Maximum likelihood estimation of oncogenetic tree models

ANJA VON HEYDEBRECK
Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics,
D-14195 Berlin, Germany
anja.von.heydebreck@merck.de

BASTIAN GUNAWAN, LÁSZLÓ FÜZESI
Institute of Pathology, Georg August University Hospital, D–37075 Göttingen, Germany

SUMMARY
We present a new approach for modelling the dependences between genetic changes in human tumours. In solid tumours, 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 tumour 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 dependences between genetic alterations in a given tumour type. We illustrate the use of the proposed method by applying it to cytogenetic data from 173 cases of clear cell renal cell carcinoma, arriving at a model for the karyotypic evolution of this tumour.

Keywords: Maximum likelihood; Oncogenesis; Tree model.

1. 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 behaviour (Nowell, 1976; Klein and Klein, 1985). For many tumour types, above all hematologic malignancies and to a lesser degree solid mesenchymal tumours, 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 tumours, which frequently show highly aberrant karyotypes characterized by a large variety of chromosomal losses and gains as well as structural rearrangements (Albertson et al., 2003; 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 tumours extremely difficult. Thus, it is a challenging problem to reconstruct dependences between genetic changes that are relevant for tumour development and progression, such as preferred sequential orders of their occurrence.

The modelling of tumourigenesis as a sequence of genetic changes was pioneered by Fearon and Vogelstein (1990). They analysed colorectal tumours of different stages, from early adenomas to metastatic carcinomas, and found certain genetic alterations predominantly in advanced tumours, whereas others were observed also in earlier disease stages. From these observations, Fearon and Vogelstein inferred a qualitative path model describing the typical sequential order of genetic changes. However, they 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 modelling techniques in the analysis of genetic tumour development and progression.

Attempts to model the sequential order of genetic changes in other types of solid tumours have led to less conclusive results. The analysis of karyotype data from solid tumours has suggested that simple path models may often be insufficient to represent the complex dependences between non-random alterations (Simon et al., 2000; Jiang et al., 2000; Radmacher et al., 2001; Höglund et al., 2001, 2002). 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 fixed sequential order of genetic events. Situations like these are naturally captured by probabilistic tree models, which were introduced in this context by Desper et al. (1999, 2000) as a generalization of path models. However, for the reconstruction of their tree models from observed genetic data, Desper et al., as well as Szabo and Boucher (2002), employed algorithms from computer science and phylogenetics that do not explicitly take the probabilistic nature of the model into account. In this article, we show how oncogenetic tree models as described by Desper et al. (2000) can be estimated using the maximum likelihood method. In particular, we demonstrate how the maximum likelihood parameters for a given binary tree topology can be obtained in closed form. We apply the proposed 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.

2. The model

We use oncogenetic tree models as proposed by Desper et al. (2000) to model the occurrence of genetic or cytogenetic alterations in a given tumour type. We start with a set of genetic alterations that are considered as relevant for the development and progression of this tumour type. Our goal is to model the statistical dependences between these alterations. A hypothetical example for an oncogenetic tree model is shown in Figure 1. The genetic alterations (in the example, these are gains or losses of chromosome arms) correspond bijectively 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 of the tree, a probability \( p_e \) is assigned. The evolution of each individual tumour 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 conditional probabilities specified by the model: given that an event corresponding to an inner node \( X \) of the tree has occurred, the event at each child \( X' \) of the node occurs with the corresponding edge probability \( p_e \), independently of the events at all nodes that are not descendants of \( X' \); 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 will not occur. Finally, the outcome of the random experiment, which corresponds to a single tumour, is recorded as the subset of leaf events that have occurred.

More formally, the model of Desper et al. (2000) is defined as follows. An oncogenetic tree model

\[ T = (V, E, r, (p_e), L) \]

is given by a vertex set \( V \), an edge set \( E \), a root vertex \( r \in V \), and probabilities \( p_e > 0 \) assigned to the edges \( e \in E, L \subseteq V \setminus \{r\} \) is the set of leaves of the tree. For each vertex \( v \in V \), we denote the set of children vertices by \( Ch(v) \) and the parent by \( Par(v) \). Likewise, \( Ch(e) \) and \( Par(e) \) denote the children and parent edges of an edge \( e \in E \). Each edge is given as an ordered pair of vertices, \( e = (v_1(e), v_2(e)) \), and we use the notation \( e = (v_1(e), v_2(e)) \) to refer to the vertices of an edge \( e \). Degenerate edges are allowed; they are characterized by \( p_e = 1 \).
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

$$
P(X_v = 1|X_{Par(v)} = 1) = p_v \quad \text{and} \quad P(X_v = 1|X_{Par(v)} = 0) = 0.
$$

In addition, $X_r = 1$. Furthermore, we assume that the joint probability distribution of $(X_v)_{v \in V}$ is given as the product of the conditional distributions in (2.1):

$$
P(X_v)_{v \in V} = \prod_{v \in V \setminus \{r\}} P(X_v|X_{Par(v)}).
$$

Due to this conditional independence property, the tree models considered here are a special class of Bayesian networks with hidden variables. Their special properties are (i) that the underlying graph is a tree, and (ii) the special form of the conditional distributions, see (2.1).

Within an oncogenetic tree model, the marginal probability of each (hidden or observed) event is given as the product of the edge probabilities along the path from the root to the corresponding node. It is natural to assign a branch length of $-\log(p_e)$ to an edge $e$ with conditional probability $p_e$, because the marginal probability of any event is then encoded by the length of the path between the root and the corresponding node. For two nodes $v_1$ and $v_2$,

$$
P(X_{v_1} = 1 \land X_{v_2} = 1) = \frac{P(X_{v_1} = 1)P(X_{v_2} = 1)}{P(X_{v_3} = 1)},
$$

where $v_3$ is the most recent common ancestor of the nodes $v_1$ and $v_2$. Thus two events are independent if the corresponding paths to the root do not overlap; otherwise they are positively correlated.

Now it is easily seen that an oncogenetic tree model is identifiable if each node corresponding to a hidden event has at least two children. Suppose that $(X_v)_{v \in L}$ is the family of leaf events of an oncogenetic tree $T$. The marginal probabilities of the observed events determine the lengths of the paths from the root to the leaves (defined as in the preceding paragraph), and the distribution of any pair of events determines the length of the intersection of the two paths from the root to these leaves. Thus, the branching pattern of the tree and the branch lengths are uniquely determined by the distribution of $(X_v)_{v \in L}$. 
3. Maximum likelihood estimation

We now want to infer a tree model describing the genetic evolution of a specific tumour type from given genetic data. We assume that each tumour in a data set is genetically characterized only at one point in time, and thus we cannot directly observe the sequential order of alterations in an individual tumour. Instead we are going to model the genetic tumour evolution using genetic information on different tumours at various stages, i.e. primary tumours of different disease stages and local or distant relapse tumours. We assume that a set of genetic alterations that might be relevant for this tumour 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 (Brodeur et al., 1982). In the data sets under consideration, each individual tumour is classified with respect to the 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 tumours: $x_{ij} = 1$ if and only if alteration $i$ is observed in tumour $j$. We call the binary vector $x_j$ assigned to a tumour $j$ its profile.

As we are going to model the genetic evolution of tumours of a certain type, one might first analyse the data for the presence of different subtypes that should be modelled separately, using for instance the method of Newton (2002).

Desper et al. (2000) estimated distances among each pair of genetic events based on their pairwise frequencies and used distance-based methods from phylogenetics 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.

3.1 Computing the likelihood of a tree

Assume we are given an oncogenetic tree $T = (V, E, r, (p_e), L)$ and observed data $X$. We now describe how to compute the likelihood

$$L(T; X) = P(X|T)$$

of the tree model efficiently. As we regard different tumours as independent realizations of the model, the likelihood of the tree is given as the product of the probabilities of the individual tumours’ 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 tumour’s profile can be computed by summing over these state vectors. However, the probability of a profile can also be computed more efficiently. The algorithm we describe in the following is a variant of the peeling algorithm for computing the likelihood of a phylogenetic tree (Felsenstein, 1981).

For each vertex $v \in V \setminus \{r\}$, let $L_v$ be the set of leaves of the subtree $T_v$ of $T$ that is rooted at $v$. For the corresponding edge $e = e_v$, let

$$q_e := P(X_l = 0 \ \forall l \in L_v | X_{Par(v)} = 1).$$

Thus, $q_e$ is the conditional probability that none of the leaf events corresponding to the subtree $T_v$ occurs, given that the hidden event associated to $Par(v)$ 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,

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

Note that
\[ p_e > 0 \quad \forall e \in E \quad \Rightarrow \quad q_e < 1 \quad \forall e \in E, \quad \text{and} \]
\[ p_e < 1 \quad \forall e \in E \quad \Rightarrow \quad q_e > 0 \quad \forall e \in E. \quad (3.4) \]

Now the probability of a tumour's profile \( x_j \) can be computed as a product of the parameters \( p_e \) and \( q_e \). If tumour \( j \) has at least one alteration, let \( T_j \) be the subtree of \( T \) 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 tumour \( j \) (here we identify the set \( L \) with the set of row indices of \( X \)). Let \( E_{j1} \) be the set of edges of \( T_j \), and \( E_{j2} = \{ e \in E \setminus E_{j1} | \text{Par}(e) \in E_{j1} \} \) the set of edges not in \( T_j \), for which the parent edge belongs to \( T_j \). In the case that a tumour has no alterations, i.e. \( L_j = \emptyset \), we set \( E_{j1} = \emptyset \) and \( E_{j2} = \{ e \in E : v_1(e) = r \} \). Now,
\[
\Pr(x_j) = \prod_{e \in E_{j1}} p_e \prod_{e \in E_{j2}} q_e. \quad (3.5)
\]

For instance, in the example tree in Figure 1, the probability of observing the aberrations +Xq and -10p (and no others) in a tumour is given by the product \( p_1p_3p_8p_9q_2 \).

As we assume \( p_e > 0 \) for all edges \( e \in E \), every leaf event occurs with a positive probability. If, moreover, \( p_e < 1 \) for all edges \( e \in E \), then every subset of leaf events has a positive probability of being observed, and none of the visible genetic events is a necessary precursor of any other—this is easily seen from (3.4) and (3.5). On the other hand, branching tree models as described by Desper et al. (1999), 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. Branching tree models are much more restricted, as each event is supposed to have a necessary direct precursor, implying that subsets of aberrations that do not respect this sequential order have probability zero of being observed.

The likelihood \( L(T; X) \) of the tree model \( T \) is now given as the product of the probabilities in (3.5) over all tumours \( j \):
\[
L(T; X) = \prod_j \left( \prod_{e \in E_{j1}} p_e \prod_{e \in E_{j2}} q_e \right) = \prod_{e \in E} (p^m_e q^n_e), \quad (3.6)
\]

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

### 3.2 Maximizing the likelihood for a fixed tree topology

In (3.6), we have expressed the likelihood of a tree as a function of the edge parameters \( p_e \) and \( q_e \). The recursion in (3.3) allows us 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 \( q_e \). Equation (3.3) yields
\[
p_e = \begin{cases} 
(1 - q_e)/(1 - \prod_{k \in \text{Ch}(e)} q_k) & : v_2(e) \notin L, \\
1 - q_e & : v_2(e) \in L
\end{cases} \quad (3.7)
\]
Note that this is always well-defined because \( q_e < 1 \) for all edges \( e \). Now we can write the likelihood of \( T \) as a function of the parameters \( q_e \):

\[
L(T; X) = \prod_{e \in E} p_e^{m_e} q_e^{n_e}
\]

\[
= \prod_{e : v_2(e) \notin L} \left( \frac{1 - q_e}{1 - \prod_{k \in Ch(e)} q_k} \right)^{m_e} q_e^{n_e} \prod_{e : v_2(e) \in L} ((1 - q_e)^{m_e} q_e^{n_e})
\]

\[
= \prod_{e : v_2(e) \notin L} \prod_{k \in Ch(e)} ((1 - q_k)q_k^{n_k}) \prod_{e : v_2(e) \in L} ((1 - q_k)^{m_k} q_k^{n_k}). \tag{3.8}
\]

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 \( L(T; X) \) is maximized by maximizing each of the factors

\[
f_e(q) = \prod_{k \in Ch(e)} ((1 - q_k)^{m_k} q_k^{n_k}) \quad \text{for any non-leaf edges } e,
\]

\[
g_k(q_k) = (1 - q_k)^{m_k} q_k^{n_k}
\]

for the edges \( k \) at the root of \( T \).

For any edge \( k \) with \( v_1(k) = r \), the value \( \hat{q}_k \) maximizing \( g_k \) is easily seen to be \( n_k / (m_k + n_k) \). Now suppose a vertex \( v \) has exactly two children \( v_1, v_2 \), with corresponding edges \( e, e_1, e_2 \). If \( n_{e_1} < m_{e_2} \) and \( n_{e_2} < m_{e_1} \), the maximum of \( f_e \) is attained at

\[
(\hat{q}_{e_1}, \hat{q}_{e_2}) = (n_{e_1} / m_{e_1}, n_{e_2} / m_{e_2}) \tag{3.10}
\]

(see Appendix). These values can be interpreted as conditional relative frequencies: \( m_{e_2} \) is the number of tumours for which we know that \( X_{v_2} = 1 \), and \( n_{e_1} \leq m_{e_2} \) is the number of tumours among these for which additionally none of the leaf events of the subtree rooted at \( X_{v_1} \) 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 the parameters \( \hat{p}_e \) by applying (3.7):

\[
\hat{p}_e = \begin{cases} 
(1 - \hat{q}_e) / (1 - \prod_{k \in Ch(e)} \hat{q}_k) & : \ v_2(e) \notin L \\
1 - \hat{q}_e & : \ v_2(e) \in L.
\end{cases} \tag{3.11}
\]

If \( 0 < \hat{p}_e \leq 1 \) for all \( e \in E \), these are the maximum likelihood parameters for the given tree. However, in applications we have occasionally obtained \( \hat{p}_e > 1 \) for some \( e \). For instance, for an edge \( e \) with two children \( e_1, e_2 \), it may be that \( \hat{q}_e = 0 \), meaning that every tumour with an aberration from the subtree rooted at the sibling \( e' \) of \( e \) also has at least one aberration belonging to the subtree rooted at \( e \) (\( n_e = 0 \)), while at the same time \( \hat{q}_{e_1}, \hat{q}_{e_2} \geq 0 \), meaning that there are tumours with aberrations from only one of the subtrees rooted at \( e_1 \) and \( e_2 \), but not from both. We consider this as evidence against the presence of edge \( e \) in the model, suggesting that \( p_e = 1 \). Furthermore, the parameter estimate \( \hat{p}_e \) for an edge \( e \) with two children \( e_1, e_2 \) obtained through (3.11) is not defined if \( n_{e_1} = m_{e_2} \) and thus also \( n_{e_2} = m_{e_1} \) (see preceding paragraph). This is the case if any given tumour has only aberrations belonging to one of the subtrees rooted at \( e \), but not from both. As the \( \hat{q}_e \) are consistent estimators of the \( q_e \), the probability
of obtaining values of $\hat{p}_e$ outside the allowable range converges to zero with increasing sample size if
the data are indeed generated from the tree model under consideration. If necessary, we shrink all edges
corresponding to the above-mentioned exceptional cases to length zero, arriving at a non-binary tree, and
again compute the maximum likelihood parameters for this tree, using numerical optimization for those
edges that have more than two children. If necessary, this procedure is iterated until we arrive at a tree
with $0 < \hat{p}_e \leq 1$ for all edges $e$.

### 3.3 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
(Schröder, 1870)

$$\prod_{i=3}^{n} (2i - 3).$$  \hspace{1cm} (3.12)

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 (Felsenstein, 1981). Generally, we
consider only binary trees, although in some cases, estimated edge lengths of zero (corresponding to
parameters $\hat{p}_e = 1$) may occur. Considering only binary trees is actually no restriction, because all non-
binary trees can be seen as degenerate binary trees with certain edge lengths equal to zero. 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 $T_1$ (containing the root of $T$) and
$T_2$. The subtrees are then merged by joining the root of $T_2$ with an edge of $T_1$. 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 might get stuck in a
local optimum. However, with the data we analysed, we have almost always found the same resulting tree,
irrespective of the order of the leaves to be inserted, suggesting that this is indeed the maximum likelihood
tree. Alternatively, probabilistic optimization algorithms, e.g. the Metropolis algorithm, can be applied.

### 4. RESULTS

We applied our method to cytogenetic data from 173 cases of clear cell RCC, comprising 151 primary
tumours of different stages, 19 metastases and three local recurrences. Karyotypes were determined by
classical cytogenetic analysis as described previously (Gunawan et al., 2001). 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. Furthermore,
polyploidization was included in the set of genetic events. For the following analysis, we selected all
stemline alterations (that is, those occurring in all observed cells of an individual tumour) observed in more
than 10% of all tumours. 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 to consider a tree model for the present data, we analysed
whether statistically significant dependences between aberrations are present. 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 tumour 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 tumours 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 Figure 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 was observed in 170 out of the 173 tumours, is placed close to the root of the tree. The next branching in the tree separates the gain of 5q from the remaining events. Gain of 5q was often observed together with loss of 3p in the form of an unbalanced translocation der(3)t(3;5) (Gunawan et al., 2001), suggesting that this may also often be an early event. In an earlier study comprising 118 primary tumours of the present series, gain of 5q was shown to be significantly correlated with longer patient survival (Gunawan et al., 2001). 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 an RCC variant with distinct genetic and clinico-pathological features (Kovacs et al., 1997). The cluster comprising chromosomal losses contains several events, including losses of 8p, 9p, 14q, and 18q, which have been associated with the progression and clinical outcome of clear cell RCCs (Moch et al., 1996; Schullerus et al., 1997; Gunawan et al., 2001).

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.
4.1 Validation

Now we consider whether the tree models described here can adequately model the evolution of genetic changes during tumour development and progression. First, we analysed whether the dependences 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 $\chi^2$-statistic, resulting in deviation measures $\chi^2_p$ and $\chi^2_t$ 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 $\chi^2_p*$ and $\chi^2_t*$ 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 $\chi^2_p$ to be smaller than 94% of the $\chi^2_p*$, and $\chi^2_t$ to be smaller than 47% of the $\chi^2_t*$, indicating no significant deviation from the model.

With data sets of limited size, the question arises whether a reliable estimation of the dependences between the genetic variables is possible. We used the nonparametric bootstrap (Felsenstein, 1985) 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 tumours. 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. In particular, 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 Figure 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 caution. With the given number of tumours, an accurate estimation of the complex dependences between 14 genetic events is hardly possible. Nevertheless we think that the model in Figure 2 can at least serve an exploratory purpose, allowing us to formulate hypotheses about the evolution of karyotypes in clear cell RCC.

5. DISCUSSION

We have described a mathematical model for the occurrence of characteristic changes in tumour 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 tumour cells. The method is implemented within the statistical computing environment R (Ihaka and Gentleman, 1996) as a package oncomodel, which is available at http://www.molgen.mpg.de/~heydebre.

We have chosen to work with the tree models introduced by Desper et al. (2000), where the observed genetic alterations are represented as leaves of a tree, in contrast to branching tree models (Desper et al., 1999), where also the inner nodes of the tree represent observed events. While the latter models have the appealing property of explicitly proposing sequential orders of genetic events, at the same time this yields problems for a likelihood-based analysis. If only a single tumour in a data set contradicts the partial order of events defined by a tree, the likelihood of the tree is zero. This problem may be partially overcome by introducing measurement error into the model (Szabo and Boucher, 2002), but still the assumption of fixed sequential orders of genetic alterations seems too strong without support from prior biological knowledge.

On the other hand, the tree models according to Desper et al. (2000), which we use in our approach, do not necessarily allow for temporal inference. However, they are able to reflect fixed or strongly preferred sequential orders of aberrations if these are suggested by the data. If an event $A$ occurs almost always
before event $B$, the leaf of the tree representing $A$ will be close to the path from the root to $B$. Therefore we consider this family of models as useful for (i) representing the correlations between aberrations and (ii) detecting preferred sequential orders of events without presupposing them.

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 is available in the case of phylogenetic trees.

One limitation of the tree models considered here is that they cannot explain negative correlation between aberrations. According to an oncogenetic tree model, two aberrations are uncorrelated if the paths from the corresponding leaves to the root of the tree do not overlap, otherwise they are positively correlated. A strong negative correlation between aberrations in different subtrees could indicate different tumour subtypes that should be modelled separately. However, in the example of the RCC data we found no significant negative correlation among pairs of aberrations; the Bonferroni-corrected minimal $p$-value from the respective one-sided Fisher tests for all pairs of aberrations is 0.28.

One may wonder how the estimation of oncogenetic tree models behaves in the presence of errors in the experimental detection of genetic aberrations. Szabo and Boucher (2002) have investigated this question in the case of branching tree models (Desper et al., 1999). Let $\epsilon_i^+$ = $P(\text{event } i \text{ is observed } | \text{ event } i \text{ has not occurred})$ and $\epsilon_i^− = P(\text{event } i \text{ is not observed } | \text{ event } i \text{ has occurred})$ be the false positive/false negative probabilities for aberration $i$. Assuming independence between errors of different aberrations, false negative errors are easily incorporated into our tree models. Suppose the occurrence of aberrations follows a tree model $T$. False negatives may be represented by adding an edge to each leaf node $i$ with conditional probability $1 − \epsilon_i^−$, the probability that $i$ is observed, given that it occurred. Merging the new leaf edges with their respective parents, we arrive at a new tree $T'$ that differs from $T$ only in the parameters of the leaf edges, $p'_i = p_i(1 − \epsilon_i^−)$. Thus, the topology of the tree to be inferred does not change due to the possibility of false negative observations. For a tree estimated from given data, the conditional probability at a leaf edge may therefore be interpreted as the product of the true conditional probability for the occurrence of the event and the detection probability $1 − \epsilon_i^−$. If in addition false positive observations are to be considered, the situation becomes more complicated. One may follow the lines of Szabo and Boucher (2002) and compute upper bounds for the error probabilities which guarantee that the joint distribution of the family of observed aberrations is represented by the same tree topology as the underlying distribution of occurred aberrations. Furthermore, the robustness of estimation procedures to measurement error in the case of practically relevant sample sizes could be evaluated using simulated data.

An important question is whether tree models are an adequate family of models for the genetic evolution of tumours, or whether more complex models are needed. Indeed one may imagine situations where the dependences 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 tumours characterized by different early aberrations (Höglund et al., 2001, 2002). This could be a reason for considering a wider family of graphical models. However, 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 tumours, where a large variety 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 graphical models if there were evidence that tree models could not explain the dependences 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.
Oncogenetic tree models

ACKNOWLEDGMENTS

We thank Wolfgang Huber and Martin Vingron for helpful discussions and Dirk Buschmann for critical reading of the manuscript. AvH was supported by a grant from the German Ministry of Science (BMBF) within the German Human Genome Project (DHGP).

APPENDIX A

Let $e$ be an edge of the tree $T$ 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 (3.6).

**PROPOSITION A.1** If $n_{e_1} < m_{e_2}$ and $n_{e_2} < m_{e_1}$, the maximum of

$$f_e : [0, 1)^2 \to \mathbb{R}$$

$$(x_1, x_2) \mapsto \frac{(1 - x_1)^{m_{e_1}} x_1^{n_{e_1}} (1 - x_2)^{m_{e_2}} x_2^{n_{e_2}}}{(1 - x_1 x_2)^m_e}$$

(A.1)

(see (3.9)) 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:

$$n_{e_1} / x_1 - m_{e_1} / (1 - x_1) + m_e x_2 / (1 - x_1 x_2) = 0$$

$$n_{e_2} / x_2 - m_{e_2} / (1 - x_2) + m_e x_1 / (1 - x_1 x_2) = 0.$$  

(A.2)

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 $n_{e_2} = 0$).

REFERENCES


[Received 27 June 2003; revised 23 December 2003; accepted for publication 17 February 2004]