Construction of Evolutionary Tree Models for Renal Cell Carcinoma from Comparative Genomic Hybridization Data

Feng Jiang, Richard Desper, Christos H. Papadimitriou, Alejandro A. Schäffer, Olli-P. Kallioniemi, Jan Richter, Peter Schraml, Guido Sauter, Michael J. Mihatsch, and Holger Moch

Institute for Pathology, University of Basel, CH-4004 Basel, Switzerland [F. J., R. D., P. S., G. S., M. J. M., H. M.]; Deutsches Krebsforschungszentrum, Abt. Theoretische Bioinformatik, D-69120 Heidelberg, Germany [R. D.]; Department of Electrical Engineering and Computer Science, University of California at Berkeley, Berkeley, California 94720 [C. H. P.]; Computational Biology Branch, National Center for Biotechnology Information, NIH, Bethesda, Maryland 20894 [A. A. S.], and Cancer Genetics Branch, National Human Genome Research Institute, NIH, Bethesda, Maryland 20892 [O. P. K.]

ABSTRACT

Renal cell carcinoma is characterized by an accumulation of complex chromosomal alterations during tumor progression. Chromosome 3p deletions are known to occur early in the carcinogenesis, but the nature of subsequent events, their interrelationships, and their sequence is poorly understood, as one usually only obtains a single “view” of the dynamic process of tumor development in a particular cancer patient. To address this limitation, we used comparative genomic hybridization analysis in combination with a distance-based and a branching-tree method to search for tree models of the oncogenesis process of 116 conventional (clear cell) renal carcinomas. This provides a means to analyze and model cancer development processes based on a more dynamic model, including the presence of multiple pathways, as compared with the fixed linear model first proposed by Vogelstein et al. (N. Engl. J. Med., 319: 525–532, 1988) for colorectal cancer. The most common DNA losses involved 3p (61%), 4q (50%), 6q (40%), 9p (35%), 13q (37%), and Xq (21%). The most common gains were seen at chromosome 17p and 17q (20%). The tree model derived from the distance-based method is consistent with the established theory that 3p is an important early event in conventional (clear cell) renal cancer and supports the prediction made from the branching tree that 4q is another important early event. Both tree models suggest that there may be two groups of clear cell renal cancers: one characterized by 4q−, 7p+, and 1q+, and another by 9p−, 13q−, and 18q−. Putative prognostic parameters were −9p and −13q. The distance-based tree clarifies that −8p (present in 12% of tumors) is a late event, largely independent of other events. In summary, tree modeling of comparative genomic hybridization data provided new information on the interrelationships of genetic changes in renal cancer and their possible order, as well as a clustering of these events. Using tree analysis, one can derive a more in-depth understanding of the renal cancer development process than is possible by simply focusing on the frequencies of genetic events in a given cancer type.

INTRODUCTION

Initiation and progression of cancer is due to genetic alterations. Recent studies characterizing genetic aberrations in RCC have implicated a number of chromosomal loci as important in cancer development (1–3). It has been suggested that an accumulation of genetic events is responsible for tumor progression in RCC, although the details of genetic changes and their order of occurrence in renal tumorigenesis are not well understood (3, 4). The study of colorectal cancer by Vogelstein et al. (5) suggested that the progression of colorectal cancer can be described by a chain of four genetic events: (a) mutation or deletion of the adenopolyposis coli gene on chromosome 5q; (b) mutation of the K-ras gene on chromosome 12p; (c) deletion of the deleted in colon cancer gene on chromosome 18q; and (d) mutation or deletion of the p53 gene on chromosome 17p. When the first of these events occurs, the chance of the second event occurring increases, and when the second event occurs the chances of the third event increases, and so on. These events are irreversible, in that once an event occurs, it is never reversed in the future. In mathematical terms, this provides a path model for tumor progression. This model suggests that the cell starts as normal and proceeds along a path with four vertices representing different genetic changes. The presence of all four events appears to be an indicator of colorectal cancer because adenomas do not show all four alterations (5). Whereas the path model suggests a most likely order of occurrence, colorectal cancer is really associated with the accumulation of the genetic changes on the path. In fact, the genetic changes will not always occur in the order of the path, but the path defines a preferred order.

Models for tumor progression pathways would be of obvious value to define gene loci relevant for the early diagnosis or treatment of cancer. However, attempts to find similar path models for other types of cancer have not been successful. The main reason for this is that cancer is genetically heterogeneous, even in tumors that are considered to be clinically or pathologically homogeneous by all current tests (6). Previous molecular studies of RCC have demonstrated a high degree of genetic heterogeneity (7–9). Therefore, it is important to explore models that are not dependent on a single linear progression. The natural way of generalizing path models to capture heterogeneity leads to tree models (Fig. 1). Previously, two classes of tree models for oncogenesis, called branching and distance-based trees, have been proposed by Desper et al. (10, 11). In these models, it is believed that genetic events do not occur in a random fashion. Once an event occurs, it increases the probability of other events occurring. However, the connection between one event and the next will be specific and directly causal in some instances, whereas in other cases, the later events occur seemingly at random because of the basic genetic instability in a tumor cell.

Many locus-specific molecular analyses have been applied in the past to study the frequency and biological relevance of genetic alterations in RCC. A comparison between these studies is difficult because different molecular markers were used, and because some analyses were performed on tumor-derived cell lines or did not use the most recent histological classification of RCCs. The frequencies of some chromosomal alterations obtained from selected FISH, CGH, and LOH studies is given in Table 1. This summary shows the broad frequency range obtained from these analyses for some chromosomal loci.

The technique of CGH allows a genome-wide survey of all significant DNA sequence copy number gains and losses per tumor in a single experiment (12). CGH can detect DNA sequence copy number changes if the affected region spans more than 10 Mb (12–14).

Received 1/27/00; accepted 9/12/00.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

1 Supported by the Swiss National Science Foundation (97.3100-050752), Krebsforschung Schweiz (KFS 367-9-1996), Basler Krebsliga (3979) and the National Science Foundation Grant CCR-9626361 (to C. H. P.)

2 To whom requests for reprints should be addressed, at Institute for Pathology, University of Basel, Schneonbeinstrasse 40, CH-4003 Basel, Switzerland. Phone: 061-265 28 90; Fax: 061-265 31 94; E-mail: hmoch@uhbs.ch.

3 RCC, renal cell carcinoma; LOH, loss of heterozygosity; FISH, fluorescence in situ hybridization; CGH, comparative genomic hybridization; CRC, conventional (clear cell) renal carcinoma; CNA, copy number aberration.

6503
MATERIALS AND METHODS

Patients

One hundred and sixteen consecutive CRCCs were selected from the archives of the Institute for Pathology, University of Basel. Histological grading and tumor staging were done according to Fuhrman et al. (18) and the International Union Against Cancer (UICC) (19). There were 5 grade 1, 39 grade 2, 56 grade 3, and 16 grade 4 tumors. Forty tumors were stage pT1, 10 were pT2, 65 were pT3, and 1 was pT4. Tumor size ranged from 2 to 16 cm (median, 6.5 cm) in largest diameter. Survival data were obtained by reviewing the hospital records, by direct communication with the attending physicians, and from the Cancer Registry of Basel. Patients were evaluated from the time of biopsy diagnosis to the last known follow-up.

Comparative Genomic Hybridization

DNA Preparation. Tumor DNA was extracted from formalin-fixed and paraffin-embedded primary tumors. Specimens were trimmed to ensure a minimum of 75% tumor cells in the sample. Tissue preparation and DNA extraction was as described (9, 20, 21). One µg of tumor DNA was nick-translated by using a commercial kit (BioNick Kit; Life Technologies, Gaithersburg, MD) and Spectrum Green- dUTP’s (Vysis, Inc., Downers Grove, IL) for direct labeling of tumor DNA. Spectrum Red-labeled tumor reference DNA (Vysis) was used for cobyridization.

CGH and Digital Image Analysis. The hybridization mixture consisted of 200 ng Spectrum Green-labeled tumor DNA, 200 ng Spectrum Red-labeled normal reference DNA, and 20 µg of Cot-1 DNA (Life Technologies) dissolved in 10 µl hybridization buffer [50% formamide, 10% dextran sulfate, 2 × SSC (pH 7.0)]. Hybridization, image acquisition, image analysis, and control experiments were exactly as described (20, 21). At least four observations/autosome and two observations/sex chromosome were included in each analysis. Each CGH experiment included a tumor cell line (Spectrum Green MPE-600, Vysis) with known aberrations (positive control) and a hybridization of two differentially labeled sex-mismatched normal DNAs to each other (negative control). Sex-mismatched normal controls were also used to test the ability of each metaphase batch to allow for a linear relationship between fluorescence intensities and DNA sequence copy numbers. Metaphases were used only if the color ratio of sex-mismatched normal DNAs was ≤0.66 at the X chromosome (14).

The thresholds used for definition of DNA sequence copy number gains and losses were based on results of CGH analyses of formalin-fixed normal tissues (14, 20, 22, 23). A gain of DNA sequences was recorded at chromosomal regions where the hybridization resulted in a tumor:normal ratio >1.20. Overrepresentations were considered amplifications when the fluorescence ratio values exceeded 1.5 in a subregion of a chromosome arm. A loss of DNA sequences was recorded when the tumor:normal ratio was <0.80. To define an aberration, additionally it was required that the first SD was above (gain) or below (deletion) 1.00. Because in negative control hybridizations the mean green:red ratio occasionally exceeded the fixed 1.2 cutoff in normal tissues at 1p, 16p, 19, and 22, these guanine-cytosine-rich regions, known to produce false positive results by CGH, were excluded from all analyses (14). CGH profiles were only evaluated if the quality of the CGH experiment was appropriate (14).

Statistics

Contingency table analysis was used to analyze the relationship between genomic alterations, grade, stage, and cell type. A U test was applied to compare the number of genomic alterations between different grades and stages using StatView 5.0 PPC (SAS Institute, Cary, NC). Ps < 0.05 were considered significant. The most common individual aberrations (occurring in >20% of tumors) were analyzed for their association with patient survival using the Kaplan-Meier method. Statistical differences between the groups were determined with the log-rank test (24). A Cox proportional hazards analysis was used to test for independent prognostic information (25).

Oncogenic Trees

Selecting Events. The result of a CGH test is a set of genetic events (CNAs). The input for the tree modeling procedure is a set of lists of chromosomal gains and losses, one list per tumor. Each gain or loss is recorded at the level of the chromosome arm because we had difficulty deciding whether events labeled by more precise band boundaries correspond to the same event or to different events. We excluded arms 1p, 16p, 19p, and 22q because these guanine-cytosine-rich regions are known to produce false positives in CGH analysis (14), and we ignored the Y chromosome because it cannot be reliably analyzed by CGH. This left 36 chromosome arms to analyze, and there
may have been a gain and/or a loss on each arm, for a total of 72 possible events. Each of the 116 tumors was associated with a list, possibly empty, of the 72 possible events.

It is neither practical nor useful to make a model of all 72 possible events. Most of these events occur only in a few tumors and are probably truly random because of the inherent instability of tumor cells or because of false positives attributable to imperfections in the CGH analysis. Therefore, it is preferable to choose a small set of events that occur sufficiently frequently to appear nonrandom. We implemented the well-known method of Brodeur et al. (26) to select a set of events that appear to occur nonrandomly in this data set of 116 RCC tumors. The method of Brodeur et al. uses random simulation to derive a distribution of event lists for tumors under the null hypothesis that all events occur randomly. To generate a sample from the null distribution, one needs a prior probability for each event to occur. We derived a prior distribution by assuming that gains and losses are equally likely, and the probability of a gain/loss on an arm is proportional to the length of that arm; we used one of the tables of chromosome arm lengths provided in Ref. 27. We generated 10,000 replicates of sets of 116 tumors sampled from the null distribution. For each replicate, a score is computed relating each event to its prior probability, and the maximum score, which represents a measure of how often the most frequent event occurs in random data, is recorded. Scores are also computed for the 72 events in the real data, and an event is considered nonrandom if its score in the real data exceeds the 95th percentile of the 10,000 maximum scores from the random replicates. The method of Brodeur et al. (26) selected 12 events (3p, 4p, 4q, 6q, 6p, 9p, 13q, 18q, Xp, 17p, 17q, and Xp) as nonrandom events in the set of 116 RCC tumors.

Branching Trees. Recently, we proposed a method for reconstructing a tree model of oncogenesis (10) based on a computer science technique called a maximum weight branching (28). The tree models derived by this method are therefore called “branching trees.” In a branching tree, there is one node called a root, and every other node is one of the events. An edge i → j represents a hypothetical cause-and-effect relationship meaning that the occurrence of event i makes the occurrence of j more likely. Branching trees are natural generalizations of Vogelstein’s path model (5) because there can be multiple edges coming out of each node that intuitively represent the heterogeneous possibilities for how oncogenesis can progress. The choice of which edges to include in the branching tree is based on a weight function that takes into account how often each event occurs and how often each pair of events occurs together in the same tumor. If there are n events and 1 root, the tree will always have n edges. We choose the tree for which the sum of the weights on the n edges is maximized. Although the number of possible trees grows exponentially as a function of n, it is possible to choose the maximum weight tree quickly using an algorithm developed in the field of combinatorial optimization (28–31).

The central theoretical result of Desper et al. (10) is that under plausible assumptions about the stochastic process of oncogenesis, and given enough samples, the maximum weight branching method provably converges to the correct tree model. Although in practice we generally do not have enough tumor samples for the convergence result to apply, we still would expect that the maximum weight branching tree has most of the edges correct. The technical assumptions needed to prove that the convergence results were inspired by some work on similar phylogenetic trees, now called Cavender-Farris trees because of the seminal papers (32, 33).

Distance-based Trees. A different class of tree models that uses existing phylogenetic tools more directly was described previously by Desper et al. (11). The trees constructed by this approach have all of the events at leaves, whereas the internal nodes are hidden, unnamed events, much like a phylogenetic tree has the existing species as leaves, and the hypothetical common ancestors as internal nodes (Fig. 1). Another significant difference is that, in these trees, each edge has a length, and we draw the trees so that horizontal distance along an edge is proportional to length. The trees are constructed by first defining an n × n distance matrix that describes for each pair of events whether they tend to occur together or not. The second step is to use existing phylogenetic methods to find the phylogenetic tree that best fits the distance matrix. For this reason the trees constructed by this method are called “distance-based trees.” The problem of fitting a tree to a matrix of distances has been extensively studied in phylogenetics (34–37). The distance between two events i and j is the sum of the distances on the edges between them in the tree.

Typically, distance-based methods produce unrooted trees, or root the resulting phylogeny by an outgroup. We added an extra leaf, to serve as the root of the tree. In our model, each edge e = (i,j) had an associated probability p_e, representing the conditional probability that event j occurs, given that event i occurs. To use a distance method, there must be an underlying additive metric. As probabilities are multiplicative in this model when directed away from the root, we used the logarithm to transform the edge probabilities to an additive metric, such that d_e = −log p_e for each edge e.

Under this transformation, the distance from the root to any other event i is −log p_i, where p_i is the probability of observing event i. Letting p_i be the probability of observing both events i and j, the distance from i to j is defined as −2 log p_i + log p_i + log p_j.

Although distance-based trees are not a direct generalization of the Vogelstein path model, they may provide a statistically robust answer to the basic questions:

(a) Which events are early? Those near the root. The distance to the root is the negative logarithm of the probability, so a shorter distance means a higher probability, i.e., an earlier event.

(b) Which events mark subclasses of tumors? Those which cluster together in subtrees.

Previous advantages and disadvantages of the branching trees and distance-based trees are described at length in the “Discussion” section. In practice, we find it useful to construct both trees and to compare the predictions they make about oncogenesis.

---

**Table 1** Frequencies of genetic alterations in conventional (clear cell or non-papillary) RCCs obtained from selected previous FISH, LOH and CGH analyses

<table>
<thead>
<tr>
<th>Author</th>
<th>Method</th>
<th>Tumor no.</th>
<th>3p−</th>
<th>7+</th>
<th>8p−</th>
<th>9p−</th>
<th>13q−</th>
<th>14q−</th>
<th>17p−</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wu</td>
<td>FISH</td>
<td>30</td>
<td>96</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>37</td>
</tr>
<tr>
<td>el-Naggar</td>
<td>FISH</td>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Siebert</td>
<td>FISH</td>
<td>26</td>
<td>79</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Renshaw</td>
<td>FISH</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>88</td>
</tr>
<tr>
<td>Moch</td>
<td>FISH</td>
<td>41</td>
<td>69</td>
<td></td>
<td></td>
<td></td>
<td>18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schullerus</td>
<td>LOH</td>
<td>105</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>33</td>
<td>33</td>
<td>45</td>
</tr>
<tr>
<td>Herberse</td>
<td>LOH</td>
<td>88</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>48</td>
</tr>
<tr>
<td>Wilhelm</td>
<td>LOH</td>
<td>44</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thrash-Bingham</td>
<td>LOH</td>
<td>22</td>
<td>41</td>
<td></td>
<td></td>
<td></td>
<td>18</td>
<td>16</td>
<td>4</td>
</tr>
<tr>
<td>Premi</td>
<td>LOH</td>
<td>60</td>
<td>75</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>26</td>
</tr>
<tr>
<td>Cairns</td>
<td>LOH</td>
<td>42</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lai</td>
<td>LOH</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>57</td>
</tr>
<tr>
<td>Brooks</td>
<td>LOH</td>
<td>42</td>
<td>47</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>Anglade</td>
<td>LOH</td>
<td>60</td>
<td>88</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>33</td>
</tr>
<tr>
<td>Mock</td>
<td>CGH</td>
<td>16</td>
<td>41</td>
<td>56</td>
<td>15</td>
<td>20</td>
<td>24</td>
<td>24</td>
<td>12</td>
</tr>
<tr>
<td>Gronwald</td>
<td>CGH</td>
<td>25</td>
<td>60</td>
<td>24</td>
<td>20</td>
<td>12</td>
<td>4</td>
<td>28</td>
<td>0</td>
</tr>
</tbody>
</table>

* 1, 3, 7, 44, 49, 52, 55–64.
* a Results obtained from tumor-derived cell lines.
* b Only PT3, RCC without metastases analyzed.
* c RCC and metastases analyzed (only data of primary tumors given).
The number of aberrations was not associated with histological grade or pT stage. To screen for specific chromosomal alterations that might be linked to tumor progression, the association between the most frequent aberrations, histological grade, and pT stage was analyzed. Chromosome 6q loss was detected in 14 of 48 pT1/2 tumors (29%) and in 31 of 67 pT3/4 RCCs (46%), but this trend did not reach significance (P = 0.06). Interestingly, gains of chromosomes 5q, 17p, and 17q were associated with good tumor differentiation. Chromosome 5q gains were detected in 10 of 46 grade 1/2 tumors (22%), and in 6 of 70 grade 3/4 tumors (9%; P = 0.05). Chromosome 17p gains were seen in 67% of grade 1/2 tumors, but in only 38% of grade 3/4 tumors (P = 0.008). Gains of chromosome 17q were detected in 49% of grade 1/2 tumors and in 29% of grade 3/4 tumors (P = 0.03). Chromosome 9p, 13q, and 18q losses were not associated with tumor grade or stage (contingency table analysis).

Construction of Oncogenetic Trees for Renal Carcinoma Progression

Using the CGH data of 116 CRCC, a branching tree and a distance-based tree were constructed. The branching tree shown in Fig. 2 is based on a maximum branching in a weighted graph. A branching tree has a root vertex representing a normal cell, whereas the other vertices represent observed CNAs of interest. An edge from CNA i to CNA j indicates that the occurrence of i increases the probability for the occurrence of j in branching trees. The branching tree is generated purely from considering individual probabilities and pairwise joint probabilities. A strongly correlated pair is usually presented in the tree with an edge from one to the other. The placement of leaves adjacent to the root, such as −3p and −8p, indicates that these two events are not more strongly correlated with some events than with other events. It leaves unclear whether −3p and −8p are important early events or not. In contrast, −4q, which is next to the root and the root of a large subtree, is clearly predicted to be an important early event. The edge lengths in the branching tree are irrelevant.

In the distance-based tree (Fig. 3), the CNAs of interest are all leaves of the tree, whereas the internal vertices are hypothetical hidden events. The observed lists of CNAs were used to define the distances between each pair of CNAs, with the CNAs on the leaves and lengths on the edges, that preserves the computed distances as closely as possible. Therefore, centrally placed leaves imply an important aberration, and leaves placed close to the root are predicted to be early events. The length associated with an edge is proportional to the horizontal length of that edge on the page; the vertical segments are used simply to separate the leaves for clarity. The distance-based model is consistent with the established theory that a loss on 3p is an early important event for CRCC, and suggests that it is not causatively associated with specific other gains or losses. The subtrees predict that there may be at least two subclasses of RCC: one subclass marked by the events −6q, +17q, and +17p, and the other by the event −9p, −13q, and −18q. The group with −6q, +17q, and +17p showed no association with tumor grade (P = 0.5) or stage (P = 0.8). Also, the group with −9p, −13q, and −18q showed no relationship to grade (P = 0.6) or stage (P = 0.2).

Clinical Outcome

Tumor-specific survival was available for 81 patients, with a mean follow-up of 49 ± 33 months (range 0.5 to 131 months). There were 29 pT1/2 tumors and 52 pT3/4 tumors. The tumor-specific survival in patients with pT3/4 tumors was worse compared with patients with pT1/2 tumors (P < 0.01). Tumors with/without −6q, +17q, and/or +17p, and tumors with/without −9p, −13q, and/or −18q were tested to see if the CRCC subclasses predicted by the mathematical models were linked to patient prognosis within the group of pT3/4 tumors. There were too few tumor-related deaths in the group of pT1/2 tumors for a statistical analysis within this group. Within pT3/4 CRCCs, tumors with −9p, −13q, or −18q had a worse prognosis than tumors without these lesions, although this trend did not reach significance (P = 0.06). The subclass with −6q, +17q, and/or +17p was not associated with prognosis (P = 0.5). Once an association was seen between survival and tumors with a group of CGH aberrations, individual loci were tested to see which were significantly linked to patient outcome. Chromosome 13q deletion was the only individual locus significantly associated with clinical outcome within pT3/4 tumors. The tumor-specific survival of patients with 13q loss was
worse compared with patients having tumors without 13q loss in univariate analysis \( (P = 0.03; \text{log-rank test}) \). There was also a trend for an association of chromosome 9p losses with worse prognosis in pT3 RCC, but this trend did not reach significance \( (P = 0.09) \). Not associated with patient prognosis were 18q losses \( (P = 0.9) \). Cox proportional hazards analysis with the prognostic parameters tumor stage, and histological grade indicated that neither the combination of \(-9p, -13q, \text{ and } -18q\) as a group, nor the single changes, were independent predictors of prognosis.

**DISCUSSION**

Previous molecular and cytogenetic analyses of CRCC showed that loss of 3p is an early event in tumor initiation \((39, 40)\), but the relevance of other alterations for tumor progression is unclear. Most of the previous analyses of CGH data have simply counted the frequencies of different events and noticed that some events occurred much more frequently than would be expected at random \((16)\). Few of these studies have attempted to analyze the interrelationships between the genetic alterations, and often the number of cases analyzed has not been large enough to make this possible. The collection of this large set of CGH data on CRCC provided the opportunity to analyze co-occurrences of multiple events in looking for a more complex and comprehensive model of the genetic changes that mark CRCC development.

Mathematical models for tumor progression are of interest to characterize pathways whereby oncogenesis proceeds. For example, the path model is a well-established hypothesis for tumor progression of colorectal cancer \((5)\). However, analysis of the complex CGH data suggests that a path model is not useful for other types of cancer \((41)\). It has been assumed that the genetic events take place in a more tree-like than path-like pattern \((10)\). The analogy between evolving tumors and evolving species had been observed by Buetow et al. \((42)\), who used phylogenetic methods to genetically classify a set of liver tumor samples. We recently developed a tree model to analyze genetic progression of solid tumors, and showed that an algorithm based on maximum-weight branching in a graph correctly infers the topology of the tree. This group of tree models has been called branching trees \((10)\). Using tree-fitting algorithms developed by researchers in phylogeny, we constructed a different type of tree model called distance-based trees, in which events are leaves of the tree \((11)\).

Remarkably, both tree models shared many relevant properties and allowed us to derive new hypotheses for CRCC progression: \( (a) \) there are important subevents emanating from a common root; \( (b) \) a subtree with \(-13q, -9p, \text{ and } -18q\); \( (c) \) a subtree with \(-6q, +17q, +17p\); \( (d) \) a close relationship between \(-4q \text{ and } -4p\); and \( (e) \) \(-3p \text{ and } -4q\) are important early events near the root.

The subevents in our models predict that there may be at least two subclasses of CRCC loosely associated with \(-4q\): one subclass in which \(+17q\) is tightly linked with and precedes \(+17p\) in a side branch closely related to \(-6q\); and another subclass that is characterized by the events \(-9p, -13q, \text{ and } -18q\).

The event \(-4q\) is closer to the first subclass, but may represent a third subclass because it appears in its own subtree, or it may be closer to independent as indicated by the long branch leading to it. Losses of \(4q\) were observed in 5–8% of RCC by LOH studies \((43, 44)\) but in 10–50% by CGHs \((3, 16, 21, 45)\). Interestingly, losses on \(4q\) were detected in 46–77% of colon, cervix, and hepatocellular carcinoma, suggesting the presence of tumor suppressor genes on this chromosomal arm \((41, 46, 47)\).

Interestingly, this CGH analysis revealed a significant association between 13q losses and poor prognosis. Previous studies examining LOH have shown that 13q allelic imbalances can occur in CRCC. The rates of 13q LOH ranged from 4 to 57% \((3, 45, 48–52)\). Our finding that chromosome 13q loss is associated with poor patient outcome is consistent with a previous study of Lai et al. \((52)\). This group found that LOH of the retinoblastoma gene, located on chromosome 13q14, was detectable in 4 of 7 informative RCCs (57%) and was associated with high tumor grade, high stage, and poor prognosis. This raises the possibilities that inactivation of one or more gene(s) on 13q might be relevant for CRCC progression.

In this CGH analysis of 116 CRCC, we confirmed previous studies suggesting that 9p losses are related to short recurrence-free survival in pT3 CRCC \((3, 53)\). In this study, 9p losses were only weakly related to poor tumor-specific survival in pT3 CRCCs by CGH. Although this trend did not reach significance, a microsatellite analysis from our laboratory demonstrated a clear relationship between LOH on 9p and poor prognosis in papillary RCC \((54)\) and in CRCC.\(^4\) According to our preliminary mapping analyses, a gene on 9p13-14 might be relevant for both types of RCC.

Losses of the short arm of chromosome 8 have been recently considered as important events in RCC progression because 8p losses were associated with higher tumor grade in nonpapillary CRCC \((55)\). LOH on 8p was detected in 18–33% of RCC by microsatellite analysis \((44, 55)\). The extremely long branch to \(-8p\) in the distance-based tree suggests that this event is more likely to be a late effect than an early cause. This finding also corresponds to previous biological observations. We have recently studied CRCC metastases and their primary tumors by CGH \((9)\). Loss of 8p was a frequent finding in metastases but was quite rare in corresponding primary tumors. This finding is consistent with the observation derived from the distance-based tree model that 8p loss is a late event in primary CRCC. If one assumes that 8p losses are late events according to the present tree analysis, and our previous study found them associated with metastases, it is tempting to speculate that the presence of 8p loss in the primary tumors would be predictive of metastatic progression. However, cell clones with 8p loss may represent minor cell populations in the primary tumor and are therefore not detectable by CGH because CGH allows only detection of clonal copy number aberrations, which are present in >50–75% of the tumor cells \((13, 14)\). Therefore, one would need other techniques (such as FISH) to detect these individual, possibly metastasis-prone cells.

Both tree models have important advantages and disadvantages. The distance-based method considers all pairwise correlations simultaneously \((11)\). If an event is very weakly correlated with a large number of events, it will be pulled to the subtree containing those events. Such a phenomenon explains the fact that \(-3p\) is pulled toward the center of the distance-based tree. The relatively long length of the edge from the root to the first central node suggests a general phenomenon that the occurrence of any one event makes any other event more likely. Comparing how we use the branching and distance-based tree models to make predictions illustrates some of the strengths of the distance-based method. For both types of trees, early events should be close to the root. The distance-based approach quantifies this precisely, and allows us to infer an order of events by ranking according to distance to the root.

In the branching tree approach \((10)\), distances are measured less precisely by numbers of edges. A CNA is adjacent to the root in the branching, but has no subbranches (such as \(-8p\) in the renal cancer branching tree). When a CNA is a branch adjacent to the root in the branching tree (such as \(-3p \text{ and } -8p\)) it is unclear whether this

---

is because it is an early event that precedes the others but is not correlated especially well with any of them, or whether it is a relatively late event that is also not well correlated with a subclass of tumors. In this case, the distance-based tree clarifies that −3p is of the first type, and −8p is of the second type.

Both methods tend to cluster CNAs that occur together in subtrees, but the distance-based method is more robust in that it preserves information about the co-occurrence of all pairs of CNAs, whereas the branching tree only shows the best-correlated pairs. One advantage of the branching tree is that it defines edges between CNAs leading to direct predictions of cause-and-effect relationships. In practice, we constructed both types of trees. We looked for similarities in which events were near the root to predict the early CNAs. We looked for similar clustering in subtrees to predict which events tend to occur together and may mark genetically homogeneous subsets. The two tree construction methods are encoded in free software.5 The branching tree method is entirely contained in the software we distribute. For the distance-based method, our software goes as far as producing a distance matrix suitable for input to PHYLP (38).

Since the publication in 1988 by Vogelstein et al. (5) of the path model for colorectal cancer, it has been hoped that similar models could be found for other types of cancer; but this search has been largely fruitless. Because the CGH technique allowed a genome-wide view of genomic alterations in solid tumors, it has been hoped that CGH will clarify which genetic changes are nonrandom; this has had mixed results (16). In both cases it appears that a central problem is that many types of solid tumors, such as CRCC, are very heterogeneous and appear to have complex patterns of genetic changes. Furthermore, it cannot be excluded that small deletions <10 Mb are missed by CGH (12). However, mathematical modeling based on the primary CGH data are desirable to propose models more general than path models, to elucidate which genetic changes are most worthy of additional study, and to suggest some good hypotheses for additional investigation. To get useful results from mathematical modeling, it is necessary to collect large tumor data sets such as the one described herein, so that small numbers of truly random genetic events do not obscure the basic patterns.

In summary, we proposed in this study tree models for CRCC oncogenesis that lead to very similar predictions, although they are a relatively late event that is also not well correlated with a subclass of tumors. In this case, the distance-based tree clarifies that −3p is of the first type, and −8p is of the second type.

Both methods tend to cluster CNAs that occur together in subtrees, but the distance-based method is more robust in that it preserves information about the co-occurrence of all pairs of CNAs, whereas the branching tree only shows the best-correlated pairs. One advantage of the branching tree is that it defines edges between CNAs leading to direct predictions of cause-and-effect relationships. In practice, we constructed both types of trees. We looked for similarities in which events were near the root to predict the early CNAs. We looked for similar clustering in subtrees to predict which events tend to occur together and may mark genetically homogeneous subsets. The two tree construction methods are encoded in free software. The branching tree method is entirely contained in the software we distribute. For the distance-based method, our software goes as far as producing a distance matrix suitable for input to PHYLP (38).
Construction of Evolutionary Tree Models for Renal Cell Carcinoma from Comparative Genomic Hybridization Data


Updated version
Access the most recent version of this article at:
http://cancerres.aacrjournals.org/content/60/22/6503

Cited articles
This article cites 57 articles, 20 of which you can access for free at:
http://cancerres.aacrjournals.org/content/60/22/6503.full.html#ref-list-1

Citing articles
This article has been cited by 14 HighWire-hosted articles. Access the articles at:
/content/60/22/6503.full.html#related-urls

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.