

Estimating an oncogenetic tree when false negatives and positives are present

Aniko Szabo^{*}, Kenneth Boucher

Huntsman Cancer Institute and Department of Oncological Sciences, University of Utah, 2000 Circle of Hope, Salt Lake City, UT 84112-5550

Received 13 March 2001; revised 27 November 2001; accepted 15 December 2001

Abstract

Human solid tumors are believed to be caused by a sequence of genetic abnormalities arising in the tumor cells. The understanding of these sequences is extremely important for improving cancer treatment. Models for the occurrence of the abnormalities include linear structure and a recently proposed tree-based structure. In this paper we extend the pure oncogenetic tree model by introducing false positive and false negative observations. We state conditions sufficient for the reconstruction of the generating tree. As an example we analyze a comparative genomic hybridization (CGH) dataset and show that addition of the error model significantly improves the ability of the model to describe the data.

Key words: cancer genetics, carcinogenesis, error model, preferred sequence, tree

1 Introduction

Cancer is believed to be a genetic disease in which gene-level abnormalities change the normal behavior of cells, resulting in the cancerous phenotype of rapid growth and invasiveness [1]. Human solid tumors generally contain many genetic alterations and their occurrence is thought to be responsible for the progression through the histopathological stages from normal cell to dysplasia to local tumor to metastasis. There are two main mechanisms through which genetic changes influence tumor progression: the deactivation of tumor suppressor genes (e.g. the mismatch repair mechanism) that increases the probability of further genetic changes, and the activation of oncogenes that give

^{*} Corresponding author. Ph: (801)585-5182, fax: (801)585-5357
Email address: aniko.szabo@hci.utah.edu (Aniko Szabo).

the cells the cancerous properties. An understanding of the role of each of the genetic abnormalities and the identification of a preferred order of occurrence would greatly improve the staging of tumors and allow to choose the most appropriate treatment for each patient.

The first effort in describing the steps involved in carcinogenesis was a study of colorectal tumor development by Vogelstein et al. [2]. The authors have shown that while the genetic profile of individual tumors varies widely and there is no single mutation present in all tumors, certain changes tend to occur early in the development, and other ones relatively late. In a subsequent paper [3] the authors propose a linear genetic model for colorectal tumorigenesis as a preferred order of occurrence of the genetic abnormalities while acknowledging the existence of other pathways. Possible biases and estimation problems for such linear pathways are discussed by Szabo and Yakovlev [4].

As a way to combine several pathways in one model, Desper et al. [5] introduced the concept of an oncogenetic tree in which certain genetic abnormalities can lead to several others (as opposed to one in the linear model) by increasing their chance of occurrence. This model allows for multiple possible pathways and even for parallel progression along several pathways in the same tumor. Oncogenetic trees include the linear model proposed [3] as a special case, however they are more flexible and appear to be more realistic. Our paper is devoted to further investigation of the oncogenetic tree model. While we consider an application to comparative genomic hybridization data, the concept of the oncogenetic tree is not limited to this area, genetic alterations detected by a variety of techniques can be modeled.

Desper et al. [5] considered the problem of inferring the structure of an oncogenetic tree from comparative genomic hybridization (CGH) data. They proposed an algorithm that recovers the correct tree structure when the probability distribution generated by the tree is sampled. Though the authors motivated the need for a specialized algorithm by the presence of noise in the data, in the derivation of their results they allow only for sampling error (variability between samples taken under similar conditions), but no false negatives or positives. They also fail to evaluate how well the proposed model fits the actual data.

In practice there are several possible sources of error. One source of error is the imperfection of the technique (any technique) used to detect the genetic abnormalities of interest. While there has great progress in the area of biotechnology, the possibility of incorrect conclusions is very real. Such an error could produce both a false negative and false positive observation, though the methodology is generally aimed at ensuring that the detected changes are real, thus false negative mistakes are much more likely. Some alterations present in the tumor might be missed because of the spatial heterogeneity of

tumors, resulting in a false negative. Also some genetic events occur truly randomly, outside the model implied by the oncogenetic tree. Such occurrences are false positives from the point of view of the tree model. Many sources of error will be greatly reduced or eliminated with the progression of technology, however some of them are intrinsic to the problem and always will be present. In this paper we propose an extension of the pure oncogenetic tree that leads to a model that allows for the occurrence of false positive and false negative observations. We give an algorithm (that is equivalent to the more complex algorithm given in [5]) that reconstructs the oncogenetic tree that generated the data even when such errors are allowed. As an illustration we apply the methodology to a clear cell renal cell carcinoma CGH dataset, estimate the error rates and evaluate the fit. Our results show that our model explains many features of the data not predicted by the pure tree model.

2 Methods

2.1 Definitions and notations

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 [5]. 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 word *tree* for a directed graph T with vertex set $\{v_0\} \cup V = \{v_1, \dots, v_n\}$ that does not contain any cycles and such that for every vertex $v_i \in V$ there is a unique directed path from v_0 to v_i along the edges of T . In the literature such a structure is often called a branching or arborescence. Intuitively, vertex v_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.

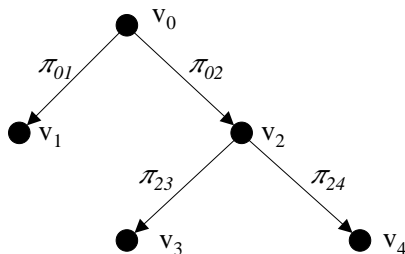


Fig. 1. An example of an untimed oncogenetic tree with four possible alterations.

First we give an intuitive description of the oncogenetic tree using a simple example given in Figure 1; here v_1, v_2, v_3 and v_4 represent four hypothetical alterations of interest. The development of a tumor according this tree could be the following: the tumor starts as $\{v_0\}$, that is none of the alterations have occurred. Now the events v_1 and v_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 v_2 has occurred and so the status of the tumor becomes $\{v_0, v_2\}$. Now in addition to v_1 , the alterations v_3 and v_4 can also occur, so the tumor can move to the status $\{v_0, v_1, v_2\}$, $\{v_0, v_2, v_3\}$ or $\{v_0, v_2, v_4\}$ 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 one to find the model-based probability of observing any combination of the alterations in a tumor; for example, $P(\{v_0, v_1, v_2, v_3\}) = \pi_{01}\pi_{02}\pi_{23}(1 - \pi_{24})$ and $P(\{v_0, v_4\}) = 0$ as according to the tree v_2 had to occur before v_4 could. This intuitive description is formalized by the following definition:

Definition 1 ([5]) An *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 v_0 gives the set of observed genetic alterations.

Another, more realistic model incorporates time as well.

Definition 2 ([5]) A *timed oncogenetic tree* is a tree T with a parameter $\lambda(e)$ attached to each edge e together with a distribution ϕ on positive real numbers. This tree defines the following sampling process: for each edge $t(e)$, the time of progression along it, is generated as an exponential random variable with parameter $\lambda(e)$, then the time of observation t_{tot} is drawn from the time-of-observation distribution ϕ ; a mutation is observed only if the corresponding vertex can be reached from v_0 in time less than t_{tot} , that is the sum of all $t(e)$'s along the path is at most t_{tot} .

We will use the term *oncogenetic tree* when a statement is true for both the untimed and timed versions.

To model observational errors suppose that though the genetic alterations in the tumor occur according to the oncogenetic tree model, each present/absent measurement has a some chance of being observed incorrectly. Specifically, suppose an absent alteration is observed with probability ϵ_+ (a false positive) and a present alteration is not observed with probability ϵ_- (a false negative). We assume that all these errors are independent of each other (within and between tumors) and that the sample size is large enough, so that the

probabilities of the considered events are estimated with sufficient precision.

We will need the following notations for the description and analysis of the reconstruction algorithm.

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

Using these notations, the probability $\pi(e)$ for an edge e from v_j to v_i in Definition 1 corresponds to the conditional probability $p_{i|j}$.

2.2 Reconstruction of the generating tree

First we describe the algorithm for reconstructing the oncogenetic tree.

Algorithm 1 (Tree reconstruction) To reconstruct the generating oncogenetic tree:

- (1) Estimate p_i and $p_{ij}, i, j = 0, \dots, n$ from the data using the above definitions.
- (2) Construct a complete directed graph on vertices $\{v_0, v_1, \dots, v_n\}$ representing the occurrence of individual events (v_0 is the artificial vertex for no events) with edge weights $w(v_i, v_j) = \log \frac{p_{ij}}{p_j(p_i + p_j)}$.
- (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 $v_i \notin S$ corresponding to any of the vertices with the smallest probability p_i .
 - (c) Let its parent in B be the vertex $v_j \notin S$ such that $w(v_j, v_i)$ is maximal. Set $S = S \cup \{v_i\}$.
 - (d) Repeat steps 3b-3c until all vertices have an assigned parent, that is $S = V$ (vertex v_0 does not need a parent).

The choice of the edge weights $w(v_i, v_j)$ follows [5]; intuitively they place large weights on the edges $v_i \rightarrow v_j$ for which $p_{i|j} = p_{ij}/p_j$ is large compared to the individual probabilities p_i and p_j . Desper et al. [5] define B as the maximum weight branching and use a general algorithm ([6–9]) to find it. We will show that under the same conditions the above algorithm also reconstructs the original tree T , so it actually finds the same maximum branching (as long as ties are broken consistently). While both algorithms have the same $O(n^2)$

complexity, our version is much more transparent and intuitive. It is interesting that this algorithm is invariant under any monotone increasing transformation of the edge weights, while the general algorithm is not. So the use of logarithm in the definition of w is optional, we have done it only to be consistent with the previous paper.

In the absence of false observations, the condition needed for reconstruction is that T is *not skewed*, that is for any two vertices v_i, v_j and their least common ancestor v_k (denoted by $lca(v_i, v_j)$) we should have $p_{i|j} < p_{i \vee j|k}$ (see [5]). This condition is always satisfied for the untimed oncogenetic trees, however timed trees can be skewed. It can be easily seen that this condition is equivalent to having

$$p_{i|j} < \frac{p_i + p_j}{p_k + p_j}. \quad (1)$$

Let $\alpha = \min_{i,j,k} \left(\frac{p_i + p_j}{p_k + p_j} - p_{i|j} \right)$, where the minimum is taken over all triples (i, j, k) such that $v_k = lca(v_i, v_j)$. Also let $p_{min} = \min_i p_i$.

2.2.1 Uniform error probabilities

Recall, that we denoted $\epsilon_+ = Pr(\text{alteration } i \text{ is observed} \mid \text{alteration } i \text{ has not occurred})$ and $\epsilon_- = Pr(\text{alteration } i \text{ is not observed} \mid \text{alteration } i \text{ has occurred})$. Note that for now we assume that the error probabilities are the same for all alterations. We also assume that a false positive or negative observation for one alteration occurs independently from observation errors for other alterations in the same tumor. The following theorem gives the restrictions on the systematic errors sufficient for the reconstruction of the oncogenetic tree:

Reconstruction Theorem 1 (Uniform errors) *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. If $\epsilon_+ + \epsilon_- < 1$ and $\epsilon_+ < (p_{min})^{1/2}(1 - \epsilon_+ - \epsilon_-)$, then the branching B given by the tree reconstruction algorithm is exactly T .*

While we will not directly prove that Algorithm 1 produces a maximum branching, we will show that $B = T$. Combined with Theorem 3.1 of Desper et al. it follows that B happens to be a maximal branching.

First note that after incorporating false positives and negatives, the observed probabilities will become

$$p_i^* = p_i(1 - \epsilon_-) + (1 - p_i)\epsilon_+ \quad (2a)$$

$$p_{ij}^* = p_{ij}(1 - \epsilon_-)^2 + (p_{i \vee j} - p_{ij})(1 - \epsilon_-)\epsilon_+ + (1 - p_{i \vee j})\epsilon_+^2 \quad (2b)$$

Lemma 1 *If v_j is a parent of v_i in T , then $p_j^* > p_i^*$.*

PROOF. Since $p_j > p_i$, the statement easily follows from eq. (2) unless $\epsilon_- + \epsilon_+ = 1$ \square

Lemma 2 *If v_j is not an ancestor of v_i in T , then $w(v_k, v_i) > w(v_j, v_i)$, where $v_k = lca_T(v_i, v_j)$.*

PROOF. From the definition $w(v_k, v_i) - w(v_j, v_i) = \log \frac{p_{ki}^*(p_i^* + p_j^*)}{p_{ji}^*(p_k^* + p_i^*)}$. Without the observation errors $p_{ki} = p_i$, so the non-skewness assumption (1) would ensure that the above expression is positive, proving the lemma. In the Appendix we show that under the assumptions of this theorem the non-skewness inequality is maintained even after the introduction of observational errors. \square

Lemma 3 *If v_j is the parent of v_i in T and v_k is any other ancestor of v_i in T then $w(v_k, v_i) < w(v_j, v_i)$.*

PROOF. $w(v_j, v_i) - w(v_k, v_i) = \log \frac{p_{ji}^*(p_k^* + p_i^*)}{p_{ki}^*(p_j^* + p_i^*)}$. Without observational errors $p_{ji} = p_{ki} = p_i$ and this expression is positive as $p_k > p_j$. In the Appendix we show that this statement holds in the presence of the errors as well by invoking the condition $\epsilon_+ < \sqrt{p_{min}}(1 - \epsilon_+ - \epsilon_-)$. \square

PROOF. (of Reconstruction Theorem 1) Combining together the results of these lemmas, we have proven that the vertex v_i chosen in step 3b of the reconstruction algorithm 1 cannot be the parent of any other vertex in S (Lemma 1); and the vertex v_j chosen in step 3c is its parent in T (Lemmas 2, 3). Hence B coincides with T . \square

A surprising feature of Reconstruction Theorem 1 is that for typical values of p_{min} the conditions imposed on the observational errors are not very restrictive. Figure 2 shows the region of the allowed (ϵ_+, ϵ_-) pairs when $p_{min} = 0.07$ as estimated in the clear cell renal carcinoma dataset considered later. This plot demonstrates that the restriction on the false negative error rate ϵ_- are especially mild: its value can be arbitrarily large provided that ϵ_+ is small enough. Intuitively, false negative errors decrease the frequency with which a certain genetic event is observed, however apparently they do not change the relationship of the events. Of course, there must be some penalty for such errors - by decreasing the probability of observing a genetic event, the sample size necessary for sufficiently precise estimates of the probabilities increases. The effect of the error probabilities on the sample size required for a high-confidence reconstruction of the oncogenetic tree is considered in more detail in Section 2.3.

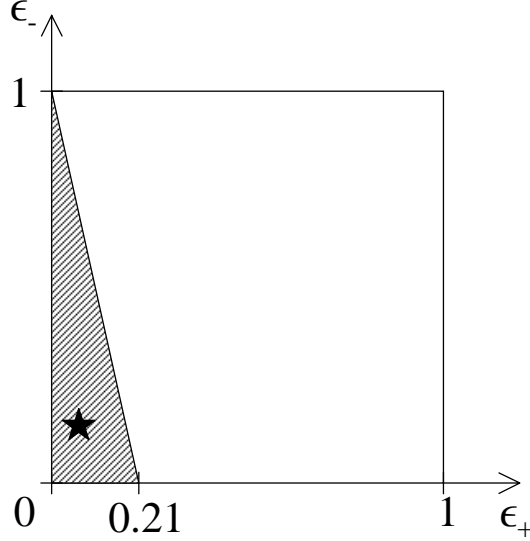


Fig. 2. The shaded area shows the allowable range for (ϵ_+, ϵ_-) when $p_{min} = 0.07$. The star denotes the value estimated for the clear cell renal carcinoma CGH data (Section 3).

2.2.2 Properties of the observed probabilities: consequences of (2)

The relationship between the underlying probabilities of occurrence and their observed counterparts defined in equation (2) has several noteworthy consequences. Equality (2a) hold for all $i = 1, \dots, n$, thus specifically for the smallest and the largest values as well. Hence

$$\begin{aligned}
 p_{min}^* &= (1 - \epsilon_+ - \epsilon_-)p_{min} + \epsilon_+ \\
 p_{max}^* &= (1 - \epsilon_+ - \epsilon_-)p_{max} + \epsilon_+ \\
 p_{max}^* - p_{min}^* &= (1 - \epsilon_+ - \epsilon_-)(p_{max} - p_{min})
 \end{aligned} \tag{3}$$

Thus the minimal and maximal observed frequencies change (up or down) according to the linear relationship, however their range is never larger than the range of the underlying probabilities of occurrence.

As mentioned above, the effect of the observational errors on the probability of observing an event is linear, but neither a straightforward increase nor a decrease – it depends on the underlying probability of occurrence. Indeed, using (2) it can be easily shown that $p_i^* = p_i$ if and only if $p_i = \epsilon_+ / (\epsilon_+ + \epsilon_-)$. Let π_0 denote this *equilibrium point*. Then the following relationship holds:

$$\begin{aligned}
 (p_i < \pi_0 &\Leftrightarrow p_i^* < \pi_0) &\Rightarrow p_i^* > p_i \\
 (p_i = \pi_0 &\Leftrightarrow p_i^* = \pi_0) &\Rightarrow p_i^* = p_i \\
 (p_i > \pi_0 &\Leftrightarrow p_i^* > \pi_0) &\Rightarrow p_i^* < p_i
 \end{aligned} \tag{4}$$

The effect of the observational errors on the joint probabilities of occurrence defined by (2b) is more complicated and cannot be simply summarized. A

relationship worth noting is the following:

$$p_{ij}^* - p_i^* p_j^* = (1 - \epsilon_+ - \epsilon_-)^2 (p_{ij} - p_i p_j). \quad (5)$$

Consequently, if the alterations i and j occur independently (thus $p_{ij} = p_i p_j$), then the events of observing these observations are also independent, as $p_{ij}^* = p_i^* p_j^*$. This is, of course, expected, as the observational errors are assumed to occur independently from each other and the genetic alterations.

2.2.3 Non-uniform error probabilities

When the error probabilities are allowed to vary, additional assumptions are necessary to reconstruct the tree. The previous theorem suggests that the reconstruction is sensitive to false positive errors, so for our next result we will consider only false negative observations. We assume that each alteration that has occurred is not observed with some probability ϵ_i that can be different for different alterations; we still, however assume that these errors are independent from each other within each tumor.

Recall, that we denoted $\alpha = \min_{i,j,k} \left(\frac{p_i + p_j}{p_k + p_j} - p_{i|j} \right)$, where the minimum is taken over all triples (i, j, k) such that $v_k = lca(v_i, v_j)$; it reflects how strictly the non-skewness inequality (1) is satisfied. Let $\beta = \min_{i,j} (p_j - p_i)$, where the minimum is taken over all pairs (i, j) such that v_j is the parent of v_i in T . Intuitively, for Lemmas 1 and 3 to hold, we must ensure that the observed frequency of a parent is higher than that of its child – thus the variability of ϵ_i should have a restriction connected to β . In the proof of Lemma 2 the non-skewness inequality is the key, restrictions to ensure its continued validity are expected to contain α .

Reconstruction Theorem 2 (No false positives) *Let T be a non-skewed oncogenetic tree with no false positive errors ($\epsilon_+ = 0$) and let ϵ_i be the probability of a false negative observation for the i^{th} alteration. If the error probabilities are “almost equal”:*

$$\max_i \frac{|\epsilon_i - \epsilon_-|}{1 - \epsilon_-} = \gamma < \min \left\{ \frac{\alpha}{\alpha + 2}, \frac{\beta}{\beta + 2} \right\} \quad (6)$$

for all i and some (fixed) ϵ_- , then the branching B given by the reconstruction algorithm 1 is exactly T .

PROOF. We need to prove the same three lemmas as in Reconstruction Theorem 1. Under the conditions of this theorem the modified probabilities

are

$$\begin{aligned} p_i^* &= p_i(1 - \epsilon_i) \\ p_{ij}^* &= p_{ij}(1 - \epsilon_i)(1 - \epsilon_j) \end{aligned} \quad (7)$$

From the definition of γ we have $\epsilon_- - \gamma(1 - \epsilon_-) \leq \epsilon_i \leq \epsilon_- + \gamma(1 - \epsilon_-)$ for any i .

1. If v_j is the parent of v_i in T , then $p_j^* - p_i^* = p_j(1 - \epsilon_j) - p_i(1 - \epsilon_i) \geq p_j(1 - \epsilon_- - \gamma(1 - \epsilon_-)) - p_i(1 - \epsilon_- + \gamma(1 - \epsilon_-)) = (1 - \epsilon_-)(p_j - p_i) - \gamma(1 - \epsilon_-)(p_j + p_i) > (1 - \epsilon_-)(\beta - 2\gamma)$. Since $\gamma \leq \beta/(\beta + 2) < \beta/2$, we get $p_j^* > p_i^*$, so the statement of Lemma 1 holds.

2. and 3. The analogs of Lemmas 2 and 3 are proven in the Appendix using the restrictions $\frac{|\epsilon_i - \epsilon_-|}{1 - \epsilon_-} < \frac{\alpha}{\alpha + 2}$ and $< \frac{\beta}{\beta + 2}$ respectively. \square

2.3 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 Theorems 1 and 2 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 estimate the sample size for which the statements of Lemmas 1 through 3 (and hence the reconstruction theorems) are true with a (large) predefined probability $1 - \xi$.

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 $\delta = \max_{i,j} (\hat{\delta}_i, \hat{\delta}_{ij})$. In the Appendix we show that Reconstruction Theorem 1 remains valid if

$$\delta < \frac{1}{9p_{max}^*} (1 - \epsilon_+ - \epsilon_-)^3 \min \left[\alpha p_{min}^2 + \beta \chi^2, \beta (p_{min} - \chi^2) \right], \quad (8)$$

where $\chi = \epsilon_+ / (1 - \epsilon_+ - \epsilon_-)$ and $p_{max}^* = \max_i p_i^* = (1 - \epsilon_+ - \epsilon_-) p_{max} + \epsilon_+$.

The sample size estimation will be based on the *Chernoff inequality* [10, Ch.3]: if $X \sim \text{Binomial}(N, p)$ and $\hat{p}_N = X/N$ denotes the estimated response probability, then for any $u > 0$

$$Pr \left(\hat{p}_N - p > \frac{u\sqrt{p}}{\sqrt{N}} \right) \leq e^{-u^2} \quad (9)$$

Specifically,

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

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

Similarly,

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

hence

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

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 we select the sample size N in accordance with the requirement (8), that is

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

Thus we have proven the quantitative version of Reconstruction Theorem 1:

Theorem 3 *Let T be a non-skewed oncogenetic tree (timed or untimed) with n vertices (not including the root v_0) and ϵ_+, ϵ_- be the probabilities of, respectively, a false positive and false negative observation. If $\epsilon_+ + \epsilon_- < 1$, $\chi = \epsilon_+/(1 - \epsilon_+ - \epsilon_-) < \sqrt{p_{min}}$ and the sample size*

$$N \geq \frac{81(p_{max} + \chi)^3 (\ln n(n+1) - \ln 2\xi)}{(1 - \epsilon_+ - \epsilon_-)^3 \min[\alpha p_{min}^2 + \beta\chi^2, \beta(p_{min} - \chi^2)]^2}, \quad (10)$$

then with probability at least $1 - \xi$ the branching B given by the tree reconstruction algorithm is exactly T .

Using similar arguments, a quantitative version of Reconstruction Theorem 2 can be proven:

Theorem 4 *Let T be a non-skewed oncogenetic tree with no false positive errors ($\epsilon_+ = 0$) and let ϵ_i be the probability of a false negative observation for the i^{th} alteration. If the error probabilities are “almost equal”:*

$$\max_i \frac{|\epsilon_i - \epsilon_-|}{1 - \epsilon_-} = \gamma < \min\left\{\frac{\alpha}{\alpha + 2}, \frac{\beta}{\beta + 2}\right\} \quad (11)$$

for all i and some (fixed) ϵ_- , and the sample size

$$N \geq \frac{81p_{max}^3(\ln n(n+1) - \ln 2\xi)}{p_{min}^2(1 - \epsilon_-)^3(1 - \gamma)^4 \min [2p_{min}(\alpha - \gamma(\alpha + 2)), \beta - \gamma(\beta + 2)]^2}, \quad (12)$$

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

The limits given in (10) and (11) are, unfortunately, too high, as during the derivation the worst possible case was used everywhere. Even using optimistic values $\epsilon_+ = \epsilon_- = 0.02$, $p_{min} = 0.15$, $p_{max} = 0.7$, $\alpha = \beta = 0.2$, the lower limit given by (10) is $N \geq 7,823,460$ to have a 95% confidence in the correctness of the tree ($\xi = 0.05$). Thus the values cannot really give guidance for sample size. While with reasonably available sample sizes the (high probability) correctness of all of the edges cannot be guaranteed, one would hope that most of the edges are still correct.

Despite the problems, these formulae do give an indication how the various parameters of the model affect the required sample size. It is encouraging to see that, like for the no-error situation of Desper et al. [5], the dependence on the number of genetic events n is logarithmic. As expected, the introduction of observational errors increases the required sample size, it is inversely proportional to the cube of the probability of no error $1 - \epsilon_+ - \epsilon_-$. It is well known that for binomial trials it is easiest to estimate probabilities close to 0.5 (in terms of sample size required to achieve a given precision); interestingly, this phenomenon surfaces here as well: the lower limit is decreased when p_{min} is increased and/or p_{max} is decreased, that is when the probabilities of occurrence of all the events are not too large and not too small.

3 Results

3.1 Application to renal cell carcinoma CGH data

In this section we investigate the properties of an oncogenetic tree fitted to a dataset of 124 cases of clear cell renal cell carcinoma. The comparative genomic hybridization technique (CGH) developed by Kallioniemi et al. [11] was used on each of the samples as described in [12] to obtain information on chromosome number aberrations (CNAs) on each of the arms of the chromosomes. More detailed description of the dataset can be found in [5,13]. The human genome consists of chromosomes 1 to 22 and the sex-linked X/Y chromosomes. All chromosomes have a long arm q and most (except 13, 14, 15,

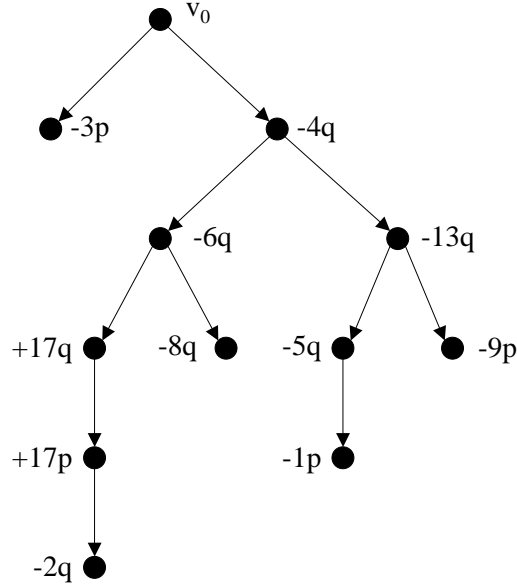


Fig. 3. The oncogenetic tree inferred from the renal carcinoma CGH data.

21 and 22) have a significant short arm p . The CGH technique uses fluorescent staining to detect abnormal (increased or decreased) number of DNA copies. In contrast to the cDNA 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 cell carcinoma dataset 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). To reduce the total number of events while keeping the most relevant ones, we used a simple heuristic similar to the one used in [5]: we identified the largest set of events such that every pair has occurred together at least 5 times. This resulted in 11 events (listed in decreasing order of frequency): $-3p$, $-4q$, $-6q$, $-13q$, $-9p$, $+17q$, $+17p$, $-8q$, $-5q$, $-2q$, $-1p$. Other approaches to the selection of relevant events have also been proposed, including the method due to Brodeur et al. [14] that allows to adjust for a priori probabilities of alteration on the chromosome arms. The later approach was used in [13] and resulted in a list of 12 events; the two lists have 7 events in common. An application of the reconstruction algorithm 1 results in the tree shown in Figure 3.

While such a tree may have an appealing interpretation, it is very important to assess how well it “fits” the generating dataset. A tree model not only predicts the expected frequencies of occurrence of mutations (marginally and pairwise), it also places severe restrictions on the possible combinations of events that can

be observed: for example, according to the model $-6q$ cannot occur without $-4q$, however in 10 tumors in the dataset this situation is observed. Altogether 53 observations out of 124 are inconsistent with the pure tree model, so it cannot be described as a well-fitting model. If we investigate the inconsistent observations more closely it turns out that 40 of them can be explained by changing the outcome of one event, suggesting that the error model introduced in this paper might significantly improve the fit of the model. There are many aspects of the data that an appropriate model should be able to explain: the joint distribution of the occurrence of the CNAs, the frequency distribution of the number of mutations, the number of inconsistent observations and the proportion of them that can be explained by error in one event, etc. We have decided to estimate the probabilities of false negative and false positive observations (ϵ_- and ϵ_+) by fitting only the marginal probabilities of occurrence for each CNA and then see how the other quantities of interest compare to the real data. The tree model allows for the calculation of the expected probability of occurrence of each event, so we estimated the error probabilities by minimizing (numerically) the χ^2 -type quantity $Q = \sum_i (\hat{p}_i(\epsilon_+, \epsilon_-) - p_i)^2 / p_i$, where p_i is the observed frequency of occurrence of alteration v_i and \hat{p}_i is its model based estimate. This method results in the estimates $\hat{\epsilon}_+ = 0.07$ and $\hat{\epsilon}_- = 0.14$ with $Q = 0.028$ (for comparison, under the assumption of no errors $Q = 0.374$). Other criteria for fit (as minimizing $\max_i |\hat{p}_i(\epsilon_+, \epsilon_-) - p_i|$) yield similar results. The improved prediction of the number of tumors with a given CNA is demonstrated in Table 1.

Using the estimates of the error probabilities additional quantities of interest like the probability of occurrence of the individual events can be estimated. The *equilibrium point* π_0 discussed in section 2.2 for which the observed and underlying probabilities of occurrence coincide is $\hat{\pi}_0 = \hat{\epsilon}_+ / (\hat{\epsilon}_+ + \hat{\epsilon}_-) = 0.33$. Events that occur with lower frequency will be observed too often and events with higher frequency will appear to occur less often than they actually do. In particular, $p_{min}^* = 16/124 = 0.13$ and $p_{max}^* = 75/124 = 0.60$ result in estimated parameters $p_{min} = 0.07 < p_{min}^*$ and $p_{max} = 0.67 > p_{max}^*$.

To evaluate how well various aspects of the data are fitted by the model, we generated data from the tree model (1000 sets of 124 samples each) using the

Table 1

Comparison of observed CNA frequencies with estimates based on the oncogenetic tree with and without the false positive/negative error model.

CNA	+17p	+17q	-1p	-2q	-3p	-4q	-5q	-6q	-8q	-9p	-13q
Observed	25	44	16	16	75	60	17	49	18	44	47
Without errors	12.3	24.7	5.7	4.9	75.0	60.0	10.8	39.0	10.3	24.9	39.0
With errors	25.7	40.4	17.2	15.4	83.9	64.9	15.5	51.2	14.3	41.2	51.4

Table 2

Comparison of data features implied by the model with the observed data.

	Observed data	Oncotree	95% CI	Oncotree	95% CI
		without errors		with errors	
Number inconsistent	53	0	(0, 0)	58.11	(50.0, 66.0)
% correctable with 1 change	75.5	not applicable		77.8	(68.5, 86.4)
Number with no CNA's	22	25.1	(18.0, 39.0)	11.4	(6.0, 17.0)

estimated error probabilities and using no errors. Table 2 shows that the number of observations that contain no CNA's is predicted better by the model without observational errors. However it fails to predict the large number of observations that are inconsistent with the pure tree model, that is observations in which a “later” alteration is observed without its “precursor”. In contrast, the inclusion of the observational errors results in a good prediction of both the number of inconsistent observations and the percentage of those that are correctable with one change.

Figure 4 compares the distribution of the number of events per tumor observed in the data with the (individual) 95% confidence intervals for the same values as expected from the oncogenetic tree with (panel B) or without (panel A) the error model. Neither model fits the observations perfectly, but the fit is visibly better with the error model. The misfit occurs at small values for the number of events. The biggest discrepancy is at the root node: the model predicts an insufficient number of samples with no CNAs and too many with one CNA. Such effect could be caused by incorrectly assuming independence among the branches in the model when the events $-3p$ and $-4q$ are actually positively correlated. To investigate this possibility, consider the following (conditional)

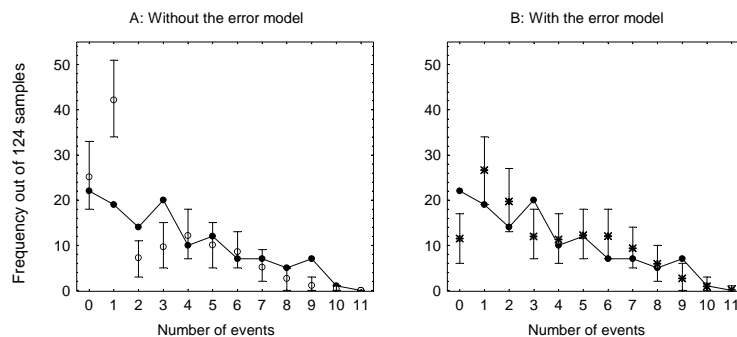


Fig. 4. Comparison of data features implied by the model with the observed data. The solid lines represent the observed values; the intervals are the 95% confidence intervals from the simulated data (see text), with a mark at the estimated mean. In panel A the simulation was performed assuming no false observations; in panel B the estimated error probabilities of $\hat{\epsilon}_+ = 0.07$ and $\hat{\epsilon}_- = 0.14$ were used.

correlation measure:

$$\text{Corr}(v_i, v_j | v_k) = \frac{p_{ij|k} - p_{i|k}p_{j|k}}{\sqrt{p_{i|k}(1 - p_{i|k})p_{j|k}(1 - p_{j|k})}}.$$

If v_k is the parent of v_i and v_j , then due to the postulated independence of the branches of the oncogenetic tree, $\text{Corr}(v_i, v_j | v_k) \approx 0$. However, from the data we estimate the correlation at the root (v_0) to be $\text{Corr}(-3p, -4q) = 0.25$; the branching at $-4q$ is also correlated ($\text{Corr}(-6q, -13q | -4q) = 0.19$) and it causes the underestimation of the frequency of samples with three mutations by the model. At the remaining branchings there is no evidence of correlation ($\text{Corr}(-8q, +17q | -6q) = -0.02$, $\text{Corr}(-5q, -9p | 13q) = -0.03$) and the model describes adequately the frequency of samples with a large number of events.

Under the false negative/positive error model each event of the observations has the same chance of being observed incorrectly. However because of structure of the tree not all such errors will result in inconsistent observations: false negative errors at the leaves ($-1p, -2q, -3p, -8q$ and $-9p$) and false positive ones at the children of the root ($-3p, -4q$) will always go undetected. Errors at other nodes can be detected with a probability that depends on the structure of the tree (mainly on the number of children and the distance from the root). When an error is detected, often it is impossible to tell whether a false positive or false negative errors has caused it: for example, if events ($-4q, -6q, +17p$) are observed in a sample, then both a false positive at $+17p$

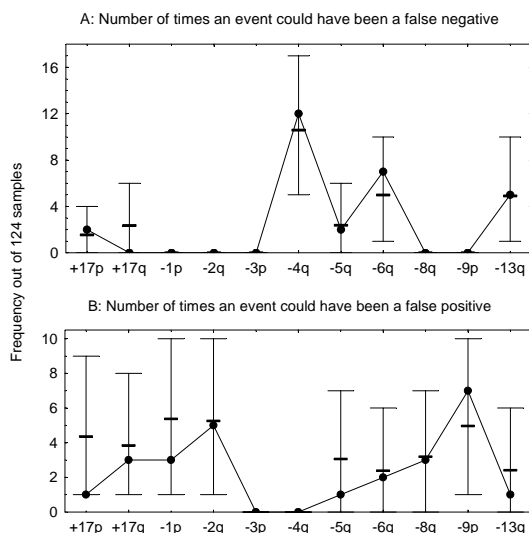


Fig. 5. Comparison of data features implied by the model with the observed data. The solid lines represent the observed values; the intervals are the 95% confidence intervals from the simulated data (see text), with a mark at the estimated mean. For each CNA the graph shows the number of times an inconsistent observation could have been made consistent with the model by assuming it was a false negative (panel A) or false positive (panel B) observation.

or a false negative at $-17q$ are plausible explanations. Figures 5A and 5B show the observed number of times that an event could “explain” an inconsistent observation by being a false negative or false positive respectively, and the 95% confidence intervals implied by the model. The observed values are in good agreement with the ones implied by the model. The largest discrepancies occur at $+17q$, where the model predicts more false negatives, and at $+17p$, where it predicts more false positives. Because of the structure of the tree, an “insufficient” (according to the model) number of samples with $-6q$ and $+17p$ but without $+17q$ explains both deviations. This effect can be caused by the assumption of independence of errors at various events; $+17p$ and $+17q$ occur on the same chromosome, so the occurrences of error might be correlated. Note that a pure oncogenetic tree (without the error model) cannot predict these values.

4 Discussion

In this paper we have extended the results of Desper et al. [5] by introducing observational errors. We gave a simpler and more intuitive algorithm for reconstructing an oncogenetic tree based on genetic alteration data. We believe that this transparency will allow to generalize the reconstruction to more general structures than trees.

The algorithm proposed in [5] was robust under sampling errors, so we expected that observational errors occurring with small probability would not interfere with the reconstruction. We have proven this to be true, but were very surprised to find such a small effect of false negative errors. If the probabilities of these errors do not differ much between the various alterations, the correct oncogenetic tree can be reconstructed regardless of the frequency of the errors (under the conditions of Reconstruction Theorem 2). The reconstruction process is much more sensitive to false positive observations. This is a serious concern since any genetic event that occurs outside the tree structure is a false positive from the point of view of the model. It is expected that no model can perfectly describe the process of mutation occurrence, so there always will be false positives.

We have also considered the issue of sample size and using theoretical considerations obtained a value for the sample size that guarantees the correctness of the reconstruction with a given probability. Unfortunately, this estimate gives values that are too high for practical applications. Further, perhaps empirical investigations of the stability of the estimated oncogenetic trees are needed to better understand the effect of lower sample sizes.

As an example we have constructed an oncogenetic tree for clear cell renal

carcinoma based on comparative genomic hybridization data and estimated the corresponding error probabilities. The proposed model resulted in a significantly improved fit over a no-error model; many features of the dataset not included in the fitting procedure were successfully predicted. The estimated error probabilities $\epsilon_+ = 0.07$ and $\epsilon_- = 0.14$ are quite reasonable considering the mechanisms that generate these errors. It is encouraging to see that the false positive error rate is small, as the reconstruction procedure is sensitive to it. The false positive error rate also gives an indication of the frequency of spontaneous mutations that occur outside the tree structure.

We have shown that oncogenetic trees augmented with an error model can be successfully and well fitted to genetic abnormality data. However our example also showed ways for further extensions: eliminating the assumption of independence at the branchings and using more sophisticated error models that allow for dependence of error rates at various locations. Another important issue is the identification of the part of the tree that is relevant to cancer development. A useful tool would be a criterion for assessing the fit of the model to the real data, so that trees of various sizes could be compared and the “best” chosen (as the Akaike criterion [15] is used for parametric models fitted by maximum likelihood). The first step in this direction has been taken by Simon et al. [16] who derived the likelihood of an oncogenetic tree under some simplifying assumptions.

Acknowledgements

We would like to thank Dr. Holger Moch for providing the renal cell clear cell carcinoma CGH dataset and Dr. Arthur Brothman for useful insights into the CGH technology. We are also thankful to the reviewers for their valuable comments and for directing us to several related papers. This research was supported, in part, by NCI Cancer Center Support Grant 2P30 CA 42014.

A Detailed proofs

Proof of Lemma 2 As v_k is an ancestor of v_i , $p_{ki} = p_i$ and $p_{k \vee i} = p_k$, so from eq. (2) we have

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

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

Thus $\frac{p_i^* + p_j^*}{p_k^* + p_i^*} > \frac{p_{ji}^*}{p_{ki}^*}$, so $w(v_k, v_i) > w(v_j, v_i)$, proving the statement. \square

Proof of Lemma 3 As v_k and v_j are ancestors of v_i , $p_{ki} = p_{ji} = p_i$ and $p_{k \vee i} = p_k$, $p_{j \vee i} = p_j$, so from eq. (2) we have

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

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

Hence $w(v_j, v_i) > w(v_k, v_i)$. \square

Proof of Reconstruction Theorem 2 (part 2) Let $v_k = lca(v_i, v_j)$ and denote $\Delta_{ijk} = p_i^2 - p_i p_{ij} + p_i p_j - p_k p_{ij}$. Then we have $|\epsilon_i - \epsilon_-| < \gamma(1 - \epsilon_-)$ for any i and $\Delta_{ijk} \geq \alpha p_i(p_i + p_k) \geq 2\alpha p_i^2$ (from the definition of α). Hence

$$\begin{aligned} p_{ki}^*(p_j^* + p_i^*) - p_{ji}^*(p_k^* + p_i^*) &= \\ \Delta_{ijk}(1 - \epsilon_i)(1 - \epsilon_j)(1 - \epsilon_k) + (1 - \epsilon_i)(1 - \epsilon_k)(\epsilon_j - \epsilon_i)p_i^2 + (1 - \epsilon_i)(1 - \epsilon_j)(\epsilon_i - \epsilon_k)p_i p_{ij} &> \\ 2\alpha p_i^2(1 - \epsilon_i)(1 - \epsilon_j)(1 - \epsilon_k) - 2\gamma(1 - \epsilon_-)p_i^2(1 - \epsilon_i)[(1 - \epsilon_j) + (1 - \epsilon_k)] &= \\ 2p_i^2(1 - \epsilon_i)(1 - \epsilon_j)(1 - \epsilon_k) \left[\alpha - \gamma \left(\frac{1 - \epsilon_-}{1 - \epsilon_k} + \frac{1 - \epsilon_-}{1 - \epsilon_j} \right) \right] &> \\ 2p_i^2(1 - \epsilon_i)(1 - \epsilon_j)(1 - \epsilon_k) \left[\alpha - \frac{2\gamma(1 - \epsilon_-)}{1 - \epsilon_- - \gamma(1 - \epsilon_-)} \right] &\geq 0, \end{aligned}$$

where the last inequality follows from $\gamma \leq \alpha/(\alpha + 2)$. Thus $w(v_k, v_i) - w(v_j, v_i) > 0$ and the statement of Lemma 2 holds. \square

Proof of Reconstruction Theorem 2 (part 3)

$$\begin{aligned}
& p_{ji}^*(p_k^* + p_i^*) - p_{ki}^*(p_j^* + p_i^*) = \\
& p_i(1 - \epsilon_i)(1 - \epsilon_j)[p_i(1 - \epsilon_i) + p_k(1 - \epsilon_k)] - p_i(1 - \epsilon_i)[p_i(1 - \epsilon_i) + p_j(1 - \epsilon_j)] = \\
& p_i(1 - \epsilon_i)[p_i(1 - \epsilon_i)(\epsilon_k - \epsilon_j) + (p_k - p_j)(1 - \epsilon_j)(1 - \epsilon_k)] = \\
& p_i(1 - \epsilon_i)(1 - \epsilon_j)(1 - \epsilon_k)\left[p_i(1 - \epsilon_i)\left(\frac{1}{1 - \epsilon_k} - \frac{1}{1 - \epsilon_j}\right) + (p_k - p_j)\right] > \\
& p_i(1 - \epsilon_i)(1 - \epsilon_j)(1 - \epsilon_k)\left[(1 - \epsilon_- + \gamma(1 - \epsilon_-))\left(\frac{1}{1 - \epsilon_- + \gamma(1 - \epsilon_-)} - \frac{1}{1 - \epsilon_- - \gamma(1 - \epsilon_-)}\right) + \beta\right] = \\
& p_i(1 - \epsilon_i)(1 - \epsilon_j)(1 - \epsilon_k)\left[\beta - \frac{2\gamma}{1 - \gamma}\right] \geq 0,
\end{aligned}$$

where the last inequality follows from $\gamma \leq \beta/(\beta + 2)$. \square

Sample size estimation for Reconstruction Theorem 1 We find the restrictions on $\delta = \max i, j(\hat{\delta}_i, \hat{\delta}_{ij})$ separately for each lemma:

Lemma 1: $\hat{p}_j^* - \hat{p}_i^* = p_j^* - p_i^* + \delta_j - \delta_i \geq (p_j - p_i)(1 - \epsilon_+ - \epsilon_-) - 2\delta \geq \beta(1 - \epsilon_+ - \epsilon_-) - 2\delta > 0$ whenever $\delta < \beta(1 - \epsilon_+ - \epsilon_-)/2$.

Lemma 2: $\hat{p}_{ki}^*(\hat{p}_i^* + \hat{p}_j^*) - \hat{p}_{ji}^*(\hat{p}_k^* + \hat{p}_i^*) = (p_{ki}^* + \delta_{ki})(p_i^* + p_j^* + \delta_i + \delta_j) - (p_{ij}^* + \delta_{ij})(p_k^* + p_i^* + \delta_k + \delta_i) \geq p_{ki}^*(p_i^* + p_j^*) - p_{ij}^*(p_k^* + p_i^*) - 8\delta p_{max}^* - 4\delta^2 > \min_{i,j,k} p_{ki}^*(p_i^* + p_j^*) - p_{ij}^*(p_k^* + p_i^*) - 9\delta p_{max}^*$, where the minimum is taken over all triples (i, j, k) such that $v_k = lca(v_i, v_j)$ and if $\delta < p_{max}^*/4$. From the proof of Lemma 2 we have $\min_{i,j,k} p_{ki}^*(p_i^* + p_j^*) - p_{ij}^*(p_k^* + p_i^*) \geq \alpha p_{min}^2(1 - \epsilon_+ - \epsilon_-)^3 + \beta \epsilon_+^2(1 - \epsilon_+ - \epsilon_-)$.

So if $\delta < \frac{1}{9p_{max}^*}(1 - \epsilon_+ - \epsilon_-)[\alpha p_{min}^2(1 - \epsilon_+ - \epsilon_-)^2 + \beta \epsilon_+^2]$ ($\delta < p_{max}^*/4$ follows from this), then $p_{ki}^*(p_i^* + p_j^*) - p_{ij}^*(p_k^* + p_i^*) > 0$ and the proof works.

Lemma 3: Similarly, just using the proof of Lemma 3, we have $\hat{p}_{ji}^*(\hat{p}_k^* + \hat{p}_i^*) - \hat{p}_{ki}^*(\hat{p}_j^* + \hat{p}_i^*) \geq (1 - \epsilon_+ - \epsilon_-)\beta[(1 - \epsilon_+ - \epsilon_-)^2 p_{min} - \epsilon_-^2] - 9\delta p_{max}^* > 0$, if $\delta < \frac{1}{9p_{max}^*}(1 - \epsilon_+ - \epsilon_-)\beta[(1 - \epsilon_+ - \epsilon_-)^2 p_{min} - \epsilon_+^2]$. Note, that the requirement for Lemma 1 follows from this condition.

Summary: Using the notation $\chi = \epsilon_+/(1 - \epsilon_+ - \epsilon_-)$ introduced in the text, we have shown that if $\delta < \frac{1}{9p_{max}^*}(1 - \epsilon_+ - \epsilon_-)^3 \min[\alpha p_{min}^2 + \beta \chi^2, \beta(p_{min} - \chi^2)]$, then the statement of Reconstruction Theorem 1 holds. \square

References

- [1] S. E. Shackney, T. V. Shankey, Common patterns of genetic evolution in human solid tumors, *Cytometry* 29 (1997) 1–27.
- [2] B. Vogelstein, E. R. Fearon, S. R. Hamilton, S. E. Kern, A. C. Preisinger, M. Leppert, Y. Nakamura, R. White, A. M. M. Smits, J. L. Bos, Genetic alterations during colorectal tumor development, *New England Journal of Medicine* 319 (1988) 525–532.
- [3] E. R. Fearon, B. Vogelstein, A genetic model for colorectal tumorigenesis., *Cell* 61 (1990) 759–767.
- [4] A. Szabo, A. Yakovlev, Preferred sequences of genetic events in carcinogenesis: Quantitative aspects of the problem, *Journal of Biological Systems* 9 (2).
- [5] R. Desper, F. Jiang, O.-P. Kallioniemi, H. Moch, C. H. Papadimitriou, A. A. Schäffer, Inferring tree models for oncogenesis from comparative genome hybridization data, *Journal of Computational Biology* 6 (1) (1999) 37–51.
- [6] Y. J. Chu, T. H. Liu, On the shortest arborescence of a directed graph, *Sci Sinica* 14 (1965) 1396–1400.
- [7] J. Edmonds, Optimum branchings, *Journal of Research of the National Bureau of Standards* 71B (1967) 233–240.
- [8] R. M. Karp, A simple derivation of Edmonds’ algorithm on optimum branching, *Networks* 1 (1971) 265–272.
- [9] R. E. Tarjan, Finding optimum branchings, *Networks* 7 (1977) 25–35.
- [10] S. M. Ross, *Probability models for computer science*, Harcourt/ Academic Press, Burlington, MA, 2002.
- [11] A. Kallioniemi, O. P. Kallioniemi, D. Sudar, D. Rutovitz, J. W. Gray, F. Waldman, D. Pinkel, Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors, *Science* 258 (5083) (1992) 818–821.
- [12] F. Jiang, J. Richter, P. Schraml, L. Bubendorf, T. Gasser, M. J. Mihatsch, G. Sauter, H. Moch, Chromosomal imbalances in papillary renal cell carcinoma: genetic differences between histological subtypes, *American Journal of Pathology* 153 (1998) 1467–1473.
- [13] F. Jiang, R. Desper, C. Papadimitriou, A. Schäffer, J. Richter, P. Schraml, O.-P. Kallioniemi, H. M. M. J. Mihatsch, Distance-based reconstruction of tree models for oncogenesis, *Cancer Research* 60 (22) (2000) 6503–6509.
- [14] G. M. Brodeur, A. A. Tsiatsis, D. L. Williams, F. W. Luthardt, A. A. Green, Statistical analysis of cytogenic abnormalities in human cancer cells, *Cancer Genet Cytogenet* 7 (1982) 137–152.
- [15] H. Akaike, A new look at the statistical model identification, *IEEE Trans Automatic Control* 19 (1974) 716–723.

- [16] R. Simon, R. Desper, C. H. Papadimitriou, A. Peng, D. S. Alberts, R. Taetle, A. A. Trent, J. M. Schäffer, Chromosome abnormalities in ovarian adenocarcinoma: III. Using breakpoint data to infer and test mathematical models for oncogenesis, *Genes, Chromosomes & Cancer* 28 (2000) 106–120.