Maximum likelihood estimation of oncogenetic tree models with an application to clear cell renal cell carcinoma

Anja von Heydebreck*, Bastian Gunawan†,
and László Füzesi†

* Max Planck Institute for Molecular Genetics, Department of Computational Molecular Biology, D–14195 Berlin, Germany
† Institute of Pathology, Georg August University Hospital, D–37075 Göttingen, Germany

Corresponding author:
Anja von Heydebreck, Phone: ++49-30-84131168, Fax: ++49-30-84131152,
E-mail: heydebre@molgen.mpg.de

Abstract

We present a new mathematical approach for modeling the occurrence of genetic changes in human tumors over time. In solid tumors, data on genetic alterations are usually only available at a single point in time, allowing no direct insight into the sequential order of genetic events. In our approach, genetic tumor development and progression is assumed to follow a probabilistic tree model. We show how maximum likelihood estimation can be used to reconstruct a tree model for the genetic evolution of a given tumor type. Applying our method to cytogenetic data from 173 cases of clear cell renal cell carcinoma, we arrive at a model of karyotypic evolution identifying primary alterations as well as two divergent pathways of secondary alterations.

Introduction

It is well established that the development of human cancers is associated with an accumulation of genetic changes resulting in deregulation of cell proliferation and behavior [1, 2]. For many tumor types, above all hematologic malignancies and to a lesser degree solid mesenchymal tumors, characteristic alterations have been identified, both on the cytogenetic and on the molecular-genetic level. Much less conclusive information, on the other hand, is available on most malignant epithelial tumors, which are frequently characterized by highly aberrant and unbalanced karyotypes dominated by a large variety of chromosomal losses and gains (Mitelman Database of Chromosome Aberrations in Cancer; http://cgap.nci.nih.gov/Chromosomes/Mitelman). The high level of karyotypic complexity often renders the identification of critical cytogenetic events and their evolutionary pathways in these tumors extremely difficult. Thus, it is a challenging problem to infer which genetic changes are typically early or late events in tumor development and progression, and whether there are preferred sequential orders of these events.

The modeling of tumorigenesis as a sequence of genetic changes was pioneered by Fearon and Vogelstein [3]. They analyzed colorectal tumors of different stages, from early adenomas to metastatic carcinomas, and found certain genetic alterations predominantly in advanced tumors, whereas others were present also in earlier disease stages. From these observations, they inferred a qualitative path model describing the typical sequential order of genetic changes. However, Fearon and Vogelstein noted that the observed genetic events did not always follow the proposed sequential order. This fact, together with the common perception that genetic changes in cancer cells are in some sense chance events, motivates the use of probabilistic modeling techniques in the analysis of genetic tumor development and progression.

Attempts to model the sequential order of genetic changes in other types of solid tumors have led to less conclusive results. The analysis of karyotype data from solid tumors has suggested that simple path models may often be insufficient to represent the complex dependencies between non–random alterations [4, 5, 6, 7, 8]. A certain alteration may alter the probability of one or several subsequent alterations, each of which may in turn influence further genetic changes. Thus, several defined genetic pathways may exist instead of a random accumulation or a mere linear order of genetic events. Situations like these are naturally captured by probabilistic tree models, which were introduced in this context by Desper et al. [9, 10] as a generalization of path models. For the reconstruction of tree models from observed data, they, as well as Szabo and Boucher [11], employed algorithms from computer science and phylogeny that do not explicitly take the probabilistic nature of the model into account. In contrast, a standard method

to fit probabilistic models to given data is the maximum likelihood estimation of their parameters. Here, we show how oncogenetic tree models as described in [10] can be reconstructed by maximum likelihood estimation. In particular, we demonstrate how the maximum likelihood parameters for a given binary tree topology can be obtained in closed form. We apply our method to cytogenetic data from 173 cases of clear cell renal cell carcinoma (RCC), arriving at a model of karyotypic evolution that identifies early cytogenetic changes, as well as divergent pathways of secondary alterations. Furthermore, we discuss how the uncertainty of the inferred models can be assessed using the bootstrap.

The model

We use tree models as proposed by Desper et al. [10] to model the occurrence of genetic or cytogenetic alterations in a given tumor type. We start with a set of genetic alterations that are considered as relevant for the development and progression of this tumor type. Our goal is to model the structure of dependencies between these alterations. An example for an oncogenetic tree model is shown in Fig. 1. The genetic alterations correspond to the leaves of the tree. The root of the tree represents the state of a normal cell, and the inner nodes can be interpreted as hidden events that cannot be observed. To each edge e of the tree, a probability p_e is assigned. The evolution of each individual tumor is regarded as a realization of the following random experiment: In the beginning, none of the events corresponding to the nodes of the tree has occurred. Then, starting at the root of the tree, the events at the following nodes occur with probabilities specified by the model: Given that an event corresponding to an inner node of the tree has occurred, the events at the children of the node occur independently with the corresponding edge probabilities p_e. On the other hand, if an event corresponding to an inner node does not occur, then also the events at the children of the node do not occur. Finally, the outcome of the random experiment, which corresponds to a single tumor, is recorded as the subset of leaf events that has occurred.

The model yields a probability distribution on the set of subsets of leaves of the tree, which correspond to the possible subsets of alterations. We assume that $p_e > 0$ for all edges e; otherwise, some leaf events would occur with probability zero. If $p_e < 1$ for all edges e, then every subset of leaves has a positive probability of being observed, and none of the visible genetic events is a necessary precursor of any other. On the other hand, tree models as described in [9], where also the inner nodes of the tree correspond to genetic events as necessary precursors of further alterations, are obtained within this framework by setting appropriate probabilities p_e equal to 1.

Positive correlation between aberrations is marked by the intersection of the paths between the corresponding leaves and the root. According to the model, events

close to the root are likely to be early events in tumor development.
We use the following notation: An *oncogenetic tree*

$$\mathcal{T} = (V, E, r, (p_e), L)$$

is given by a vertex set V, an edge set E, a root vertex $r \in V$, and probabilities p_e assigned to the edges $e \in E$. $L \subset V$ is the set of leaves of the tree. Each edge is given as an ordered pair of vertices $e = (e^1, e^2)$, with e^1 being the vertex closer to the root r. For each edge $e \in E$, we denote the set of children edges by $Ch(e)$ and the parent edge by $Par(e)$. Likewise, $Ch(v)$ and $Par(v)$ denote the children and parent vertices of a vertex $v \in V$. We encode the events corresponding to the nodes $v \neq r$ of the tree by random variables X_v with values in $\{0, 1\}$, such that

$$\begin{aligned} \mathrm{P}(X_v = 1 | X_{Par(v)} = 1) &= p_{e_v} \quad \text{and} \\ \mathrm{P}(X_v = 1 | X_{Par(v)} = 0) &= 0. \end{aligned}$$

Here, e_v denotes the edge with $e_v^2 = v$. In addition, $X_r = 1$.

Maximum likelihood estimation

We now want to infer a tree model describing the genetic evolution of a specific tumor type from given genetic data. We assume that each tumor is genetically characterized only at one point in time, and thus we cannot directly observe the sequential order of alterations in individual tumors. Instead we are going to model the genetic tumor evolution using genetic information on different tumors at various stages, i.e., primary tumors of different disease stages and local or distant relapse tumors. We assume that a set of genetic alterations that might be relevant for this tumor type has been established. These alterations, which will correspond to the leaves of the tree to be inferred, may comprise all kinds of genetic events, for instance chromosomal imbalances, i.e., net gains and losses. The selection of this set of relevant alterations may be performed on the basis of biological knowledge or by statistical criteria [12]. In the data sets under consideration, each individual tumor is classified with respect to presence or absence of these alterations. The data are summarized as a binary matrix $X = (x_{ij})$ with the rows representing alterations and the columns representing tumors: $x_{ij} = 1$ if and only if alteration i is observed in tumor j. We call the binary vector $x_{\cdot j}$ assigned to a tumor j its *profile*. As we are going to model the genetic evolution of tumors of a certain type, one might first analyze the data for the presence of different subtypes that should be modeled separately [13].
Desper et al. [10] estimated distances among each pair of genetic events based on their pairwise frequencies and used distance–based methods from phylogeny to infer tree models. We will show how the method of maximum likelihood can be

used for this purpose. The maximum likelihood estimation of tree models consists of two parts: First, we have to know how to identify the maximum likelihood parameter values for a given tree topology. Second, we have to find the tree topology maximizing the likelihood.

Computing the likelihood of a tree. We now describe how to compute the likelihood of a given tree model efficiently. The likelihood of an oncogenetic tree $\mathcal{T} = (V, E, r, (p_e), L)$ is defined as the probability of the observed data X under this model:

$$L(\mathcal{T}; X) = \mathrm{P}(X|\mathcal{T}).$$

As we regard different tumors as independent realizations of the model, the likelihood of the tree is given as the product of the probabilities of the individual tumors' profiles. In general, a certain profile may be explained by more than one state vector for the hidden events corresponding to the inner nodes of the tree, and the probability of a tumor's profile can be computed by summing over these state vectors. However, the probability of a profile can also be computed more efficiently, as we will show now. For each edge $e \in E$, let L_e be the set of leaves of the subtree $\mathcal{T}'$ of $\mathcal{T}$ that is rooted at e^2. We introduce new parameters

$$q_e := \mathrm{P}(X_l = 0 \quad \forall l \in L_e | X_{e^1} = 1).$$

Thus, q_e is the conditional probability that none of the leaf events corresponding to the subtree $\mathcal{T}'$ occurs, given that the hidden event associated to e^1 has occurred. The parameters q_e can be calculated recursively as follows. If e is a leaf edge, that is, $Ch(e) = \emptyset$, then $q_e = 1 - p_e$. Else,

$$q_e = (1 - p_e) + p_e \prod_{k \in Ch(e)} q_k. \tag{1}$$

Note that

$$p_e > 0 \quad \forall e \in E \quad \Rightarrow \quad q_e < 1 \quad \forall e \in E.$$

Now the probability of a tumor's profile $x_{\cdot j}$ can be computed as a product of the parameters p_e and q_e: If tumor j has at least one alteration, let $\mathcal{T}_j$ be the subtree of $\mathcal{T}$ rooted at r and spanned by the set $L_j = \{l \in L | x_{lj} = 1\}$ of leaves that correspond to the observed events in tumor j (here we identify the set L with the set of row indices of X). Let E_{j1} be the set of edges of $\mathcal{T}_j$, and $E_{j2} = \{e \in E \setminus E_{j1} | Par(e) \in E_{j1}\}$ the set of edges not in $\mathcal{T}_j$, for which the parent edge belongs to $\mathcal{T}_j$. In the case that a tumor has no alterations, that is, $L_j = \emptyset$, we set $E_{j1} = \emptyset$ and $E_{j2} = \{e \in E : e^1 = r\}$. Now,

$$\mathrm{P}(x_{\cdot j}) = \prod_{e \in E_{j1}} p_e \prod_{e \in E_{j2}} q_e. \tag{2}$$

For instance, in the example tree in Fig. 1, the probability of observing the aberrations +Xp and -10p (and no others) in a tumor is given by the product $p_1 p_3 p_6 p_7 q_2$. The likelihood $L(\mathcal{T}; X)$ of the tree $\mathcal{T}$ is now given as the product of the probabilities in Eqn. (2) over all tumors:

$$\begin{aligned} L(\mathcal{T}; X) &= \prod_j \Big(\prod_{e \in E_{j1}} p_e \prod_{e \in E_{j2}} q_e \Big) \\ &= \prod_{e \in E} p_e^{m_e} q_e^{n_e}, \end{aligned} \tag{3}$$

where $m_e = \#\{j | e \in E_{j1}\}$ and $n_e = \#\{j | e \in E_{j2}\}$.

Maximizing the likelihood for a fixed tree topology. In Eqn. (3), we have expressed the likelihood of a tree as a function of the edge parameters p_e and q_e. The recursion in Eqn. (1) allows to write the likelihood as a function of the p_e, which could be used in order to identify the maximum likelihood parameters. On the other hand, the parameters p_e may also be expressed in terms of the parameters q_e. Eqn. (1) yields

$$p_e = \begin{cases} \frac{1-q_e}{1-\prod_{k \in Ch(e)} q_k} & : \quad e^2 \notin L \\ 1 - q_e & : \quad e^2 \in L \end{cases} \tag{4}$$

Note that this is always well–defined because $q_e < 1$ for all edges e. Now we can write the likelihood of $\mathcal{T}$ as a function of the parameters q_e:

$$\begin{aligned} L(\mathcal{T}; X) &= \prod_{e \in E} p_e^{m_e} q_e^{n_e} \\ &= \prod_{e: e^2 \notin L} \left(\frac{1 - q_e}{1 - \prod_{k \in Ch(e)} q_k} \right)^{m_e} q_e^{n_e} \prod_{e: e^2 \in L} (1 - q_e)^{m_e} q_e^{n_e} \\ &= \prod_{e: e^2 \notin L} \frac{\prod_{k \in Ch(e)} ((1 - q_k)^{m_k} q_k^{n_k})}{(1 - \prod_{k \in Ch(e)} q_k)^{m_e}} \times \\ &\quad \times \prod_{k \in E: Par(k) = \emptyset} ((1 - q_k)^{m_k} q_k^{n_k}). \end{aligned}$$

Thus, the likelihood splits up into a product where each factor contains only the parameters q_k from a set of sibling edges. Therefore, the likelihood $L(\mathcal{T}; X)$ is maximized by maximizing each of the factors

$$f_e(\boldsymbol{q}) = \frac{\prod_{k \in Ch(e)} ((1 - q_k)^{m_k} q_k^{n_k})}{(1 - \prod_{k \in Ch(e)} q_k)^{m_e}}, \tag{5}$$

where $\boldsymbol{q} = (q_k)_{k \in Ch(e)}$, for all non–leaf edges e, and

$$g_k(q_k) = (1 - q_k)^{m_k} q_k^{n_k}$$

for the edges k at the root of $\mathcal{T}$.
For any edge k with $k^1 = r$, the value $\hat{q}_k$ maximizing g_k is easily seen to be $n_k/(m_k + n_k)$. Now suppose an edge e has exactly two children e_1, e_2. If $n_{e_1} < m_{e_2}$ and $n_{e_2} < m_{e_1}$, the maximum of f_e is attained at $(\hat{q}_1, \hat{q}_2) = (n_{e_1}/m_{e_2}, n_{e_2}/m_{e_1})$ (see Appendix). These values can be interpreted as conditional relative frequencies: m_{e_2} is the number of tumors for which we know that $X_{e_2^2} = 1$, and $n_{e_1} \leq m_{e_2}$ is the number of tumors among these for which additionally none of the leaf events of the subtree rooted at $X_{e_1^2}$ is observed.
In the case that e has more than two children, we use numerical optimization to determine the maximizing values $(\hat{q}_k)_{k \in Ch(e)}$ of f_e. From the resulting parameters $\hat{q}_k$, we obtain parameters $\hat{p}_k$ according to Eqn. (4). If $0 < \hat{p}_k \leq 1$ for all k, these are the maximum likelihood parameters for the given tree. If $\hat{p}_k > 1$ for some k, we consider this as evidence that edge k does not exist, which corresponds to $p_k = 1$. Also if $n_{e_1} = m_{e_2}$ and thus also $n_{e_2} = m_{e_1}$ for the two children e_1, e_2 of an edge e (see preceding paragraph), we conclude that there is no support for an edge e of positive length. This is the – according to the tree model unlikely – case where any given tumor has only aberrations belonging to one of the subtrees rooted at e, but not from both. If necessary, we shrink all edges corresponding to the mentioned exceptional cases to length zero, arriving at a non–binary tree, and again compute the maximum likelihood parameters for this tree. If necessary, this procedure is iterated until $0 < \hat{p}_k \leq 1$ for all edges k.

Searching for the maximum likelihood tree topology. In principle, the maximum likelihood tree model may simply be obtained by computing the maximum likelihood parameters for all possible tree topologies. The number of rooted binary trees with n leaves is [14]

$$\prod_{i=3}^{n} (2i - 3).$$

For instance, with $n = 8$ variables, there are 135,135 different binary tree topologies, whereas for $n = 12$, their number is already larger than 10^{10}.
If the number of possible tree topologies prohibits an exhaustive search, we apply a heuristic that is also commonly used in maximum likelihood phylogeny estimation [15]. Generally, we consider only binary trees, although in some cases, estimated edge lengths of zero (corresponding to parameters $\hat{p}_e = 1$) may occur. We build the tree by adding one leaf at a time in random order, where the new

leaf is joined with the edge giving the highest likelihood. After each leaf insertion step, we try to improve the obtained tree topology through rearrangements: We cut the tree at a certain edge, obtaining subtrees $\mathcal{T}_1$ (containing the root of $\mathcal{T}$) and $\mathcal{T}_2$. The subtrees are then merged by joining the root of $\mathcal{T}_2$ with an edge of $\mathcal{T}_1$. In each rearrangement step, all possible combinations of edges for cutting and merging are searched for the one giving the highest likelihood. Rearrangements are iterated until there is no more improvement in the likelihood. This procedure is not guaranteed to yield the maximum likelihood tree, because it may get stuck in a local optimum. However, for different orders of the leaves to be inserted, we have always found the same resulting tree, which suggests that this is indeed the maximum likelihood tree.

Results

We applied our method to cytogenetic data from 173 cases of clear cell RCC, comprising 151 primary tumors of different stages, 19 metastases and 3 local recurrences. Karyotypes were determined by classical cytogenetic analysis as described previously [16]. The data were summarized as net changes of chromosome arms in relation to the underlying ploidy level, where a chromosome arm was considered to be gained or lost if this was observed for at least a part of the arm (see Supporting Information). Furthermore, polyploidization was included in the set of genetic events. For the following analysis, we selected all stemline alterations observed in more than 10% of all tumors. However, in cases where the gains or the losses of both arms of a chromosome fulfilled this criterion, we only selected the more frequent of the two in order to avoid highly dependent events due to whole chromosome gains or losses.

In order to see whether there is any reason of considering a tree model for the present data, we analyzed whether statistically significant dependencies between aberrations are present at all. In particular, we wanted to see whether the observed data could be explained by a model where the probability of any given aberration depends only on the number of other aberrations, with no further dependence among pairs of aberrations. For this purpose, we randomly shuffled the entries in the data matrix X, with the restriction that the numbers of aberrations per tumor as well as the frequencies of the individual aberrations remain unchanged. Thus, any dependence between two aberrations in the randomized data sets should be due to their common tendency to occur in tumors with the same number of other aberrations. For each of 1000 shuffled data matrices X^*, we calculated the number of pairs of aberrations with $p < 0.05$ in the Fisher–test of independence. All of the shuffled data sets exhibited fewer significant pairs than the real data, which indicates a significant deviation from a conditional independence between aberrations as described above (estimated $p < 0.001$).

The resulting maximum likelihood tree model for clear cell RCC is shown in Fig. 2. The edge parameters p_e are encoded as branch lengths in the horizontal direction, with the branch length of edge e being proportional to $-\log(p_e)$. The loss of 3p, presumably a primary event, which is present in 170 out of the 173 tumors, is placed close to the root of the tree. A second early event seems to be gain of 5q, which was often observed together with loss of 3p in the form of an unbalanced translocation der(3)t(3;5) [16]. In an earlier study comprising 118 primary tumors of the present series, gain of 5q was shown to be significantly correlated with longer patient survival [16]. The next branching in the tree separates two clusters of aberrations, remarkably with one characterized exclusively by chromosomal losses and the polyploidization event, and the other one comprising only chromosomal gains. These gains are also frequently seen in the papillary type of RCC, which is recognized as a RCC variant with distinct genetic and clinico–pathological features [17]. The cluster comprising chromosomal losses contains several events, including losses of 8p, 9p, and 14q, which have been associated with the progression and clinical outcome of clear cell RCCs [18, 19].
The tree model indicates that apart from the primary loss of 3p, no aberration is a necessary precursor of any other, which would correspond to certain genetic events being very close to inner nodes of the tree. For instance, although gain of 5q may often be an early event preceding later changes, none of these further alterations is seen only in conjunction with gain of 5q. Generally, the oncogenetic tree models considered here are able to reflect how strongly the data deviate from fixed sequential orders of genetic events.

Validation. Now we address the question whether the tree models described here can adequately model the evolution of genetic changes during tumor development and progression. First, we analyzed whether the dependencies between genetic changes in the clear cell RCC data are well characterized by the inferred tree model. For this purpose, we compared the observed relative frequencies of all pairs and triplets of aberrations with the corresponding probabilities given by the maximum likelihood tree model $\hat{T}$. The deviations between observed and expected frequencies were quantified by the χ^2–statistic, resulting in deviation measures χ^2_2 and χ^2_3 for pairs and triplets of aberrations, respectively. Then we generated 1000 data sets X^* according to the inferred tree model $\hat{T}$ and computed deviation measures χ^{2*}_2 and χ^{2*}_3 in the same way as for the real data. We argue that a poor model fit would likely lead to the real data showing a larger deviation from the model probabilities than many of the data sets generated according to the inferred model. We observed χ^2_2 to be smaller than 94% of the χ^{2*}_2, and χ^2_3 to be smaller than 47% of the χ^{2*}_3, indicating no deviation from the model.
With data sets of limited size, the question arises whether a reliable estimation of

the dependencies between the genetic variables is possible. We used the nonparametric bootstrap [20] to assess the uncertainty of properties of the estimated tree model. In each bootstrap iteration, a new data set was generated by sampling with replacement from the set of tumors. Maximum likelihood trees $\hat{T}^*$ were computed for each bootstrap data set, and a bootstrap confidence value for a property of the original maximum likelihood tree $\hat{T}$ was calculated as the percentage of bootstrap trees having this property. Especially, we assigned bootstrap confidence values to the edges of $\hat{T}$, providing a measure of uncertainty for the presence of each cluster of aberrations. In Fig. 2, confidence values for the internal edges of the maximum likelihood tree for the clear cell RCC data (based on 500 bootstrap data sets) are given. We can see that the proposed tree structure has to be interpreted with some caution. With the given number of tumors, an accurate estimation of the complex dependencies between 14 genetic events is hardly possible. Nevertheless the model in Fig. 2 can at least serve an exploratory purpose, allowing to formulate hypotheses about the evolution of karyotypes in clear cell RCC.

Discussion

We have described a mathematical method to model the occurrence of characteristic changes in tumor development and progression. While we have demonstrated the use of our approach on cytogenetic data, the method is as well applicable to data sets obtained from other experimental techniques used to determine chromosomal aberrations, gene mutations, or possibly also epigenetic changes in tumor cells. A computer program is available upon request from the first author.

Our approach of maximum likelihood estimation of oncogenetic tree models is similar to the maximum likelihood estimation of phylogenetic trees. In contrast to the phylogeny problem, however, we can determine the maximum likelihood parameters for a given binary tree topology in closed form. This is due to our assumption that the hidden events occur as necessary precursors of observed genetic events, which provides us with more information about the states of the inner nodes of a tree than what is available in the case of phylogenetic trees.

An important question is whether tree models are an adequate family of models for the genetic evolution of tumors, or whether more complex models are needed. Indeed one may imagine situations where the dependencies between aberrations could not be captured by a tree model, for instance in the case of several converging pathways, where certain late changes are common to a variety of tumors characterized by different early aberrations [4, 5]. This could be a reason for considering a wider family of graph models. However, one has to bear in mind that this would lead to an increased model complexity, making it more difficult to reliably reconstruct models from data sets of limited size. This is especially a problem for karyotype data from many types of solid tumors, where a large vari-

ety of recurrent aberrations can be observed, but data are often available from at most a few hundred cases. For this reason, we would only consider a wider family of graph models if there were evidence that tree models could not explain the dependencies between alterations in a given data set. In any case, the framework of likelihood–based estimation allows for a systematic comparison of the proposed family of tree models with other classes of probabilistic models that might be used in this context.

Appendix

Let e be an edge of the tree $\mathcal{T}$ that has exactly two children e_1, e_2, with corresponding non–negative integers $m_{e_1}, n_{e_1}, m_{e_2}, n_{e_2}, m_e$ according to Eqn. (3).

Proposition. *If $n_{e_1} < m_{e_2}$ and $n_{e_2} < m_{e_1}$, the maximum of*

$$\begin{aligned} f_e : [0,1)^2 &\rightarrow \mathbb{R} \\ (x_1, x_2) &\mapsto \frac{(1-x_1)^{m_{e_1}} x_1^{n_{e_1}} (1-x_2)^{m_{e_2}} x_2^{n_{e_2}}}{(1-x_1x_2)^{m_e}} \end{aligned}$$

(see Eqn. (5)) is attained at

$$(\hat{x}_1, \hat{x}_2) = (n_{e_1}/m_{e_2}, n_{e_2}/m_{e_1}).$$

Proof. First we consider the case that $n_{e_1} > 0$ and $n_{e_2} > 0$. Setting the gradient of $\log(f_e)$ to zero yields the following equations:

$$\begin{aligned} n_{e_1}/x_1 - m_{e_1}/(1-x_1) + m_e x_2/(1-x_1x_2) &= 0 \\ n_{e_2}/x_2 - m_{e_2}/(1-x_2) + m_e x_1/(1-x_1x_2) &= 0 \end{aligned}$$

Using the relations $n_{e_1} + m_{e_1} = m_e = n_{e_2} + m_{e_2}$, the unique solution is obtained as $(\hat{x}_1, \hat{x}_2) = (n_{e_1}/m_{e_2}, n_{e_2}/m_{e_1})$. Taking into account that $m_{e_1} + m_{e_2} > m_e$, it is easily shown that f_e converges to zero at all points of the boundary of $(0,1)^2$, which proves the claim in this case. Now suppose $n_{e_1} = 0$, which implies $m_{e_1} = m_e$. It is easily seen that in this case, the maximum of f_e has to be attained at a point with $x_1 = 0$. Considering the derivative with respect to x_2 yields the solution $(0, n_{e_2}/m_{e_1})$ (analogously for $n_{e_2} = 0$). □

Acknowledgments. We thank W. Huber, D. Buschmann and M. Vingron for helpful discussions. A.v.H. was supported by a grant from the German Ministry of Science (BMBF) within the German Human Genome Project (DHGP).

References

[1] P.C. Nowell. The clonal evolution of tumor cell populations. *Science*, 194:23–28, 1976.

[2] G. Klein and E. Klein. Evolution of tumours and the impact of molecular oncology. *Nature*, 315:190–195, 1985.

[3] E.R. Fearon and B. Vogelstein. A genetic model for colorectal tumorigenesis. *Cell*, 61:759–767, 1990.

[4] M. Höglund, D. Gisselsson, N. Mandahl, B. Johansson, F. Mertens, F. Mitelman, and T. Säll. Multivariate analyses of genomic imbalances in solid tumors reveal distinct and converging pathways of karyotypic evolution. *Genes Chromosomes Cancer*, 31:156–171, 2001.

[5] M. Höglund, D. Gisselsson, T. Säll, and F. Mitelman. Coping with complexity: multivariate analysis of tumor karyotypes. *Cancer Genet. Cytogenet.*, 135:103–109, 2002.

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

[7] F. Jiang, R. Desper, C.H. Papadimitriou, A.A. Schäffer, O.P. Kallioniemi, J. Richter, P. Schraml, G. Sauter, M.J. Mihatsch, and H. Moch. Construction of evolutionary tree models for renal cell carcinoma from comparative genomic hybridization data. *Cancer Res.*, 60:6503–6509, 2000.

[8] M.D. Radmacher, R. Simon, R. Desper, R. Taetle, A.A. Schäffer, and M.A. Nelson. Graph models of oncogenesis with an application to melanoma. *J. Theor. Biol.*, 212:535–548, 2001.

[9] R. Desper, F. Jiang, O.P. Kallioniemi, H. Moch, C.H. Papadimitriou, and A.A. Schäffer. Inferring tree models for oncogenesis from comparative genome hybridization data. *J. Comput. Biol.*, 6:37–51, 1999.

[10] R. Desper, F. Jiang, O.P. Kallioniemi, H. Moch, C.H. Papadimitriou, and A.A. Schäffer. Distance-based reconstruction of tree models for oncogenesis. *J. Comput. Biol.*, 7:789–803, 2000.

[11] A. Szabo and K. Boucher. Estimating an oncogenetic tree when false negatives and positives are present. *Math. Biosci.*, 176:219–236, 2002.

[12] G.M. Brodeur, A.A. Tsiatis, D.L. Williams, F.W. Luthardt, and A.A. Green. Statistical analysis of cytogenetic abnormalities in human cancer cells. *Cancer Genet. Cytogenet.*, 7:137–152, 1982.

[13] M.A. Newton. Discovering combinations of genomic alterations associated with cancer. *J. Am. Stat. Assoc.*, 97:931–942, 2002.

[14] E. Schröder. Vier combinatorische Probleme. *Z. Math. Phys.*, 15:361–376, 1870.

[15] J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17:368–376, 1981.

[16] B. Gunawan, W. Huber, M. Holtrup, A. v.Heydebreck, T. Efferth, A. Poustka, R.-H. Ringert, G. Jakse, and L. Füzesi. Prognostic impacts of cytogenetic findings in clear cell renal cell carcinoma: Gain of 5q31–qter predicts a distinct clinical phenotype with favorable prognosis. *Cancer Res.*, 61:7731–7738, 2001.

[17] G. Kovacs, M. Akhtar, B.J. Beckwith, P. Bugert, C.S. Cooper, B. Delahunt, J.N. Eble, S. Fleming, B. Ljungberg, and L.J. Medeiros et al. The Heidelberg classification of renal cell tumours. *J. Pathol.*, 183:131–133, 1997.

[18] H. Moch, J.C. Presti, G. Sauter, N. Buchholz, P. Jordan, M.J. Mihatsch, and F.M. Waldman. Genetic aberrations detected by comparative genomic hybridization are associated with clinical outcome in renal cell carcinoma. *Cancer Res*, 56:27–30, 1996.

[19] D. Schullerus, J. Herbers, J. Chudek, H. Kanamaru, and G. Kovacs. Loss of heterozygosity at chromosomes 8p, 9p, and 14q is associated with stage and grade of non-papillary renal cell carcinomas. *J. Pathol.*, 183:151–155, 1997.

[20] B. Efron, E. Halloran, and S. Holmes. Bootstrap confidence levels for phylogenetic trees. *Proc. Natl. Acad. Sci. USA*, 93:13429–13434, 1996.

Figure legends

Figure 1. Example of an oncogenetic tree. The leaves correspond to genetic alterations, whereas the inner nodes can be interpreted as hidden events preceding the visible genetic changes. The model is parameterized by conditional probabilities assigned to the edges (see text).

Figure 2. Maximum likelihood tree model for the karyotypic evolution of clear cell RCC, based on the chromosomal aberrations seen in more than 10% of the 173 cases. The length of each horizontal edge e is proportional to $-\log(p_e)$. Bootstrap confidence values (in percent) for the inner edges are given.

Figure 1

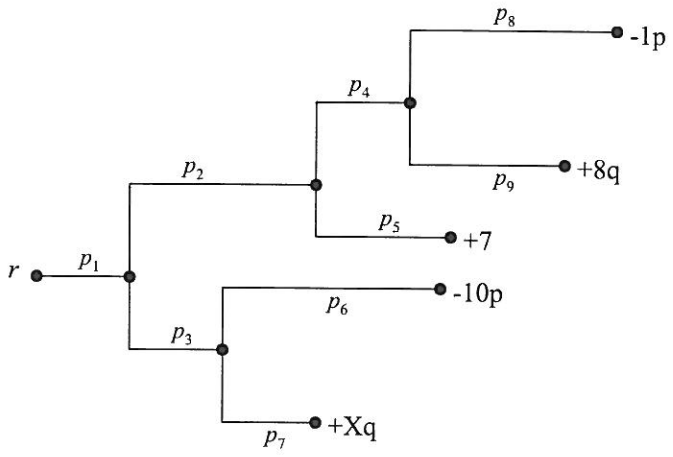
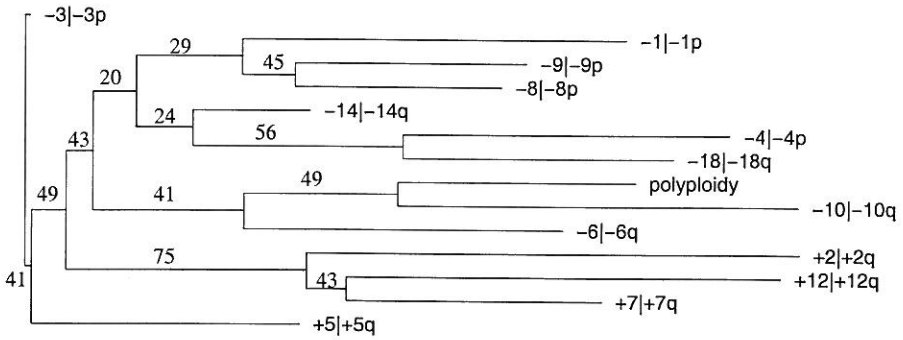


Figure 2


−3|−3p
29
−1|−1p
45
−9|−9p
20
−8|−8p
24
−14|−14q
43
56
−4|−4p
−18|−18q
49
49
polyploidy
41
−10|−10q
−6|−6q
75
+2|+2q
41
43
+12|+12q
+7|+7q
+5|+5q