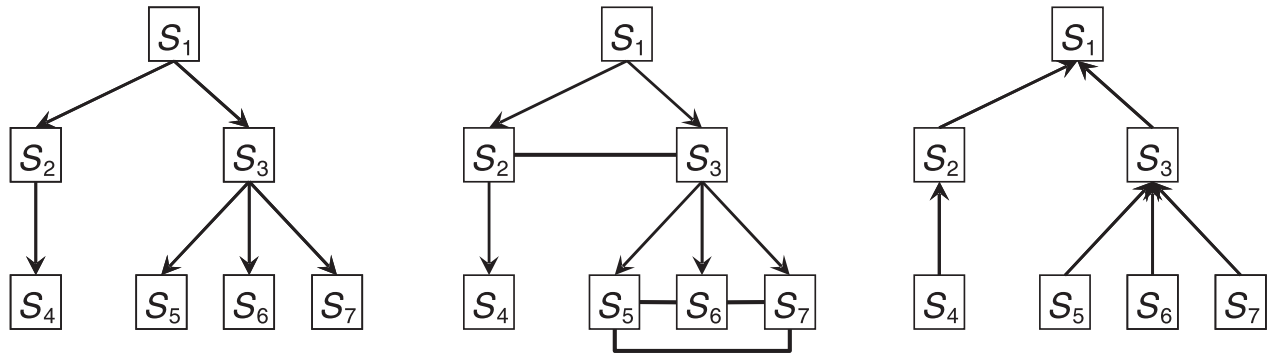

Fig. 6 Families of hidden Markov tree models and their parameters.

tractable. The application of this enhanced model to the apple tree data set could illustrate its importance, and complete the proposed analysis of the state succession (see apple tree results). A further issue is the periodicity of flowering occurrence, which remains to be analysed at a local scale as well as globally, for example using the tree segmentation obtained by state restoration.

Perspectives on methodologies and applications

Analysis of the periodicity of flowering occurrence requires computation of the distribution of empirical and theoretical characteristics of trees, such as number of GUs between flowering occurrences. This approach has been used for the analysis of sequences, and the results are quite helpful for model building and exploratory analysis, as well as for model validation (Guédon *et al.*, 2001). However, in the case of tree-structured data, the loss of the notion of unique successor makes the definition of such characteristics more difficult.

Concerning model building, the choice of the number of states is based, ideally, on the exploratory analysis and on statistical criteria, although in this paper greater importance is given to these criteria. As a consequence, a high number of hidden states cannot be statistically supported by small data sets, thus discrimination by the HMT models of many morphological changes requires large amounts of data. However, the peripheral parts of individuals are numerous and tend to have lower variability than the central parts, which have greater variability. This should lead to an adaptive sampling in the protocol of measurement. In the case of partial measurements, the deletion of structural data (part of the tree topology) can be handled by the HMIT model. Concerning the issue of missing attributes for given entities, the pruning strategy described in this paper for the bush willow data set can be avoided by resorting to a dedicated algorithm, developed by Celeux & Durand (2002) in the context of hidden Markov chains.

From the viewpoint of application, dominating or equivalent paths have been handled here by HMT modelling only. The approach is based on a heuristic cost function that favours small values of the states at basal positions. However, this *ad hoc* method can barely be extended to the determination of equivalent (or partially equivalent) branching systems. Such a family of problems can be approached by dedicated methods that rely on an edit distance between tree structures (Ferraro & Godin, 2000, 2003). These algorithms implement the comparison of branching systems on the basis of both topology and variables – in our case, the state variable. Furthermore, the existence of an order on the state values would provide a natural local cost function for this variable. As a result, the combination of the state tree restoration provided by HMT modelling, with tree comparison algorithms, is a promising way of determining reiterated complexes or identifying some hierarchical levels among branching systems.

As the HMT models provide a general framework for the identification of homogeneous zones within a plant – for its automatic segmentation into several parts of similar nature, or for the extraction of remarkable paths – this approach proves especially useful in the context of perennial plants. In this case (or any other context of usually huge and complex crowns), the amount of available entities per plant is large. Hence the trees may need to be sampled. For this purpose, the homogeneous zones are expected to provide some guidelines for the selection of branching systems for the reconstruction of whole individuals. Lastly, a number of models (not necessarily stochastic) are valid under the assumption that the local characteristics of the plant do not vary abruptly. Consequently, homogeneous data sets obtained by such a sampling method could be analysed independently using different models.

More generally, the segmentation of plants into a small number of states will offer new possibilities for quantifying

physiological age. As a perspective, this should lead to a methodology for the validation of this notion.

References

Adam B, Sinoquet H, Godin C, Dones N. 1999. *3A – A Software for the Acquisition of Plant Architecture, Version 2.0.* Clermont-Ferrand, France: UMR PIAF INRA-UBP.

Barlow PW. 1994. From cell to system: repetitive units of growth in the development of roots and shoots. In: Iqbal M, ed. *Growth Patterns in Vascular Plants*. Portland, OR, USA: Dioscorides Press, 19–58.

Barthélémy D. 1991. Levels of organization and repetition phenomena in seed plants. *Acta Biotheoretica* **39**: 309–323.

Barthélémy D, Caraglio Y, Costes E. 1997. Architecture, gradients morphogénétiques etâge physiologique chez les végétaux. In: Bouchon J, de Reffye P, Barthélémy D, eds. *Modélisation et Simulation de L'architecture des Végétaux*. INRA Editions, 89–136.

Bell A. 1991. *Plant Form – An Illustrated Guide to Flowering Plant Morphology*. Oxford, New York, Tokyo: Oxford University Press.

Biernacki C, Celeux G, Govaert G. 2000. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**: 719–725.

Bonnet P. 2002. Évaluation de la productivité du *Combretum adenogonium* Steud. ex. A. Rich à partir d'une approche architecturale. *DESS Gestion des Systèmes Agro-Sylvo-Pastoraux en Zone Tropicale*. Université Paris XII.

Bühlmann P, Wyner AJ. 1999. Variable length Markov chains. *Annals of Statistics* **27**: 480–513.

Celeux G, Durand J-B. 2002. Choosing the order of a hidden Markov chain through cross-validated likelihood. In: Klinke S, Ahrend P, Richter L, eds. *Compstat 2002.* Berlin, Germany: Humboldt-Universität zu Berlin. http://ise.wiwi.hu-berlin.de/~sigbert/compstat2002/paper/short/C_02_celeux.pdf

Costes E, Guédon Y. 2002. Modelling branching patterns on 1-year-old trunks of six apple cultivars. *Annals of Botany* **89**: 513–524.

Costes E, Sinoquet H, Kelner JJ, Godin C. 2003. Exploring within-tree architectural development of two apple tree cultivars over 6 years. *Annals of Botany* **91**: 91–104.

Crabbé J, Escobedo-Alvarez JA. 1991. Activités méristématiques et cadre temporel assurant la transformation florale des bourgeons chez le pommier (*Malus* × *domestica* Borkh., cv. Golden Delicious). *L'Arbre: Biologie et Développement. Naturalia Monspeliensia*, 369–379.

Crouse MS, Nowak RD, Baraniuk RG. 1998. Wavelet-based signal processing using hidden Markov models. *IEEE Transactions on Signal Processing* **46**: 886–902.

Durand J-B, Gonçalvès P, Guédon Y. 2004. Computational methods for hidden Markov trees – an application to wavelet trees. *IEEE Transactions on Signal Processing* **52**: 2551–2560.

Ephraim Y, Merhav N. 2002. Hidden Markov processes. *IEEE Transactions on Information Theory* **48**: 1518–1569.

Ferraro P, Godin C. 2000. A distance measure between plant architectures. *Annals of Forest Sciences* **57**: 445–461.

Ferraro P, Godin C. 2003. An edit distance between quotiented trees. *Algorithmica* **36**: 1–39.

Gatsuk LE, Smirnova OV, Vorontzova LI, Zaugolnova LB, Zhukova LA. 1980. Age states of plants of various growth forms: a review. *Journal of Ecology* **68**: 675–696.

Geiger D, Heckerman D, King H, Mekk C. 2001. Stratified exponential families: graphical models and model selection. *Annals of Statistics* **29**: 505–529.

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

Godin C, Guédon Y, Costes E, Caraglio Y. 1997. Measuring and analysing plants with the AMAPmod software. In: Michalewicz MT, ed. *Plants to Ecosystems – Advances in Computational Life Sciences*, Vol. I. Melbourne, Australia: CSIRO Publishing, 53–84.

Guédon Y, Barthélémy D, Caraglio Y, Costes E. 2001. Pattern analysis in branching and axillary flowering sequences. *Journal of Theoretical Biology* **212**: 481–520.

Lauri PÉ, Térouanne É, Lespinasse JM. 1997. Relationship between the early development of apple fruiting branches and the regularity of bearing – an approach to the strategies of various cultivars. *Journal of Horticultural Science* **72**: 519–530.

Nozeran R, Bancilhon L, Neville P. 1971. Intervention of internal correlations in the morphogenesis of higher plants. *Advances in Morphology* **9**: 1–66.

Smyth P, Heckerman D, Jordan MI. 1997. Probabilistic independence networks for hidden Markov probability models. *Neural Computation* **9**: 227–270.

White J. 1979. The plant as a metapopulation. *Annual Review of Ecology and Systematics* **10**: 109–145.