

Chapter 1

ONCOGENETIC TREES

Aniko Szabo and Kenneth M. Boucher

Human solid tumors are believed to be caused by a sequence of genetic abnormalities arising in normal and premalignant cells. The understanding of these sequences is important for improving cancer treatment. Models for the occurrence of the abnormalities include linear structure and a recently proposed tree-based structure. We will describe the oncogenetic tree model and an efficient algorithm for its estimation. We also discuss methods for estimating the reliability and goodness-of-fit of this reconstruction. An R package “Oncotree” implementing the described methodology is available from the authors.

Keywords: Oncogenesis, branching, robustness, bootstrap.

1. INTRODUCTION

A seminal effort in describing the steps involved in carcinogenesis was a study of colorectal tumor development by Vogelstein *et al.* (1988). The authors have shown that while the genetic profile of individual tumors varied widely and there was no single mutation present in all tumors, certain changes tended to occur early in the development, and other ones relatively late. In a subsequent paper Fearon and Vogelstein (1990) proposed a linear genetic model for colorectal tumorigenesis as a preferred order of occurrence of the genetic abnormalities while acknowledging the existence of other pathways. Their conclusion was equivocal: “...although a preferred order for the genetic alterations...exist, the data suggest that the progressive accumulation of these alterations is the most consistent feature...” (Fearon and Vogelstein, 1990).

The idea of multiple pathways of cancer progression was introduced in the works of Zelen (Zelen, 1968; Feldstein and Zelen, 1984). Szabo and Yakovlev (2001) considered modeling the process of tumorigenesis as a mixture of all possible pathways; however, not so surprisingly, they showed that given the available data such a model is generally not identifiable in the case of three or more genetic events. Thus one linear pathway does not adequately describe the diversity of cancer development, yet the commonly available data does not contain sufficient information to identify a model with an arbitrarily large number of pathways.

As a compromise between the extremes of too few and too many pathways, Desper *et al.* (1999) introduced the oncogenetic tree model in which the possible pathways form a directed tree structure: they share a common beginning, but can “branch out” at the ends. The linear model of Fearon and Vogelstein (1990) is a special case of the oncogenetic tree model; however the latter is more flexible and appears to be more realistic. This paper is dedicated to describing oncogenetic trees and investigating their properties.

2. DEFINITIONS AND BASIC RESULTS

2.1. 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, \dots, 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 $\mathbf{x}_j = (x_{j1}, x_{j2}, \dots, x_{jn})$, where

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

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

notations:

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

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

2.2. 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 both the words *tree* and *branching* to refer to a rooted directed graph T with vertex set $\{M_0\} \cup V = \{M_0, M_1, M_2, \dots, M_n\}$ 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.

First we give an intuitive description of the oncogenetic tree using a simple example given in Figure 1; here M_1, M_2, \dots, 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

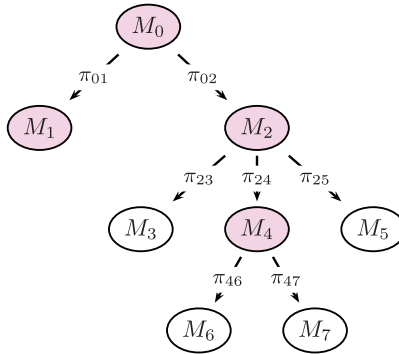


Fig. 1. An example of an untimed oncogenetic tree with seven possible alterations.

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 finding the model-based probability of observing any combination of the alterations in a tumor. For example, $P(\{M_0, M_4\}) = 0$ because, according to the tree, M_2 had to occur before M_4 could; while the probability of the set highlighted with grey is $P(\{M_0, M_1, M_2, M_4\}) = \pi_{01} \pi_{02} \pi_{24} (1 - \pi_{23})(1 - \pi_{25})(1 - \pi_{46})(1 - \pi_{47})$.

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 t 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 t 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.

2.2.1. *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_+ .

3. 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 $N \rightarrow \infty$). 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

- (i) Estimate p_i and p_{ij} , $i, j = 0, \dots, n$ from the marginal frequencies in the data using the definitions (Section 2.1).
- (ii) Construct a complete directed graph on vertices $\{M_0, M_1, \dots, M_n\}$ 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 $\overrightarrow{M_i M_j}$.
- (iii) 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 $M_i \in S$ with the smallest probability p_i (in case of a tie, choose randomly).
 - (c) Let its parent in B be the vertex $M_j \in S$ such that $w(M_j, M_i)$ is maximal. Set $S = S \cup \{M_i\}$.
 - (d) Repeat steps (b)–(c) until all vertices have an assigned parent, that is $S = V$ (vertex M_0 does not need a parent).

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

In the absence of false observations, this algorithm reconstructs the original tree T provided that it is not skewed:

Definition 3. An oncogenetic tree T is skewed if there exist two vertices M_i, M_j with a least common ancestor M_k in T such that

$$p_{i|j} \geq p_{i \vee j|k}. \quad (1)$$

Lemma 0. An untimed oncogenetic tree is not skewed.

Proof. In an untimed tree events M_i and M_j are conditionally independent given the status of their least common ancestor M_k , so $p_{i|j} = p_{i|k} < p_{i \vee j|k}$. \square

Note that in a timed tree $p_{i|j} > p_{i|k}$, so skewness can occur.

It can be easily seen that in a pure oncogenetic tree the non-skewness condition is equivalent to having

$$p_{i|j} < \frac{p_i + p_j}{p_k + p_j}. \tag{2}$$

This form will be easier to use in the proofs.

Theorem 1 (Reconstruction Theorem). *Let T be a non-skewed oncogenetic tree (timed or untimed) and ε_+ , ε_- be the probabilities of, respectively, a false positive and false negative observation, and let $p_{\min} = \min_i p_i$. If $\varepsilon_+ + \varepsilon_- < 1$ and $\varepsilon_+ < (p_{\min})^{1/2}(1 - \varepsilon_+ - \varepsilon_-)$, then the branching B given by the tree reconstruction algorithm is exactly T .*

We will prove this theorem using three lemmas. First note that after incorporating false positives and negatives, the probabilities of observing alterations will become

$$\begin{aligned} p_i^* &= p_i(1 - \varepsilon_-) + (1 - p_i)\varepsilon_+ \\ p_{ij}^* &= p_{ij}(1 - \varepsilon_-)^2 + (p_{i \vee j} - p_{ij})(1 - \varepsilon_-)\varepsilon_+ + (1 - p_{i \vee j})\varepsilon_+^2. \end{aligned} \tag{3}$$

Lemma 1.1. *If M_j is a parent of M_i in T , then $p_j^* > p_i^*$.*

Proof. Since $p_j > p_i$, the statement easily follows from the first equation in Eq. (3): $p_j^* - p_i^* = (p_j - p_i)(1 - \varepsilon_- - \varepsilon_+) > 0$ unless $\varepsilon_- + \varepsilon_+ \geq 1$. □

Lemma 1.2. *If M_j is not an ancestor of M_i in T , then $w(M_k, M_i) > w(M_j, M_i)$, where M_k is the least common ancestor of M_i and M_j .*

Proof. From the definition,

$$w(M_k, M_i) - w(M_j, M_i) = \log \frac{p_{ki}^*(p_i^* + p_j^*)}{p_{ji}^*(p_k^* + p_i^*)}.$$

As M_k is an ancestor of M_i , $p_{ki} = p_i$, so without observation errors the non-skewness assumption Eq. (2) would ensure that the above expression is positive, proving the lemma. We will show that under the assumptions of this theorem the non-skewness inequality is maintained even after the introduction of observational errors.

From Eqs. (3) we have

$$\begin{aligned}
 p_{ki}^*(p_i^* + p_j^*) - p_{ji}^*(p_k^* + p_i^*) &= [p_i(1 - \varepsilon_-)^2 + (p_k - p_i)(1 - \varepsilon_-)\varepsilon_+ \\
 &+ (1 - p_k)\varepsilon_+^2][(p_i + p_j)(1 - \varepsilon_- - \varepsilon_+) + 2\varepsilon_+] - [p_{ij}(1 - \varepsilon_-)^2 \\
 &+ (p_{i\vee j} - p_{ij})(1 - \varepsilon_-)\varepsilon_+ + (1 - p_{i\vee j})\varepsilon_+^2][(p_i + p_k)(1 - \varepsilon_- - \varepsilon_+) \\
 &+ 2\varepsilon_+] = (1 - \varepsilon_- - \varepsilon_+)[(p_i^2 - p_i p_{ij} + p_i p_j - p_{ij} p_k)(1 - \varepsilon_- - \varepsilon_+)^2 \\
 &+ 2\varepsilon_+(p_i - p_{ij})(1 - \varepsilon_- - \varepsilon_+) + (p_k - p_j)\varepsilon_+^2] > 0.
 \end{aligned}$$

The second equality can be checked by expanding both sides of the equation and the last inequality follows because $1 - \varepsilon_- - \varepsilon_+ > 0$ (assumption of the theorem), $p_i^2 - p_i p_{ij} + p_i p_j - p_{ij} p_k > 0$ (non-skewness assumption Eq. (2)), $p_i > p_{ij}$ (by definition) and $p_k > p_j$ (M_k is an ancestor of M_j).

Thus

$$\frac{p_i^* + p_j^*}{p_k^* + p_i^*} > \frac{p_{ji}^*}{p_{ki}^*},$$

so $w(M_k, M_i) > w(M_j, M_i)$, proving the statement. □

Lemma 1.3. *If M_j is the parent of M_i in T and M_k is any other ancestor of M_i in T then $w(M_k, M_i) < w(M_j, M_i)$.*

Proof.

$$w(M_j, M_i) - w(M_k, M_i) = \log \frac{p_{ji}^*(p_k^* + p_i^*)}{p_{ki}^*(p_j^* + p_i^*)}.$$

Without observational errors $p_{ji} = p_{ki} = p_i$ and this expression is positive as $p_k > p_j$. We will show that this statement holds in the presence of the errors as well by invoking the condition $\varepsilon_+ < (p_{\min})^{1/2}(1 - \varepsilon_+ - \varepsilon_-)$.

As M_k and M_j are ancestors of M_i , $p_{ji} = p_{ki} = p_i$ and $p_{k\vee i} = p_k$, so from Eq. (3) we have

$$\begin{aligned}
 p_{ji}^*(p_k^* + p_i^*) - p_{ki}^*(p_j^* + p_i^*) &= [p_i(1 - \varepsilon_-)^2 + (p_j - p_i)(1 - \varepsilon_-)\varepsilon_+ \\
 &+ (1 - p_j)\varepsilon_+^2][(p_i + p_k)(1 - \varepsilon_- - \varepsilon_+) + 2\varepsilon_+] - [p_i(1 - \varepsilon_-)^2
 \end{aligned}$$

$$\begin{aligned}
 &+ (p_k - p_i)(1 - \varepsilon_-)\varepsilon_+ + (1 - p_k)\varepsilon_+^2][(p_i + p_j)(1 - \varepsilon_- - \varepsilon_+) \\
 &+ 2\varepsilon_+] = (1 - \varepsilon_- - \varepsilon_+)(p_k - p_j)[(1 - \varepsilon_- - \varepsilon_+)^2 p_i - \varepsilon_+^2] > 0.
 \end{aligned}$$

Again, the verification of the second equality is straightforward, while the inequality follows because $1 - \varepsilon_- - \varepsilon_+ > 0$, $p_k > p_j$ (M_k is an ancestor of M_j) and $(1 - \varepsilon_- - \varepsilon_+)^2 p_i - \varepsilon_+^2 > (1 - \varepsilon_- - \varepsilon_+)^2 p_{\min} - \varepsilon_+^2 > 0$ (assumption of the theorem).

Hence $w(M_j, M_i) > w(M_k, M_i)$. □

Proof of the Reconstruction Theorem. Combining together the results of these lemmas, we have proven that the vertex M_i chosen in step (3b) of the Reconstruction Algorithm cannot be the parent of any other vertex in S (Lemma 1.1); and the vertex M_j chosen in step (3c) is its parent in T (Lemmas 1.2, 1.3). Hence B coincides with T . □

The Reconstruction Theorem can be generalized in a variety of ways. For example, Szabo and Boucher (2002) developed sufficient conditions for reconstructing the oncogenetic tree in the case when the rates of false negative errors are allowed to vary as long as they are “almost equal”. Here we will not attempt to find the most general statement possible, but rather explore other issues.

The Reconstruction Algorithm provides an intuitively appealing and computationally fast approach for estimating an oncogenetic tree. However its constructivist nature does not provide a good conceptual description of the tree. The following theorem from Desper *et al.* (1999) gives such a characterization.

Theorem 2. *The oncogenetic tree T is a maximum weight branching spanning all the vertices $\{M_0, M_1, \dots, M_n\}$.*

Proof. Assume that T is not a maximum weight branching, but instead D is. We will prove that D coincides with T in three steps, each using one of the three lemmas proved in the Reconstruction Theorem. □

Lemma 2.1. *M_0 is the root of the maximum weight branching D .*

Suppose another vertex M_i is the root of D , while the parent of M_0 in D is M_j ($j = i$ is possible). Consider the branching D' obtained by replacing

the edge $\overrightarrow{M_j M_0}$ by $\overrightarrow{M_0 M_i} : D' = D - \overrightarrow{M_j M_0} + \overrightarrow{M_0 M_i}$. D' has M_0 as a root. Then D' has a higher weight than D , since

$$w(D') - w(D) = w(M_0, M_i) - w(M_j, M_0) = -\log(1 + p_i^*) - \log(p_j^*) + \log(1 + p_j^*) > -\log(2) + \log(1 + 1/p_j^*) > 0,$$

using the fact that all probabilities are less than one.

Hence if M_0 is not the root, the branching cannot have maximal weight.

Lemma 2.2. *If $\overrightarrow{M_i M_j}$ is an edge in D , then M_i is an ancestor of M_j in T .*

Suppose the statement is false. From all the vertices M_j with parent in D not an ancestor in T , choose the one closest to M_0 in T . Let M_k be the least common ancestor of M_i and M_j in T .

Consider the branching D' obtained by replacing the edge $\overrightarrow{M_i M_j}$ by $\overrightarrow{M_k M_j} : D' = D - \overrightarrow{M_i M_j} + \overrightarrow{M_k M_j}$. Since M_k is closer to the root than M_j , the statement of the lemma holds for M_k and its ancestors. Hence M_j cannot be M_k 's ancestor and D' is really a branching. Then

$$w(D') - w(D) = w(M_k, M_i) - w(M_i, M_j) = \log \frac{p_{jk}^*(p_i^* + p_j^*)}{p_{ij}^*(p_{ij}^* + p_k^*)} > 0$$

as shown in Lemma 1.2.

Lemma 2.3. *For every edge $\overrightarrow{M_i M_j}$ of D , M_i is the parent of M_j in T as well.*

Again, suppose $M_k \neq M_i$ is the parent of M_j in T . Then from the previous step M_i is an ancestor of M_k in T . Consider the branching $D' = D - \overrightarrow{M_i M_j} + \overrightarrow{M_k M_j}$. The previous step ensures that M_j cannot be an ancestor of M_k , so D' really is a branching. Then

$$w(D') - w(D) = w(M_k, M_i) - w(M_i, M_j) = \log \frac{p_{jk}^*(p_i^* + p_j^*)}{p_{ij}^*(p_{ik}^* + p_k^*)} > 0$$

as proven in Lemma 1.3 of the Reconstruction Theorem.

Note, that unlike finding a maximum weight spanning tree, the problem of finding a maximum weight branching (a directed tree) is not simple. Polynomial-time algorithms have been developed for this problem (Edmonds, 1967; Karp, 1971; Tarjan, 1977), however none are as simple and fast as the Reconstruction Algorithm. The reason for this is that our algorithm uses the special structure of the weights that is not available in the general case.

4. SAMPLE SIZE ESTIMATION

The success of the Reconstruction Algorithm depends on the relative order of the frequencies of occurrence of the mutations and of the edge weights. In the Reconstruction Theorem we have shown that (under certain conditions) the introduction of false positive and negative errors maintains the correct ordering. However these results were proven only for the “true” probabilities p_i^* and p_{ij}^* , ignoring the variability inherent to sampling. In this section we give a lower bound on the sample size that is sufficient for reconstruction with a (large) predefined probability $1 - \xi$.

First, we introduce a few notations. Let $\alpha = \min_{i,j,k} \frac{p_i + p_j}{p_k + p_j} - p_{i|j}$, where the minimum is taken over all triples (i, j, k) such that M_k is the least common ancestor of M_i and M_j , $\beta = \min_i (p_{\text{parent}(i)} - p_i)$, $p_{\min} = \min_i p_i$ and $p_{\max} = \max_i p_i$. Intuitively, α measures the tightness of the non-skewness assumption, and β measures the ability to determine the order of “adjacent” events.

Theorem 3. *Let T be a non-skewed oncogenetic tree (timed or untimed) with n vertices (not including the root M_0) and $\varepsilon_+, \varepsilon_-$ be the probabilities of, respectively, a false positive and false negative observation. If $\varepsilon_+ + \varepsilon_- < 1$, $\chi = \varepsilon_+ / (1 - \varepsilon_+ - \varepsilon_-) < p_{\min}^{1/2}$, and the sample size*

$$N \geq \frac{81(p_{\max} + \chi)^3 (\ln[n(n + 1)] - \ln(2\xi))}{(1 - \varepsilon_+ - \varepsilon_-)^3 (\min[\alpha p_{\min}^2 + \beta \chi^2, \beta(p_{\min} - \chi^2)])^2}, \tag{4}$$

then with probability at least $1 - \xi$ the branching B given by the Reconstruction Algorithm is exactly T .

Proof. Let $\hat{\delta}_i = \hat{p}_i^* - p_i^*$ and $\hat{\delta}_{ij} = \hat{p}_{ij}^* - p_{ij}^*$, $i, j = 0, \dots, n$ denote the deviation of the observed frequencies from their theoretical counterparts, and let $\delta = \max(\max_i \hat{\delta}_i, \max_{ij} \hat{\delta}_{ij})$. With some extra work in each of the lemmas in the proof of the Reconstruction Theorem, it can be shown to remain valid if

$$\delta < \frac{1}{9p_{\max}^*} (1 - \varepsilon_+ - \varepsilon_-)^3 \min[\alpha p_{\min}^2 + \beta \chi^2, \beta(p_{\min} - \chi^2)], \quad (5)$$

where $\chi = \varepsilon_+ / (1 - \varepsilon_+ - \varepsilon_-)$, $p_{\max}^* = \max_i p_i^*$. From the first equation in Eq. (3), $p_{\max}^* = (1 - \varepsilon_+ - \varepsilon_-) p_{\max} + \varepsilon_+$.

The sample size estimation will be based on the Chernoff inequality Ross (2002): if $X \sim \text{Binomial}(N, p)$ and $\hat{p}_N = X/N$ denotes the estimated response probability, then for any $u > 0$:

$$P\left(\hat{p}_N - p > \frac{u\sqrt{p}}{\sqrt{N}}\right) \leq e^{-u^2}. \quad (6)$$

Specifically,

$$P\left(\hat{\delta}_i > \frac{u\sqrt{p_{\max}^*}}{\sqrt{N}}\right) < P\left(\hat{\delta}_i > \frac{u\sqrt{p_i^*}}{\sqrt{N}}\right) \leq e^{-u^2}$$

$$P\left(\max_i \hat{\delta}_i > \frac{u\sqrt{p_{\max}^*}}{\sqrt{N}}\right) \leq n e^{-u^2}.$$

Similarly,

$$P\left(\max_{ij} \hat{\delta}_{ij} > \frac{u\sqrt{p_{\max}^*}}{\sqrt{N}}\right) \leq \binom{n}{2} e^{-u^2},$$

hence

$$P\left(\delta > \frac{u\sqrt{p_{\max}^*}}{\sqrt{N}}\right) \leq \frac{n(n+1)}{2} e^{-u^2}.$$

We select u to ensure the desired significance level by setting $e^{-u^2} n(n+1)/2 = \xi$, that is $u^2 = \ln[n(n+1)/(2\xi)]$. On the other side of the inequality,

using the limit on the sample size N set in Eq. (4), we can see that the requirement of Eq. (5) is satisfied:

$$\frac{u\sqrt{p_{\max}^*}}{\sqrt{N}} \leq \frac{1}{9p_{\max}^*} (1 - \varepsilon_+ - \varepsilon_-)^3 \min[\alpha p_{\min}^2 + \beta \chi^2, \beta(p_{\min} - \chi^2)]. \quad \square$$

5. PARAMETER ESTIMATION

So far we have addressed only the reconstruction of the topological structure of the oncogenetic tree. Here we will discuss the estimation of the further parameters of the model for untimed oncogenetic trees, that is the estimation of the edge transition probabilities $\pi(e)$ and the error rates $\varepsilon_+, \varepsilon_-$. The estimation of these parameters is linked.

Edge transition probabilities. We propose using a method-of-moments estimator based on the relationship valid in the pure tree: $\pi(\overrightarrow{M_j M_i}) = p_{i|j}$. Equations (3) describe the effect of the error model on this relationship. Thus, in the absence of sampling variability, we can recover the edge weights by solving these equations for $p_{i|j}$:

$$p_{i|j} = \frac{p_{ij}^* - (p_i^* + p_j^*)\varepsilon_+ + \varepsilon_+^2}{(p_j^* - \varepsilon_+)(1 - \varepsilon_+ - \varepsilon_-)}. \quad (7)$$

In practice, we use the observed marginal probabilities \hat{p}_i^* and \hat{p}_j^* to estimate the edge weights.

Error rates. The tree structure, the edge weights and the error probabilities jointly define a distribution of all the possible outcome sets. Thus, we can estimate the error probabilities by fitting this estimated distribution to the observed one. However, obtaining the model-based distribution is computationally expensive (each outcome based on the error-free tree can be “corrupted” in 2^n ways), so we propose fitting the only the marginal probabilities which are much easier to compute.

The tree-based marginal probability of occurrence for event M_i is

$$p_i^*(\varepsilon_+, \varepsilon_-)_T = \varepsilon_+ + (1 - \varepsilon_-) \prod_k p_{j_{k+1}|j_k}, \quad (8)$$

where $0 = j_1, j_2, \dots, j_d = i$ is the path from the root to M_i .

Thus, combining Eqs. (7) and (8), for every value of the error rates $(\varepsilon_+, \varepsilon_-)$ we can obtain the tree-based marginal probabilities of occurrence and compare them to the observed frequencies. After defining a suitable error-function (in the application we use the squared ℓ_2 -distance $\sum_i (\hat{p}_i^* - p_i^*(\varepsilon_+, \varepsilon_-)_T)^2$), the error rates can be estimated by minimizing the error-function.

6. EXAMPLE: RENAL CARCINOMA DEVELOPMENT

In this section we show an application of the oncogenetic tree model to comparative genomic hybridization data for clear cell renal carcinoma. The comparative genomic hybridization technique (CGH) developed by Kallioniemi *et al.* (1992) was used on each of the $N = 124$ samples as described in Jiang *et al.* (1998) to obtain information on chromosome number aberrations (CNAs) on each of the arms of the chromosomes. A more detailed description of the dataset can be found in Desper *et al.* (1999); Jiang *et al.* (2000). The human genome consists of 22 autosomal and 2 sex-linked chromosomes. All chromosomes have a long arm q and most (except 13, 14, 15, 21 and 22) have a significant short arm p . The chromosome arms are denoted by attaching the letter p or q to the appropriate chromosome number. The CGH technique uses fluorescent staining to detect abnormal (increased or decreased) number of DNA copies. In contrast to microarray technology, the results cannot be narrowed down to a specific gene, only to a segment of the chromosome, called a band. However when two tumors have abnormalities along a similar region, it is often difficult to tell whether they are based on the same genetic change, so in the renal carcinoma data set the results are reported as a gain or loss on a certain arm, without further distinction for specific bands. Also, as some samples were from females, the Y chromosome was excluded from consideration. This resulted in 82 possible events from 41 locations (both a gain and a loss could occur on different bands of the same chromosomal arm). It is common to denote a change in DNA copy number on a specific chromosome arm by prefixing a “-” sign for decrease and a “+” for increase. Thus, say, $-3q$ denotes abnormally low DNA copy number on the q arm of the 3rd chromosome.

To reduce the total number of events and to keep p_{\min} reasonably large, we selected the seven most frequent events (listed in decreasing order of

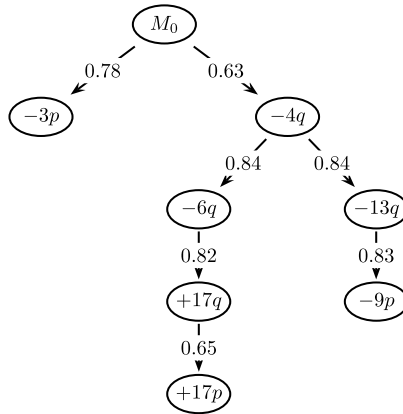


Fig. 3. Oncogenetic tree based on the renal carcinoma CGH data. The edges are labeled with the estimated transition probability $\hat{\pi}_e$.

frequency): $-3p, -4q, -6q, -13q, -9p, +17q, +17p$. Other approaches to the selection of relevant events have also been proposed, including the method due to Brodeur *et al.* (1982) that allows to adjust for prior probabilities of alteration on the chromosome arms. The later approach was used in Jiang *et al.* (2000) and resulted in a list of 12 events; the selected events are included in that list.

Upon applying the Reconstruction Algorithm we obtain the tree shown in Figure 3. This tree will be used as a basis of simulation studies in Section 7, so we introduce the notation \mathcal{T}_{CGH} for it. The error rate was estimated by fitting the tree-based marginal probabilities of occurrence $p_i(\varepsilon_+, \varepsilon_-)$ to the observed frequencies \hat{p}_i through minimizing the squared ℓ_2 -distance $\sum_i (\hat{p}_i p_i(\varepsilon_+, \varepsilon_-))^2$ as described in Section 5: $\hat{\varepsilon}_+ = 0.00, \hat{\varepsilon}_- = 0.228$.

Since the combination of the pure oncogenetic tree model with the error process defines a non-zero probability for any possible outcome (given the parameters), maximum likelihood estimation of the error rates is possible.

7. PROPERTIES OF THE ONCOGENETIC TREE ESTIMATOR: A SIMULATION STUDY

In this section we investigate the properties of the Reconstruction Algorithm as an estimator of the true oncogenetic tree structure. The Reconstruction Theorem guarantees that given a sufficiently large sample and with error

rates satisfying its restrictions, the correct tree will be reconstructed with a given probability. However, the sufficient sample size provided by Eq. (4) is extremely conservative. For example, for \mathcal{T}_{CGH} we have $\alpha = 0.032$, $\beta = 0.031$, $\varepsilon_+ = 0.00$, $\varepsilon_- = 0.228$, $p_{\min} = 0.26$, and $p_{\max} = 0.78$, so for a 95% confidence ($\xi = 0.05$) the formula requires $N \approx 210,000,000$ samples. In practice, it is likely that much smaller sample sizes are sufficient. Also, the dependence of N on the parameters is quite convoluted, their relative significance is hidden. All the investigations are empirical and are based on the oncogenetic tree \mathcal{T}_{CGH} estimated in Section 6. While many of the specific results depend on the exact structure of the tree, the values of the transition probabilities, etc. the qualitative conclusions are likely to be valid generally.

7.1. Simulating Data Based on a Given Tree

The generation of data from an oncogenetic tree is fairly straightforward based on the definitions of a pure oncogenetic tree and of the error process. First a “clean” observation is generated by retaining each edge with the associated probability and creating the set of occurred alterations from the vertices still reachable from M_0 . Then the errors are introduced independently for each alteration: each occurred alteration is not observed with probability ε_- , and each alteration that has not occurred is observed with probability ε_+ . This process is repeated until the required sample size is reached.

7.2. Probability of Correct Reconstruction

First, we investigate the dependence of the probability of correct reconstruction of the data generating tree on the error probabilities when the sample size is fixed ($N = 124$). We considered 12 combinations for $(\varepsilon_+, \varepsilon_-)$ with the parameters taking the values 0.00 through 0.10 and 0.20, respectively, and generated 1000 random data sets for each combination. The Reconstruction Algorithm was applied to each random data set and the probability of incorrect reconstruction was estimated as the proportion of sets for which the reconstructed tree had the same structure as \mathcal{T}_{CGH} . To ease interpretation in Figure 4, a logistic surface with quadratic dependence on ε_+ and ε_- was fitted to the resulting estimates.

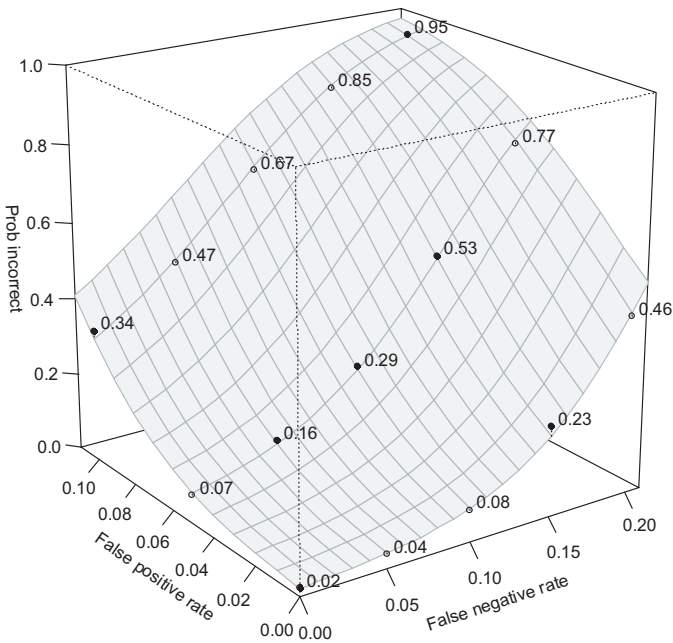


Fig. 4. Effect of error rate on the probability of incorrect reconstruction. Based on the renal carcinoma tree \mathcal{T}_{CGH} in Fig. 3, 1000 simulated samples of size 124 at each point.

From the results it is evident that the effect of the error probabilities ε_+ and ε_- on the success of reconstruction is quite different. False positive errors appear to have a significantly higher deteriorating effect than false negative errors. Unfortunately, the presence of observations that are false positive with respect to the tree model is an intrinsic feature of the problem and cannot be eliminated by technological improvement. Fortunately, in our data the false negative rate is the one that is estimated to be high. Using simulations, the probability of correct reconstruction (assuming the model is correct) was estimated to be 39%.

7.3. Sample Size for High Probability of Reconstruction

For a researcher planning to collect mutation occurrence data the important question is rather the sample size required to achieve correct reconstruction with a sufficiently high probability. For the same 15 combinations of

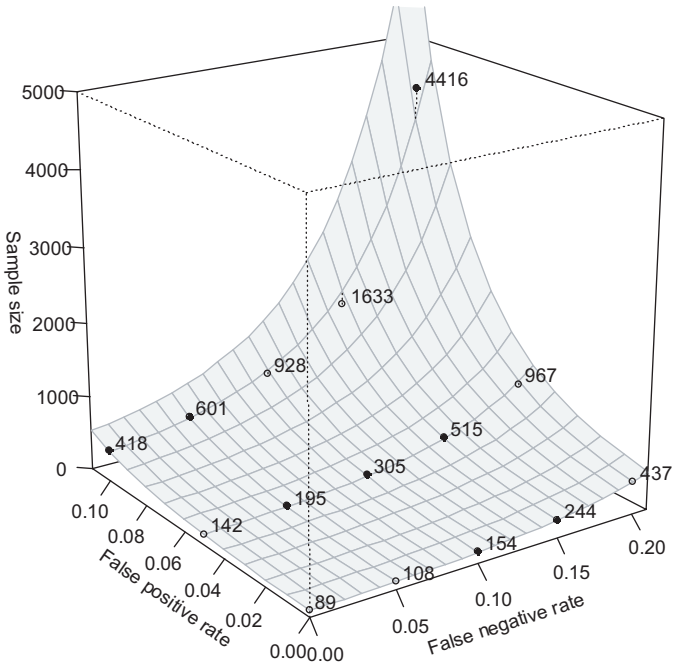


Fig. 5. Effect of error rates on the sample size $N_{0.95}(\varepsilon_+, \varepsilon_-)$ required for 95% confidence reconstruction. Based on the renal carcinoma tree \mathcal{T}_{CGH} in Fig. 3, 1000 simulated samples of size 124 at each point.

$(\varepsilon_+, \varepsilon_-) \in \{0; 0.05; 0.1\} \times \{0; 0.05; 0.1; 0.15; 0.20\}$ as above, we estimated the sample size $N_{0.95}(\varepsilon_+, \varepsilon_-)$ at which the tree reconstructed from a data set randomly generated from \mathcal{T}_{CGH} coincided with it 95% of the time. The frequency estimates were based on 1000 simulated data sets. Figure 5 shows the results for the 15 design points with a fitted surface. As previously, false positive errors have a much stronger impact on the required sample size. It is notable that unless the error rates are large, the required samples size is realistic in practice.

In Section 7.2 we found that with error rates $\varepsilon_+ = 0.00$ and $\varepsilon_- = 0.228$, the probability of correct reconstruction with sample size 124 is 39%. We used 1000 simulated data sets to estimate the sample size required for 95% confidence reconstruction of the oncogenetic tree for the renal carcinoma data: $N_{0.95}(\hat{\varepsilon}_+, \hat{\varepsilon}_-) = 658$.

8. GOODNESS OF FIT

In the previous section we used simulation to estimate the confidence of correct reconstruction *if the oncogenetic tree model is correct*. Thus the 39% confidence level is model-based. In this section we use bootstrap techniques to obtain a non-parametric estimate of the reconstruction confidence, and thus examine the goodness of fit of the oncogenetic tree model.

8.1. Bootstrap Estimate of Reconstruction Confidence

The reconstruction confidence is closely related to the variability of the estimator of the structure of the oncogenetic tree. A well-established non-parametric method to evaluate the variability of an estimator is bootstrap resampling. We generated 1000 bootstrap data sets, that is the original data set was sampled with replacement to obtain new data sets of the same size, and an oncogenetic tree \tilde{T}_i , $i = 1, \dots, 1000$ was estimated for each of these data sets. We found a very high variability among the \tilde{T}_i 's: 149 different trees were found with the most frequent (\mathcal{T}_{CGH}) occurring only 8% of the time. Figure 6 shows the nine most frequently occurring trees.

Our non-parametric estimate of the reconstruction confidence is the proportion of the bootstrap estimates \tilde{T}_i that have the same structure as the tree estimated from the original data \mathcal{T}_{CGH} , that is 8%. This is substantially lower than the parametric estimate of 39% obtained in Section 7.2. Such a disagreement raises doubt in the goodness of fit of the oncogenetic tree model.

8.2. Analysis of the Stable Portions

Despite the high variability of the trees based on bootstrap resamples of the data, there are conserved portions that are present in a large proportion of the trees. These pieces and the corresponding probabilities of occurrence are highlighted in Figure 7.

–3p is a direct descendant of the root. According to the oncogenetic tree model, the most frequently occurring event cannot be a consequence of another event, so it has to be a direct descendant of the root. In the renal carcinoma data, the –3p event is by far the most frequent ($\approx 60\%$ while

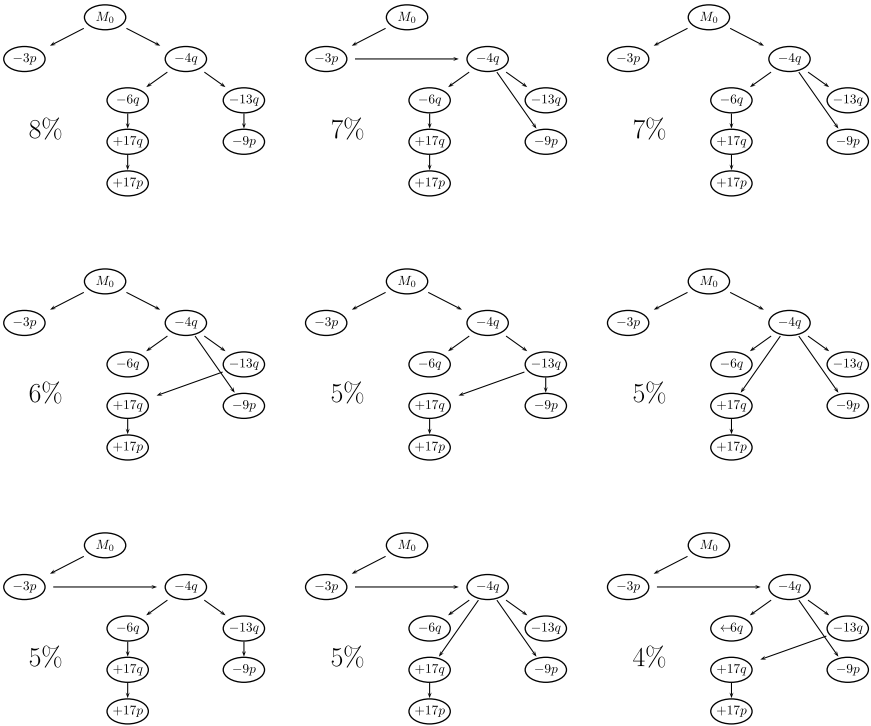


Fig. 6. The most frequent (out of 149 different) trees with associated frequencies from 1000 bootstrap resamples of size 124 of the renal carcinoma data.

for the next most frequent event $P(-4q) \approx 50\%$). Thus the stability of this edge is expected and does not provide any additional insight.

+17p is a child of +17q. This edge is more interesting, especially because the two events occur on the same chromosome and thus an association through the gain/loss of the entire chromosome is not unexpected. The occurrence of +17p jumps from almost none (4%) with normal 17q to 50% when there is a gain on 17q.

-6q and -13q are children of -4q. This cluster is the most complex and unexpected *a priori*. According to the definition of an oncogenetic tree, it implies not only increased probability of occurrence of events -6q and -13q after a loss on 4q, but also a conditional independence of the “lower” events (given -4q). For reference, in the left panel of Figure 8 we present a mosaic plot for the joint distribution of these three events expected under

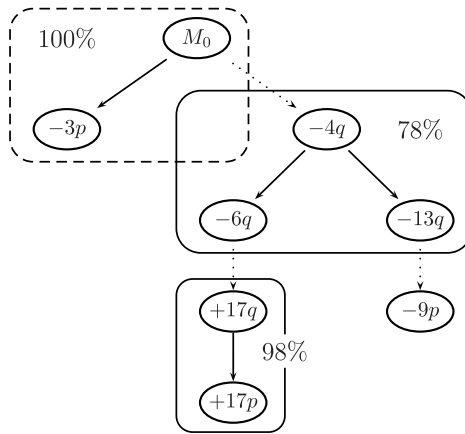


Fig. 7. Highly preserved portions of \mathcal{T}_{CGH} among the bootstrapped trees with associated frequencies of occurrence. The solid frames mark the “non-trivial” portions; the dashed frame contains an expected edge as it includes the most frequent event (see text for details).

the oncogenetic tree model with the estimated error rates $\hat{\varepsilon}_+ = 0.00$ and $\hat{\varepsilon}_- = 0.228$. In this plot the area of each cell is proportional to the frequency of occurrence of the corresponding combination of the events. When $4q$ is normal (left side), the cell with normal $6q$ and $13q$ dominates; the other cells are due to false positives. However when $-4q$ is present (right side), the probabilities of $-6q$ and $-13q$ become larger, so all the cells increase. Note that due to the presence of false negative errors (but not false positives), the observed data is not expected to follow conditional independence for normal $4q$, only the underlying “true” events would show the grid pattern that is present on the $-4q$ side. The observed mosaic plot is in the right panel of Figure 8. The cells of this plot are coded according to their deviation from the Poisson regression model adjusting for expected frequencies: the fit is very good, none of the cells have residuals with absolute value above 1.3. Comparing this plot to the expected plot on the left side we see a general agreement in the patterns and a clear effect of $-4q$ on the probabilities of occurrence of $-6q$ and $-13q$. While there is some evidence for an excess of positive correlation (cells that are concordant with respect to $-6q$ and $-13q$ are larger), this excess is not statistically significant.

In conclusion of the above discussion, it appears that different analysis of the available data supports the presence of the stable associations found

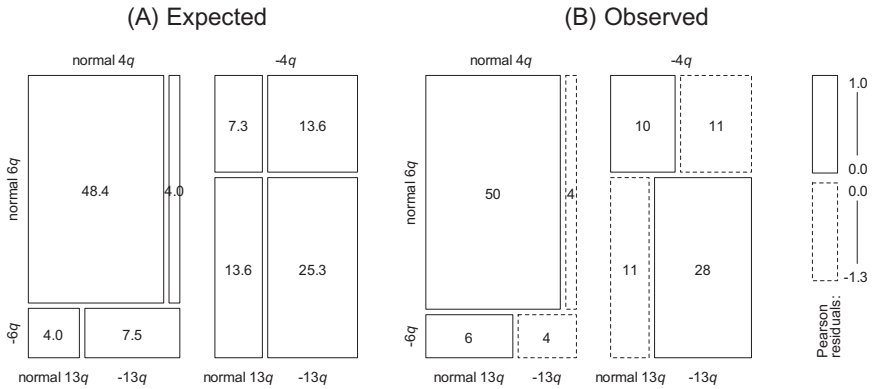


Fig. 8. The expected (A) and observed (B) mosaic plots of the joint distribution of $-4q$, $-6q$ and $-13q$. The area of the cells is proportional to the probability of occurrence; the labels are the expected/observed frequencies, respectively. In (B) the cells are coded according to the Pearson residual for a fit to the expected values in a Poisson regression model (see text).

by the oncogenetic tree model. As a next step, such hypotheses should be validated on an independent data set.

9. DISCUSSION

We have seen that oncogenetic trees provide a flexible, yet not too rich space for modeling genetic alteration data. More development is needed for the estimation of the timing of the alterations; the estimation of the parameters of the timed oncogenetic tree model is an open problem. While time information is not directly available for human tumors, stage and/or tumor size could possibly be used as surrogates.

An alternative tree-based modeling approach has been recently proposed for genetic alteration data (Desper *et al.*, 2000; von Heydebreck *et al.*, 2004). This methodology, that unfortunately is also often referred to as oncogenetic tree building, constructs a phylogenetic tree where the alterations are the leaves. These phylogenetic trees do not have the same mechanistic interpretation as the oncogenetic trees that we described. Additionally, estimation is much more difficult, probabilistic search algorithms have to be used.

References

1. Brodeur GM, Tsiatsis AA, Williams DL, Luthardt FW and Green AA. Statistical analysis of cytogenic abnormalities in human cancer cells. *Cancer Genet. Cytogenet.* 1982; **7**: 137–152.
2. Desper R, Jiang F, Kallioniemi O, Moch H, Papadimitriou C and Schäffer A. Distance-based reconstruction of tree models for oncogenesis. *J. Comput. Biol.* 2000; **7**: 789–803.
3. Desper R, Jiang F, Kallioniemi OP, Moch H, Papadimitriou CH and Schäffer AA. Inferring tree models for oncogenesis from comparative genome hybridization data. *J. Comput. Biol.* 1999; **6**: 37–51.
4. Edmonds J. Optimum branchings. *J. Res. Natl. Bur. Stand.* 1967; **71B**: 233–240.
5. Fearon ER and Vogelstein B. A genetic model for colorectal tumorigenesis. *Cell* 1990; **61**: 759–767.
6. Feldstein M and Zelen M. Inferring the natural time history of breast cancer: implications for tumor growth rate and early detection. *Breast Cancer Res. Treat.* 1984; **4**: 3–10.
7. Jiang F, Desper R, Papadimitriou C, Schäffer A, Kallioniemi O, Richter J, Schraml P, Sauter G, Mihatsch M and Moch H. Construction of evolutionary tree models for renal cell carcinoma from comparative genomic hybridization data. *Cancer Res.* 2000; **60**: 6503–6509.
8. Jiang F, Richter J, Schraml P, Bubendorf L, Gasser T, Mihatsch MJ, Sauter G and Moch H. Chromosomal imbalances in papillary renal cell carcinoma: genetic differences between histological subtypes. *Am. J. Pathol.* 1998; **153**: 1467–1473.
9. Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F and Pinkel D. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* 1992; **258**: 818–821.
10. Karp RM. A simple derivation of Edmonds' algorithm on optimum branching. *Networks* 1971; **1**: 265–272.
11. Ross SM. *Probability Models for Computer Science*. Harcourt/ Academic Press, Burlington, MA, 2002.
12. Szabo A and Boucher K. Estimating an oncogenetic tree when false negatives and positives are present. *Math. Biosci.* 2002; **176**: 219–236.
13. Szabo A and Yakovlev A. Preferred sequences of genetic events in carcinogenesis: quantitative aspects of the problem. *J. Biol. Syst.* 2001; **9**: 105–121.
14. Tarjan RE. Finding optimum branchings. *Networks* 1977; **7**: 25–35.

15. Vogelstein B, Fearon ER, Hamilton SR, Kern SE, Preisinger AC, Leppert M, Nakamura Y, White R, Smits AMM and Bos JL. Genetic alterations during colorectal tumor development. *N. Engl. J. Med.* 1988; **319**: 525–532.
16. von Heydebreck A, Gunawan B and Füzesi L. Maximum likelihood estimation of oncogenetic tree models. *Biostatistics* 2004; **5**: 545–556.
17. Zelen M. A hypothesis for the natural time history of breast cancer. *Cancer Res.* 1968; **28**: 207–216.