ONCOGENETIC TREES

ANIKO SZABO, KENNETH BOUCHER AND LISA PAPPAS

1. Definitions and basic results

1.1. **Description of the data.** Before defining the oncogenetic tree model, we first describe the data that it is designed to model and that will be used for model fitting. Let $M_1, M_2, \ldots, M_n$ denote the genetic alterations of interest. These could be point mutations, gain or loss of chromosomal regions or other genetic events. N independent specimens ("tumors") are obtained and the presence or absence of the alterations of interest is recorded as a binary vector $\mathbf{x}_j = (x_{j1}, x_{j2}, \ldots, x_{jn})$, where

$$x_{j\ell} = \begin{cases} 0, & \text{if } M_\ell \text{ is absent in the } j^{th} \text{ tumor} \\ 1, & \text{if } M_\ell \text{ is present in the } j^{th} \text{ tumor,} \end{cases} \qquad j = 1, \ldots, N;\ \ell = 1, \ldots, n$$

The reconstruction algorithm will only use the marginal and pairwise frequencies of occurrence of the alterations, so we introduce the following notations:

- $p_i = P(M_i \text{ occurs}),\ i = 1, \ldots, n;\ p_0 = 1$
- $p_{ij} = \begin{cases} P(\text{both } M_i \text{ and } M_j \text{ occur}), & i, j = 1, \ldots n;\ i \neq j \\ p_i, & i = 1, \ldots n;\ j = 0, i \end{cases}$
- $p_{i|j} = P(M_i \text{ occurs given } M_j \text{ has occurred}),\ i, j = 1, \ldots n;\ i \neq j$
- $p_{i \vee j} = P(M_i \text{ or } M_j \text{ or both occur}), i, j = 1, \ldots n;\ i \neq j$

We will assume that only actually observed alterations are modeled, so $p_i > 0$ always.

1.2. **The oncogenetic tree model.** In this section we give a short description of an oncogenetic tree and provide some pertinent definitions. For a more complete treatment we refer the reader to Desper et al. [1999]. An oncogenetic tree models the process of occurrence of genetic alterations in carcinogenesis using a directed tree structure. In this paper we will use the words *tree* and *branching* for a directed graph T with vertex set $\{M_0\} \cup V = \{M_0, M_1, \ldots, M_n\}$ such that for every vertex $M_i \in V$ there is a unique directed path from M_0 to M_i along the edges of T. In the literature such a structure is also called an arborescence. We will use the common "arrow" notation to denote the edges of the tree: $\overrightarrow{M_i M_j}$ denotes the directed edge from vertex M_i to vertex M_j.

Intuitively, vertex M_0 (the root of the tree) represents the 'no alterations' event and each of the vertices of V represent a certain mutation or other genetic alteration. Thus the alteration status of a tumor is described by a set of the vertices that correspond to the alterations that are present in the tumor.

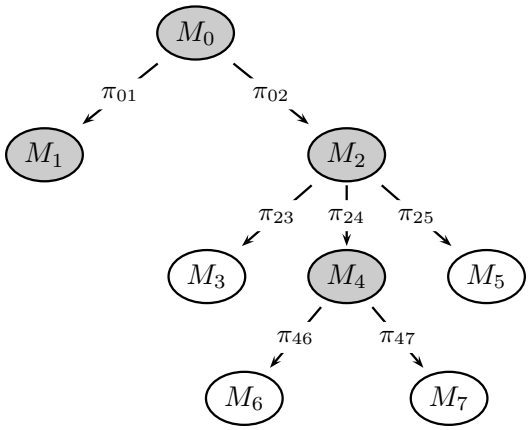

FIGURE 1. An example of an untimed oncogenetic tree with seven possible alterations.

First we give an intuitive description of the oncogenetic tree using a simple example given in Figure 1; here $M_1, M_2, \ldots, M_7$ represent hypothetical alterations of interest. The development of a tumor according to this tree could be the following: the tumor starts as $\{M_0\}$, that is none of the alterations have occurred. Now the events M_1 and M_2 can occur, and their appearance is independent of each other, that is the occurrence of one of them does not change the probability of occurrence for the other one. Suppose M_2 has occurred and so the status of the tumor becomes $\{M_0, M_2\}$. Now in addition to M_1, the alterations M_3, M_4 and M_5 can also occur, so the tumor can move to the status $\{M_0, M_1, M_2\}$, $\{M_0, M_2, M_3\}$, $\{M_0, M_2, M_4\}$ or $\{M_0, M_2, M_5\}$ and so on. The observed status of the tumor depends on the time of the observation. The values π_{ij} on the edges are the probabilities of transition along the given edge by the time of observation. These values allow to find the model-based probability of observing any combination of the alterations in a tumor; for example, the probability of the set highlighted with grey is $P(\{M_0, M_1, M_2, M_4\}) = \pi_{01}\pi_{02}\pi_{24}(1-\pi_{23})(1-\pi_{25})(1-\pi_{46})(1-\pi_{47})$ and $P(\{M_0, M_4\}) = 0$ as according to the tree M_2 had to occur before M_4 could. This intuitive description is formalized by the following definitions:

Definition 1. *A* pure untimed oncogenetic tree *is a tree T with a probability $\pi(e)$ attached to each edge e. This tree generates observations on mutation presence/absence the following way: each edge e is independently retained with probability $\pi(e)$; the set of vertices that are still reachable from M_0 gives the set of the observed genetic alterations.*

A somewhat more realistic model incorporates the progression of time.

Definition 2. *A* pure timed oncogenetic tree *is a tree T with a rate $\lambda(e)$ attached to each edge and an observation-time distribution φ on $\mathbb{R}^+$. This tree generates observations on mutation presence/absence the following way: first the time of observation τ is drawn from φ and the transition time along each edge e is drawn independently from an exponential distribution with rate $\lambda(e)$. The set of vertices that are reachable from M_0 along a path for which the sum of transition times is less than τ gives the set of the observed genetic alterations.*

While the above definition of an oncogenetic tree gives a clearly interpretable model for the process of occurrence of genetic events during carcinogenesis, *real* data never quite follows prescribed models. Thus before a tree model can be fitted, an error structure describing the character of random deviations from the model has to be defined. There are several sources of errors in the context of this model. Some of the observations $x_{j\ell}$ might be incorrect due to the imperfection of the detection technology or the spatial heterogeneity of the tumor. A more fundamental source of "errors" is the truly random occurrence of genetic alteration unrelated to the causal process of carcinogenesis. The error model introduced by Szabo and Boucher [2002] suggests combining the possible errors regardless of their source into two basic types: false positives and false negatives, and base the error model on the probabilities of occurrence of these errors.

Error model

- The tumor develops according to the pure oncogenetic tree model
- The presence/absence of each alteration is independently measured
- If the alteration is present it is not observed with probability ϵ_-.
  If the alteration is absent it is observed with probability ϵ_+.

2. Reconstruction

The main goal of the analysis is the reconstruction of the topology of the oncogenetic tree T; the estimation of the edge transition probabilities and error probabilities is of secondary importance. First we will concentrate on the conceptual aspects of reconstruction and assume that there is no sampling error (the sample size $N \to \infty$). One of the main results of the theory of oncogenetic trees is the Reconstruction Algorithm given in Figure 2 that provides an explicit construction method for T [Szabo and Boucher, 2002]. This algorithm takes a greedy bottom-up approach: it assigns the parent of each node by finding the maximum-weight in-edge starting from the leaves.

Reconstruction algorithm

1. Estimate p_i and $p_{ij}, i, j = 0, \ldots, n$ from the marginal frequencies in the data using the definitions (Section 1.1).
2. Construct a complete directed graph on vertices $\{M_0, M_1, \ldots, M_n\}$ representing the occurrence of individual events with weight $w(M_i, M_j) = \log \frac{p_{ij}}{p_j(p_i + p_j)}$ for the directed edge $\overrightarrow{M_i M_j}$.
3. Build a directed spanning tree (branching) B by defining the ancestor of each vertex the following way:
   a. Let S denote the set of vertices with assigned parent. Start with $S = \emptyset$.
   b. Find the vertex $M_i \notin S$ with the smallest probability p_i (in case of a tie, choose randomly).
   c. Let its parent in B be the vertex $M_j \notin S$ such that $w(M_j, M_i)$ is maximal. Set $S = S \cup \{M_i\}$.
   d. Repeat steps 3b-3c until all vertices have an assigned parent, that is $S = V$ (vertex M_0 does not need a parent).

FIGURE 2. Algorithm for reconstructing the oncogenetic tree from marginal and pairwise joint distribution of alterations.

REFERENCES

R. Desper, F. Jiang, O.-P. Kallioniemi, H. Moch, C.H. Papadimitriou, and A. A. Schäffer. Inferring tree models for oncogenesis from comparative genome hybridizationdata. *Journal of Computational Biology*, 6(1):37–51, 1999.

A. Szabo and K. Boucher. Estimating an oncogenetic tree when false negatives and positives are present. *Mathematical Biosciences*, 176(2):219–236, 2002.