JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 7, Number 6, 2000
Mary Ann Liebert, Inc.
Pp. 789–803

Distance-Based Reconstruction of Tree Models for Oncogenesis

RICHARD DESPER,[1] FENG JIANG,[2] OLLI-P. KALLIONIEMI,[3] HOLGER MOCH,[2]
CHRISTOS H. PAPADIMITRIOU,[4] and ALEJANDRO A. SCHÄFFER[5]

ABSTRACT

Comparative genomic hybridization (CGH) is a laboratory method to measure gains and losses in the copy number of chromosomal regions in tumor cells. It is hypothesized that certain DNA gains and losses are related to cancer progression and that the patterns of these changes are relevant to the clinical consequences of the cancer. It is therefore of interest to develop models which predict the occurrence of these events, as well as techniques for learning such models from CGH data. We continue our study of the mathematical foundations for inferring a model of tumor progression from a CGH data set that we started in Desper *et al.* (1999). In that paper, we proposed a class of probabilistic tree models and showed that an algorithm based on maximum-weight branching in a graph correctly infers the topology of the tree, under plausible assumptions. In this paper, we extend that work in the direction of the so-called distance-based trees, in which events are leaves of the tree, in the style of models common in phylogenetics. Then we show how to reconstruct the distance-based trees using tree-fitting algorithms developed by researchers in phylogenetics. The main advantages of the distance-based models are that 1) they represent information about co-occurrences of all pairs of events, instead of just some pairs, 2) they allow quantitative predictions about which events occur early in tumor progression, and 3) they bring into play the extensive methodology and software developed in the context of phylogenetics. We illustrate the distance-based tree method and how it complements the branching tree method, with a CGH data set for renal cancer.

Key words: cancer, algorithms, phylogenetic trees, comparative genomic hybridization.

[1]Deutsches Krebsforschungzentrum, Abt. Theoretische Bioinformatik, Heidelberg, Germany.
[2]Institute of Pathology, University of Basel, Basel, Switzerland.
[3]Cancer Genetics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD.
[4]Department of Electrical Engineering and Computer Science, University of California at Berkeley, Berkeley, CA.
[5]National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD.

1. INTRODUCTION

Recent research in cancer genetics focuses on identifying changes in the DNA of tumor cells that cause the processes of cell division, cell differentiation, or cell death to go out of control. For each type of cancer, it is desirable to identify genetic changes associated with that type of cancer and to understand the relationships between the submicroscopic genetic changes and the tumor phenotype and clinical outcome.

This research program is facing a number of challenges. First, a genetic change may predispose the cell to undergo more changes, so it is necessary to separate cause from effect, and it is desirable to infer the order in which the alterations occurred. However, it is difficult to get repeated samples of the same tumor over time. Therefore, we need methods that can infer progression and causality from a single genome-wide "snapshot" of each tumor. A second difficulty is that, even when a collection of samples is clinically homogeneous, and the tumor cells look similar under the microscope, the tumors may turn out to be genetically heterogeneous in their causes. Finally, collecting samples from tumors, especially in the least common types of cancer, is difficult to carry out in a large scale, and therefore analysis needs to be done on small numbers of samples.

The relationship of chromosomal abnormalities to cancer was first worked out in forms of leukemia where relatively simple exchanges of DNA between two chromosomes (called translocations) were causally associated with cancer (Nowell, 1976). The analysis of solid tumors, such as breast cancer, prostate cancer, and kidney cancer, has proven much more difficult because solid tumor cells appear to have a much larger set of genetic changes. Many of these changes may be random, while others are related causally to each other and to tumor development. Understanding the causal relationships and time ordering of these changes could shed light on their clinical consequences.

Since the genetic changes that are potentially causally related to solid tumors are many and spread all over the genome, laboratory techniques that provide a genome-wide view of genetic alterations are needed. We have applied our methodology to data collected by a technique called *Comparative Genomic Hybridization* (CGH) (Kallioniemi *et al.*, 1992). A normal cell should have two copies of the DNA for each entire chromosome. CGH provides a genome-wide survey of chromosomal regions in which tumor cells and normal/control cells have significantly dissimilar amounts of DNA. Such regional gains or losses of DNA in a tumor cell are called *copy number aberrations* (CNAs). Copy number aberrations are thought to be important indicators of cancer-related genes in the region—a gain could indicate a hyperactive oncogene, while a loss could indicate a tumor suppressor gene whose activity level is pathologically low.

A CGH experiment yields a list of chromosomal regions with gains or losses in the tumor. CGH is carried out by fluorescently labeling equal volumes of a sample of normal DNA and a sample of tumor DNA with two different colors. The two samples are then biochemically matched in a process called *hybridization* and images are taken of the hybridizing DNA. Image analysis is then used to quantify the ratio of the two colors in each region. A range of normal ratios $[1/r < 1 < r]$ is prescribed. Any region outside the normal range is reported as a gain or loss for the tumor, depending on whether the tumor color is more or less prevalent. An exposition on CGH and a survey of its applications can be found in (Forozan *et al.*, 1997).

CNA data represent a set of genetic changes that took place in some unknown order. CGH studies of various types of cancer suggest that the CNA lists are not entirely random, in that various CNAs or combinations of CNAs appear to be causally linked to the cancer and to each other. Using the CNA lists as input, our goal is to use mathematical modeling in order to identify for each cancer type:

1. A set of genetically significant events which appear to be causally related to the particular cancer type, as opposed to events that occur randomly.
2. The order of such events, and especially the early events.
3. Groups of events that tend to occur together, possibly indicating a set of alterations whose co-occurrence greatly enhances tumor development.
4. Any apparent cause-and-effect relationships between CNA events that could be further evaluated in the laboratory.

Our modeling work was inspired by the work of Vogelstein *et al.* (Fearon and Vogelstein, 1990; Vogelstein *et al.*, 1988) on a type of colorectal cancer with visible and distinguishable precancerous stages. Vogelstein *et al.* were able to associate specific genetic changes with four of the stages of cancer progression. The genetic changes are irreversible and the presence of all four changes indicates that the cell is cancerous. In mathematical terms, this provides a *path* model for tumor progression, where we think of the cell starting from a healthy, normal state and proceeding down a path with four vertices representing different genetic changes. (In fact, the genetic changes will not always occur in the order of the path, but the path defines a preferred order.)

Unfortunately, attempts to find similar path models for other types of cancer have not been successful. CGH studies such as (Kuukasjärvi *et al.*, 1997) suggest that this is because many cancers are genetically heterogeneous, in that clinically similar cancers have different genetic causes. Furthermore, many types of cancer do not have clearly distinguishable premalignant steps like those found in the colorectal cancer studied by Vogelstein *et al.* A tree model is the simplest natural generalization of a path model, so we use trees in order to capture this heterogeneity.

In this paper, we further develop the mathematical models presented in Desper *et al.* (1999) and present algorithms for identifying early genetic changes and classes of genetic changes that tend to occur together in a tumor. In Desper *et al.* (1999), we proposed a class of tree models for oncogenesis, called in this paper *branching trees*. A branching tree has a root vertex representing a normal cell, while 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. We proposed a method for reconstructing the branching tree by the maximum-weight branching algorithm (Karp, 1971), and proved that, under plausible assumptions, this algorithm converges to the correct branching tree.

One of the tools used in the convergence proof in Desper *et al.* (1999) was inspired by research on phylogenetic trees. The analogy between evolving tumors and evolving species had been observed by Buetow *et al.* who used phylogenetic methods to genetically classify a set of liver tumor samples (Buetow *et al.*, 1998). In this paper we expand on this analogy by proposing a different class of tree models for oncogenesis that enable us to use phylogenetic tools more directly and extensively. The tree models considered here are called *distance-based trees*, as distinct from the branching trees of (Desper *et al.* 1999). In the distance-based trees, CNAs of interest are all *leaves* of the tree, while the internal vertices are hypothetical hidden events (extinct species in the phylogenetic analogy). The observed lists of CNAs are used to define distances between each pair of CNAs. We seek a tree whose leaves represent the CNAs and lengths for the edges of this tree, such that the leaf-to-leaf distances in the tree approximate the computed distances as closely as possible. The problem of finding a tree that best fits a given set of pairwise distances has been studied extensively in phylogenetics (Swofford and Olsen, 1990; Barthélemy and Guénoche, 1991), thus, we can take advantage of a well-developed body of literature and algorithmic methods. In the same spirit as in Desper *et al.* (1999), we prove that, using one particular method of tree reconstruction, we can find a tree that is provably close to optimal.

In Section 2, we define the branching and distance-based tree models. Section 3 reviews relevant material from the phylogenetics literature, presents the distance-based tree reconstruction method, and includes our main theorem showing that the tree we reconstruct is near-optimal. In Section 4, we illustrate the application of the distance-based method on a relatively large renal cancer data set, and we compare it to a branching tree on the same data set, as per Desper *et al.* (1999). Although the two methods are very different mathematically, the trees share many topological properties and suggest several similar predictions about which CNAs are most relevant to renal cancer. More comparisons of the two types of tree models and comments on future work are presented in the Discussion.

2. TREE MODELS FOR ONCOGENESIS

2.1. The model

The input to our analysis is a set of lists of copy number aberrations (CNAs), one list from each tumor. Each CNA list may include gains and losses from each arm of each human chromosome (we ignore the

Y chromosome because it contains mostly repetitive DNA that cannot be easily analyzed by CGH). The other 23 human chromosomes are denoted 1, 2, . . . , 22, and X. All chromosomes have a long arm, denoted q, and all except 13, 14, 15, 21, 22 have a substantial short arm denoted p, so there are 41 arms. The actual gains and losses may not necessarily span an entire chromosome arm, but we find it difficult to decide when two subarm intervals should be treated as the same or not. Nevertheless, it is possible for a tumor to have both a gain and a loss (on different sub-intervals) of one chromosome arm. Therefore, there is a total of $41 \cdot 2 = 82$ possible CNAs.

For our analysis, we focus on a much smaller subset, because most of the 82 possible CNAs either do not appear at all or appear very few times in a manner that seems to be random noise. In Desper *et al.* (1999) we used a clique heuristic for selecting the relevant CNAs; here we use an established statistical heuristic of Brodeur *et al.* (1982) to select a set of apparently nonrandom CNAs to model.

Let L be this reduced set of CNAs. Each list of CNAs is thus a subset of L. In fact, we can consider the given set of lists as samples from a probabilistic distribution over 2^L, the set of all subsets of L. We use the term *model of oncogenesis* to refer to a random process that generates subsets of L and therefore defines a probability distribution over 2^L. In this section, we briefly review the tree-based models of oncogenesis proposed in Desper *et al.* (1999) and define the extension studied in the present paper.

A *probability distribution on* 2^L is a function p defined on the subsets of L such that $p(X) \geq 0$ for all X, and $\sum_{S \subseteq L} p[S] = 1$.

If p is such a distribution, then for any $x, y \in L$ we define the probability of an event

$$p_x = \sum_{X \subset L, x \in X} p(X),$$

and of a pair of events,

$$p_{xy} = \sum_{X \subset L, \{x,y\} \subset X} p(X).$$

We define $p_{rx} = p_x$. We use the notation $p_{x|y}$ to denote the conditional probability $p_{x|y} = p_{xy}/p_y$.

We shall consider trees that define distributions on 2^L. The vertex set V of these trees includes a root vertex r (an extra vertex not in L added to denote the null event, or an initial normal state), the genetic events in L, and possibly other nodes. In Desper *et al.* (1999), we required that there be no other nodes besides r and those in L; i.e., $V = L \cup \{r\}$. In the trees considered here, we require that $L \subset V$ be precisely the set of leaves of the tree.

A *rooted tree* on V is a triple $T = (V, E, r)$, where $r \in V$ is a special vertex called the *root*, and E is a set of pairs of vertices such that

1. for each vertex $v \in V$ there is at most one edge $(u, v) \in E$ with v as its second component,
2. there is no edge (u, r) entering r,
3. there is no *cycle*, that is, no sequence of edges in E of the form $((v_0, v_1), (v_1, v_2), \ldots, (v_{k-1}, v_k), (v_k, v_0))$,
4. for each $v \in V \backslash \{r\}$ there is a sequence of edges $\mathcal{P}_v = ((r, v_1), (v_1, v_2), \ldots, (v_{k-1}, v_k))$, such that $v_k = v$, and $(v_{i-1}, v_i) \in E$ for all i.

An *oncogenetic tree* $T = (V, E, r, p, L)$ is a rooted tree with a positive real number $p(e) \leq 1$ associated with each edge $e \in E$ and with a distinguished nonempty set of nodes $L \subseteq V$.

Each oncogenetic tree $T = (V, E, r, p, L)$ generates a distribution on 2^L, as follows. We select a subset $E' \subset E$ by including each edge e independently with probability $p(e)$. We then define the set of *active* edges E'' to be all those edges (u, v) in E', such that all the edges in the path $\mathcal{P}_u$ are also in E'. We then look at the resulting graph (V, E'') and consider the set of all vertices in L that are reachable from the root. This set $S \subseteq L$ is the outcome of this experiment. Let p_T denote the resulting distribution on 2^L.

Oncogenetic trees are a natural generalization of the path-like models considered previously by Vogelstein *et al.* Intuitively, trees are more expressive models when the cancer under study is heterogeneous in genetic origin, and this appears to be a common characteristic of clinically defined cancers. The probability labels

on edges and the method of mapping a tree to a distribution formally capture the geneticists' intuition that the occurrence of some CNAs may predispose others to occur.

In Fig. 1, we show a sample oncogenetic tree, T, rooted on the left, where L is the set of leaves. The leaves in L correspond to the set of CNAs, $L = \{+1q, -8p, Xq\}$, while the two interior nodes $I1$ and $I2$ intuitively stand for hidden states which cannot be directly observed. Let $e_0 = (\text{Root}, I1), e_1 = (I1, +1q), e_2 = (I1, I2), e_3 = (I2, -8p)$, and $e_4 = (I2, +Xq)$. Table 1 shows the possible sets of active edges, the resulting output leaf sets, and their corresponding probabilities. Output probabilities can be read from Table 1 as the sum of probabilities for all possible sets E'' which yield a given output S. For example, $p_T(\{+1q\}) = .056 + .0784 = .1344$. This example demonstrates how edge probabilities induce a distribution on the set 2^L.

2.2. *The reconstruction problem*

In this paper, as in Desper *et al.* (1999), we are interested in the problem of *reconstructing* oncogenetic trees from samples of a distribution p over 2^L. This approach follows in the tradition of learning theory, where a model is proposed to fit a sampled probability distribution. In other words, our problem is the following:

Input:

- A set L of genetic events.
- k samples from a distribution p over 2^L.

Output: an oncogenetic tree $T = (V, E, r, p, L)$ with $L \subset V$, such that p_T is an approximation of p

What we have not specified above is the relationship between L and the set of vertices in the trees that we consider in the reconstruction problem. In Desper *et al.* (1999) we allow all trees over the vertex set $V = \{r\} \cup L$. That is, we assumed implicitly that there are no other genetic events other than those in L. We call this variation of the reconstruction problem *the branching model.* We also proposed an efficient reconstruction algorithm and proved rigorously that, under realistic assumptions, it indeed converges to the correct model.

In this paper, we consider the class of all trees rooted at r whose set of leaves is precisely L. In other words, we allow arbitrary unknown ("hidden") genetic events to be the internal nodes of the oncogenetic tree. In doing so, however, we implicitly make the assumption that the observed genetic events in L are all leaves of the tree. Thus, no single event is a necessary precursor to any other event (an assumption which is justified by the data.) We call this the *distance-based model*, because in our reconstruction methodology

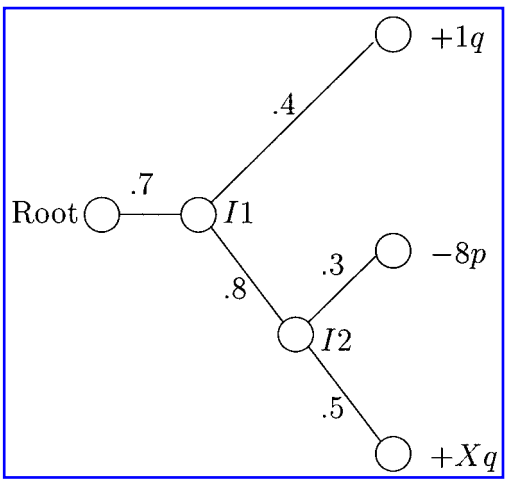

FIG. 1. Sample oncogenetic tree.

TABLE 1. OUTPUT SETS FOR SAMPLE TREE

Active edges E''	Active leaf set S	Probability
$\emptyset$	$\emptyset$	.3
$\{e_0\}$	$\emptyset$	$(.7)(.6)(.2) = .084$
$\{e_0, e_1\}$	$\{+1q\}$	$(.7)(.4)(.2) = .056$
$\{e_0, e_2\}$	$\emptyset$	$(.7)(.6)(.8)(.7)(.5) = .1176$
$\{e_0, e_1, e_2\}$	$\{+1q\}$	$(.7)(.4)(.8)(.7)(.5) = .0784$
$\{e_0, e_2, e_3\}$	$\{-8p\}$	$(.7)(.6)(.8)(.3)(.5) = .0504$
$\{e_0, e_1, e_2, e_3\}$	$\{+1q, -8p\}$	$(.7)(.4)(.8)(.3)(.5) = .0336$
$\{e_0, e_2, e_4\}$	$\{+Xq\}$	$(.7)(.6)(.8)(.7)(.5) = .1176$
$\{e_0, e_1, e_2, e_4\}$	$\{+1q, +Xq\}$	$(.7)(.4)(.8)(.7)(.5) = .0784$
$\{e_0, e_2, e_3, e_4\}$	$\{-8p, +Xq\}$	$(.7)(.6)(.8)(.3)(.5) = .0504$
$\{e_0, e_1, e_2, e_3, e_4\}$	$\{+1q, -8p, +Xq\}$	$(.7)(.4)(.8)(.3)(.5) = .0336$

we use distance-based algorithms from the phylogenetic literature. In the next section, we explain our main results. As with the branching model, there are efficient algorithms for fitting distance-based trees which can be proved, under plausible assumptions, to approximate the correct tree. One of the main attractions of the distance-based model is that it allows us to utilize for oncogenetic trees the extensive methodology (and software) that has developed over decades for the phylogenetic tree reconstruction problem.

Since the reconstruction problem requires learning a probability distribution from samples, no algorithm can guarantee that it reconstructs the "exact" tree, so our task is essentially one of approximation. Our strategy is the following: from the samples of p we first infer a "distance" metric $\mathcal{T}_L$ between the events in L—intuitively, the distance between two events i and j captures the extent to which i and j are correlated. We then find a tree whose path metric is indeed close to the inferred $\mathcal{T}_L$. This approach is inspired by the Cavender-Farris (CF) tree reconstruction algorithm of Farach and Kannan (1996). Therefore, we review some relevant material about CF trees.

2.3. *Cavender-Farris trees*

Molecular biologists and statisticians model evolution as a Markov process proceeding through a tree structure, which represents the evolutionary history of a set of species. Cavender (1978) and Farris (1973) proposed a very simplified model, in which we follow the evolution of a binary value (the presence or absence of a genetic element, say) through the set of species.

A *Cavender-Farris tree*, or *CF tree*, is a weighted, rooted tree $T = (V, E, r, p)$ with an edge probability function $p : E \to (0, .5)$. Associated with such a tree is the following probabilistic experiment. One assigns the bit 0 or 1 to each node of the tree inductively. The bit 1 is assigned at the root, and for each edge $e = (u, v) \in E$, the bit at v equals the bit at u with probability $1 - p_e$ and differs with probability p_e. The outcome is the set of all leaves assigned 1. Denote the resulting probability distribution as p_T^{CF}. An oncogenetic tree can be viewed as a one-sided CF tree, where the bits can flip from 1 to 0 (if the corresponding edge is not selected in E') but a bit cannot flip from 0 back to 1. (Another difference is that oncogenetic trees allow a greater range of edge probabilities.)

Let $T = (V, E, r, p)$ be a CF tree with leaf set $L \subset V$ and associated probability distribution p_T^{CF} on 2^L. The *CF tree reconstruction problem* is this:

Input: k samples of p_T^{CF}.

Output: A tree $T^* = (V^*, E^*, r^*, p^*)$ with leaf set $L \subset V^*$, such that $p_{T^*}^{\mathrm{CF}}$ is an approximation for p_T^{CF}.

Farach and Kannan (1996) proposed measuring the quality of an algorithm's approximation via the L_1 distance between the distribution of the output tree and the distribution of the true tree: $\|T, T^*\|_1 = \sum_{v \in \{0,1\}^L} |p_T^{\mathrm{CF}}(v) - p_{T^*}^{\mathrm{CF}}(v)|$. We shall use the same measure to show convergence for our oncogenetic tree reconstruction algorithm.

2.4. Path metrics

Suppose $T = (V, E, d)$ is an undirected tree with edge weights $d : E \to \mathbf{R}^+$. For each $x, y \in V$, define $\mathcal{P}_{xy}$ to be the unique path in T from x to y. Then d induces a metric on V via $d_T(x, y) = \sum_{e \in \mathcal{P}_{xy}} d(e)$. For $S \subset V$, define $d_{T|S}$ to be the restriction of d_T to S. We call d_T a *path* metric (it is also called an *additive* metric or a *tree* metric in the literature).

Given a finite set L and a metric $d : L^2 \to \mathbf{R}^+$ on it, it is of interest to tell if there is a tree T with L as its set of leaves such that $d_{T|L} = d$. In the phylogenetic application, the vertices of T are all common ancestors of the extant species L, and T then represents the evolutionary history of L; see Barthélemy and Guénoche (1991), Swofford and Olsen (1990), Waterman *et al.* (1977), and references therein. Usually, one does not know d exactly, but rather some approximation d' of d, such that $N(d, d')$ is small for some norm N on the space of path metrics. Recovering d, or finding another tree-like approximation to d', is called the *numerical taxonomy problem:*

Input: $d' : S^2 \to \mathbf{R}^+$, a distance matrix, N a norm on distance matrices.

Output: A weighted tree $T = (V, E, d)$ with $S \subseteq V$ such that $N(d_{T|S}, d')$ is small.

When N is the L_1 or L_2 norm, Day (1987) showed that minimizing the distance $N(d_T, d')$ is NP-complete. Agarwala *et al.* (1999) developed an approximation algorithm when N is the L_∞ norm. Commonly used numerical taxonomy algorithms include neighbor-joining (Saitou and Nei, 1987), the single- and double-pivot method (Cohen and Farach, 1997), and the Fitch-Margoliash (1967) least-squares method.

Many algorithms in phylogenetics work by reducing the CF tree recovery problem to the numerical taxonomy problem, as follows. If $T = (V, E, r, p)$ is a CF tree, we define a path metric by taking $d_e = -\log(1 - 2p_e)$ for each edge e. Then, if u and v are any two leaves of T and we set p_{uv} to the probability that u and v have different bits, we observe the relationship

$$d(u, v) = \sum_{e \in \mathcal{P}_{uv}} d_e = -\log(1 - 2p_{uv}). \tag{1}$$

3. RECONSTRUCTING ONCOGENETIC TREES

3.1. The algorithm

Let T be an unknown oncogenetic tree with leaf set L. We are given a sample D of k subsets of L generated by T. The reconstruction problem asks us to produce an estimate tree T^* with leaf set L such that the probability distribution p_{T^*} approximates the distribution p_T induced by the true oncogenetic tree T.

To use tree-fitting for oncogenetic trees, we need to compute a path metric from the distribution p_T. By analogy with (1), and since probabilities are multiplicative from root to leaf, a logical choice for a path edge weight is the negative logarithm of edge probabilities. Thus, we define a path metric d_T on T by assigning the distance $d(e) = -\log p(e)$ to each edge e. Notice that, if x and y are two leaves of T, their distance can be easily verified to be

$$d_T(x, y) = -2 \log p_{xy} + \log p_x + \log p_y.$$

We can now state our reconstruction algorithm:

1. For each event $x \in L$, let $\hat{p}_x$ be the observed probability of the event x occurring, that is, the number of subsets in D containing x, divided by k.
2. For each pair of events, x and y, let $\hat{p}_{xy}$ be the observed joint probability of x and y, that is, the number of subsets in D containing both x and y, divided by k.
3. For each x, y, let $\hat{d}(x, y) = -2 \log \hat{p}_{xy} + \log \hat{p}_x + \log \hat{p}_y$.
4. Use a tree-fitting algorithm to find a tree T^* whose associated metric $d^* = d_{T^*}$ is close to $\hat{d}$.
5. Return T^*.

Suppose that the obtained samples come indeed from an oncogenetic tree T. As the sample size increases, $\hat{d}$ converges to d_T, the true metric of T. Since the path metric d^* resulting from Step 4 is close to $\hat{d}$, it is also close to d_T. For our theoretical analysis (see the next subsection) we use in Step 4 the pivot method of Agarwala *et al.* (1999), which allows us to formally prove that the tree obtained indeed approximates the correct oncogenetic tree. In practice, our software feeds the distance metric $\hat{d}$ computed in Step 3 into several existing phylogenetics software packages (which have been used for many years and are widely believed to work well in practice) to compute a tree, and we choose the best of these trees as T^*.

3.2. Approximation proof

We measure the distance between two trees by the L_1 distance (Farach and Kannan, 1996; Ambainis *et al.*, 1997) between their corresponding distributions.

Let $p_{\min} > 0$ be the smallest value for p_x among all $x \in L$. Our main result is the following:

Theorem 3.1. *Suppose that the input data are indeed k samples from the distribution p_T of an oncogenetic tree. T. Our oncogenetic tree reconstruction algorithm converges to a tree T^* and distribution p_{T^*} such that the expected L_1 distance between p_T and p_{T^*} is $O(|L|^2/\sqrt{kp_{\min}})$.*

Note that our theorem implies that, when the number of samples becomes large in comparison to $\frac{|L|^4}{p_{\min}}$, the reconstructed oncogenetic tree induces a distribution close to that of the true oncogenetic tree. This is useful as an indication for the number of samples needed for our method to be theoretically conclusive. As we shall see in the next section, in the use of our algorithm L is quite small and $p_{\min}$ not too miniscule, because we choose a subset of common events to work with. Therefore, the guarantees of the theorem become relevant for quite reasonable data. Of course, one hopes that, in practice, good approximations may occur with even smaller samples than guaranteed by our theorem.

To prove the theorem, we will first show that we can achieve a good value for ϵ, which will lead to a bound on the L_∞ distance between the true distance-tree and the output distance-tree. We will then show that the L_∞ bound will translate to a bound in the error on all edges shared by the true tree and the test tree and that any edges not shared by the two trees will be very short. Finally, we will show how an edgewise bound on the length estimates leads to a bound on the variational distance between the corresponding probability distributions. The general direction of the proof parallels the proofs from Farach and Kannan (1996) used for learning Cavender-Farris trees.

Lemma 3.1. *Given k samples from an oncogenetic tree T, for each $x, y \in L(T)$, let $\hat{d}_{xy}$ be the estimate for d_{xy}. Let $\epsilon = \max_{x,y} |d_{xy} - \hat{d}_{xy}|$. Then $E[\epsilon] = O\left(\frac{|L|}{\sqrt{kp_{\min}}}\right)$, where $p_{\min}$ is the minimum probability of any event in L.*

Given k samples from the tree T, for each edge $e = (u, v)$, let $\hat{p}_e$ be the observed probability of e being active, given that u was reachable from the set E'', and let $\delta_e = \hat{p}_e - p_e$. (Here, "observed" probability is something of a misnomer, as the states at u and v may both be hidden from observation.)

Consider a fixed pair of leaves, x, y, and let u be the lease common ancestor of x, y in T. Then $d_{xy} = -\log p_{x|u} - \log p_{y|u}$.

Let $\delta_{x|u} = \hat{p}_{x|u} - p_{x|u}, \delta_{y|u} = \hat{p}_{y|u} - p_{y|u}$. $\delta_x = \hat{p}_x - p_x$, and $\delta_y = \hat{p}_y - p_y$. Then

$$\begin{aligned}
\epsilon_{xy} &= |\hat{d}_{xy} - d_{xy}| \\
&= |-\log(\hat{p}_{x|u}) - \log(\hat{p}_{y|u}) + \log(p_{x|u}) + \log(p_{y|u})| \\
&\leq |-\log(p_{x|u} + \delta_{x|u}) + \log(p_{x|u})| + |-\log(p_{y|u} + \delta_{y|u}) + \log(p_{y|u})| \\
&\leq \sum_{e \in \mathcal{P}_{xy}} |-\log(p_e + \delta_e) + \log(p_e)|. \qquad (2)
\end{aligned}$$

Define $\epsilon = \max_{x,y} \epsilon_{xy}$. From Inequality (2) it follows that

$$\epsilon \leq \sum_{e \in T} |-\log(p_e + \delta_e) + \log(p_e)|.$$

Consider a given term $|-\log(p_e+\delta_e)+\log(p_e)|$. Using the general inequality $|f(t+\delta)-f(t)| \leq |f'(c)|*\delta$ for any differentiable function f, where $t < c < t+\delta$, and the fact that the derivative of $\log t$ is $1/t$, we obtain the inequality

$$|-\log(p_e + \delta_e) + \log(p_e)| = O\left(\frac{|\delta_e|}{p_e}\right).$$

A standard result is that the sample variance with k samples equals the variance divided by k (Sokal and Rohlf, 1995, p. 138). From this it follows that $E[\delta_e^2] = O\left(\frac{p_e}{k}\right)$. By the concavity of the square-root function, $E[|\delta_e|] = O\left(\sqrt{\frac{p_e}{k}}\right)$.

Since $p_{\min}$ is the minimum probability over all events, $p_e \geq p_{\min}$, and thus $E\left[\frac{|\delta_e|}{p_e}\right] = O\left(\sqrt{\frac{1}{kp_{\min}}}\right)$. Summing over the edges in T yields $E[\epsilon] = O\left(\frac{|L|}{\sqrt{kp_{\min}}}\right)$. We have used here the fact that T has no internal node not of degree one; such nodes are redundant, as the two edges adjacent to them can be combined. In such trees, the number of edges is at most twice the number $|L|$ of leaves.

We have established so far that the metric inferred from the data, $\hat{d}$, differs from the correct path metric d_T by $O\left(\frac{1}{\sqrt{kp_{\min}}}\right)$. Thus the tree T^* produces a metric d_{T^*} which observes the same bound. ■

Next, we relate the L_∞ distance from T to a bound on internal edge length errors. Given two trees, T, T', on the same leaf set L, we say the edges $e \in E(T)$ and $e' \in E(T')$ *correspond* if they determine the same partition of the leaf set, i.e., if there is a bipartition $L = L_1 \cup L_2$ such that, for every $u \in L_1$ and every $v \in L_2$, $e \in \mathcal{P}_{uv}$ in T and $e' \in \mathcal{P}_{uv}$ in T'.

Lemma 3.2. *Suppose T, T' are two trees on the same leaf set L such that the weight functions w, w' induce the path metrics D_T and $D_{T'}$, respectively. If $\|D_T - D_{T'}\|_1 \leq \delta$, then for any pair of corresponding edges, $e \in E(T)$ and $e' \in E(T')$, we observe the inequality $|w(e) - w'(e')| \leq 2\delta$. Also, if e is an edge in T which corresponds to no edge $e' \in E(T')$, then $w(e) \leq 2\delta$.*

This lemma is a generalized version of Lemmas 6 and 7 in (Farach and Kannan, 1996).

Proof of Lemma 3.2. If the edges e, e' are corresponding edges in T, T', then there is a bipartition of $L = L_1 \cup L_2$ and leaves $a, b \in L_1, c, d \in L_2$ such that

$$\begin{aligned} w(e) - w(e') = (1/2)[&(D_T(a,d) + D_T(b,c) - D_T(a,b) - D_T(c,d)) \\ &- (D_{T'}(a,d) + D_{T'}(b,c) - D_{T'}(a,b) - D_{T'}(c,d))]. \end{aligned}$$

We observe the bound $|w(e) - w'(e')| \leq 2\delta$ by comparing the like terms on the right-hand side with the L_∞ bound. Now consider an edge $e \in E(T) \backslash E(T')$. Let $L_1 \cup L_2$ be the partition of L induced by e.

There are two possibilities. It may be possible to add an edge e' to T' which separates L_1 from L_2. If so, we may consider such an edge to be part of T', with a weight 0. Then we revert to the prior argument. If it is not possible to add such an edge to T', then there must be an edge $e' \in E(T')$ inducing a partition $L'_1 \cup L'_2$ with $L_i \cap L'_j \neq \emptyset$ for $i, j - 1, 2$. Select $a \in L_1 \cap L'_1$, $b \in L_1 \cap L'_2$, $c \in L_2 \cap L'_1$, and $d \in L_2 \cap L'_2$. Then

$$\begin{aligned} w(e) &= (1/2)(D_T(a,c) + D_T(b,d) - D_T(a,b) - D_T(c,d)) \\ &\leq (1/2)((D_{T'}(a,c) + D_{T'}(b,d) - D_{T'}(a,b) - D_{T'}(c,d) + 4\delta) \\ &\leq 2\delta. \end{aligned}$$

■

It remains to show the same bound holds for the distributions p_{T^*} and p_T. This follows from the following lemma.

Lemma 3.3. *If* $\|d_{T^*} - d_T\|_\infty < \delta$, *then* $\|p_{T^*} - p_T\}_1 = O(|L|\delta)$.

Proof. We can assume that both T and T^* have leaves L and the same number of edges and vertices (by allowing edges to have length 0). The proof proceeds by showing that for each high-length edge in T, there is a corresponding edge of similar length in T^* and vice versa. First we bound the discrepancies in lengths of the corresponding edges and show that the noncorresponding edges must have small length. Then we extend the argument to show that the lengths of paths to the same leaf are either similar or small, which leads to a bound on the difference between p_{T^*} and p_T.

Each edge e in T either corresponds exactly to an edge e^* in T^* with $|w(e) - w^*(e^*)| < 2\delta$, or $|w(e)| < 2\delta$. Recall that the transformation from distances back to probabilities is $p(e) = \exp(-d(e))$. Since $d(e) \geq 0$, the mean value theorem implies that $|\exp(-(d(e) + 2\delta)) - \exp(-d(e))| \leq 2\delta$. Thus each edge e in T either corresponds to an edge e^* in T^* with $|p(e) - p^*(e^*)| < 2\delta$, or $p(e) \geq (1 - 2\delta)$. Similarly, any edge e^* in T^* not corresponding to an edge in T satisfies $p^*(e^*) \geq 1 - 2\delta$.

To bound the distance between distributions, suppose that we sample sets of events from T and T^* as described in Section 2. If edges e and e^* correspond, then we decide whether to include or exclude them by selecting a random number z_e uniformly from [0,1]. If $z_e < \min(p(e), p(e^*))$, both edges are selected, while if $z_e > \max(p(e), p^*(e^*))$, both edges are not selected. Only if z_e falls in between the two edge probabilities, is the edge selection discrepant.

For each edge e in $E(T) \cup E(T^*)$, let the random variable X_e be 0 if the edge e is not selected, and 1 if e is selected. A sufficient condition for producing the same output from T and T^* is that each pair of corresponding edges e, e^* $X_e = X_{e^*}^*$, and for every edge $e \in E(T)\backslash E(T^*)$, and every edge $e \in E(T^*)\backslash E(T)$, $X_e = 0 = X_{e^*}^*$. Conversely, the outputs differ only if either the corresponding edges have different selection status or if an edge with no corresponding partner is selected. The probabilities of these two selection discrepancies gives an upper bound on the variational distance. Specifically,

$$V(T, T^*) \leq \sum_{e \in E(T) \cap E(T^*)} |w(e) - w(e^*)| + \sum_{e \in E(T)\backslash E(T^*)} (1 - p(e)) + \sum_{e^* \in E(T^*)} (1 - p^*(e^*)) = O(|L|\delta).$$

Proof of the Theorem. The pivot method of Agarwala *et al.* (1999) for tree-fitting provides the following guarantee: if there is a path metric D such that $\|D - d\|_\infty = \epsilon$, then the pivot method returns a tree T^* with $\|d_{T^*} - \hat{d}\|_\infty \leq 3\epsilon$. Thus $\|D - d_{T^*}\|_\infty \leq 4\epsilon$. By Lemma 3.1, we know that the expected value $E[\epsilon] = O\left(\frac{|L|}{\sqrt{k p_{\min}}}\right)$. By Lemma 3.2, the same bound holds for the error of edge estimates, and by Lemma 3.2 the L_1 bound is $O(|L|\epsilon)$. ■

The analysis above is fairly loose with regard to $|L|$. However, for the data sets we have looked at, $|L|$ is usually quite small.

4. A DISTANCE-BASED TREE FOR RENAL CANCER

We illustrate the usage of our distance-based method on a set of 116 cases of clear cell renal cell carcinoma from the laboratory of H. M. that was collected using CGH as described in Jiang *et al.* (1998). Kidney cancer is very heterogeneous in its histology and its genetic origin (Erlandsson, 1998). "Clear cell" renal cell carcinoma is one histological category of nonpapillary kidney cancer. Renal cancer is known to have both familial and sporadic forms. The familial forms are typically caused by a germ-line defect (i.e., inherited at birth and present in all cells) in a tumor suppressor gene. One copy of the gene on the two homologous pairs of chromosomes is defective at conception, and the cancer occurs if the other copy becomes defective in a renal cell later on. Most renal cancers (>90% [Motzer *et al.*, 1996]) are sporadic cases where two mutations occur after birth in some renal cells. The gene responsible for clear cell renal cell carcinoma associated with the rare Von Hippel-Lindau Syndrome has been identified on chromosome

arm 3p (Latif *et al.*, 1993). About 70–80% of sporadic clear cell renal cell carcinomas have a loss of the Von Hippel-Lindau gene on chromosome arm 3p (Gnarra *et al.*, 1994; Moch *et al.*, 1998).

Studies of the role of the Von Hippel-Lindau gene in renal cancer suggest that both copies should become defective early in the oncogenesis process, but that a loss of this gene alone may not be sufficient to cause renal cancer. For example, Thrash-Bingham *et al.* (1995) used a different laboratory technique to look only for losses in 33 renal cell carcinomas, and they found a wide variety of chromosomal losses. Among 13 cases are clear cell renal carcinomas, they observed 5 losses on 3p.

The first step of our modeling, before applying the tree-fitting algorithm, is to select a small set of events that appear to be most relevant. Because cancer involves genetic instability and because CGH does have false positives, many of the possible 82 events may appear in a small percentage of tumors. To select events that appear to occur nonrandomly, we used the method of Brodeur *et al.* (1982). This method starts with prior probabilities of all events and uses simulation to derive a distribution under the null hypothesis that all events are random. For each replicate (of 116 tumors) a score is computed relating the most frequent events to their prior probabilities, and the maximum score is recorded. In the real data an event is considered nonrandom if its score is above the 95th percentile of maximum scores from the null distribution. We assumed for the prior distribution that gains and losses have equal probability, and the probability for a chromosome arm is proportional to its size. Arm sizes were derived from Morton (1991). We used 10,000 replicates. The method of Brodeur *et al.* selected 12 events: −3p, −4p, −4q, −6q, −8p, −9p, −13q, −18q, −Xp,+17p, +17q, +Xp.

The distance-based tree (Fig. 2) was constructed by a four-step process. First, the oncogenetic tree inference algorithm described above was used to compute a distance matrix. Second, this distance matrix was fed into both the Fitch and Neighbor programs from the PHYLIP (version 3.5c) package (Felsenstein, 1989) to obtain initial topologies. The Fitch program (Fitch and Margoliash, 1967) finds the tree T^* minimizing

$$\sum_{x,y} \frac{(D^*(x, y) - D(x, y))^2}{D(x, y)^p}$$

where $p = 0, 1$, or 2 depending on the setting. (All three options were tested.) The Neighbor program (Saitou and Nei, 1987) uses the *neighbor-joining* heuristic. Starting from a tree with star topology, it iteratively creates subtrees linking together leaves close to each other (or, at later stages, subtrees relatively

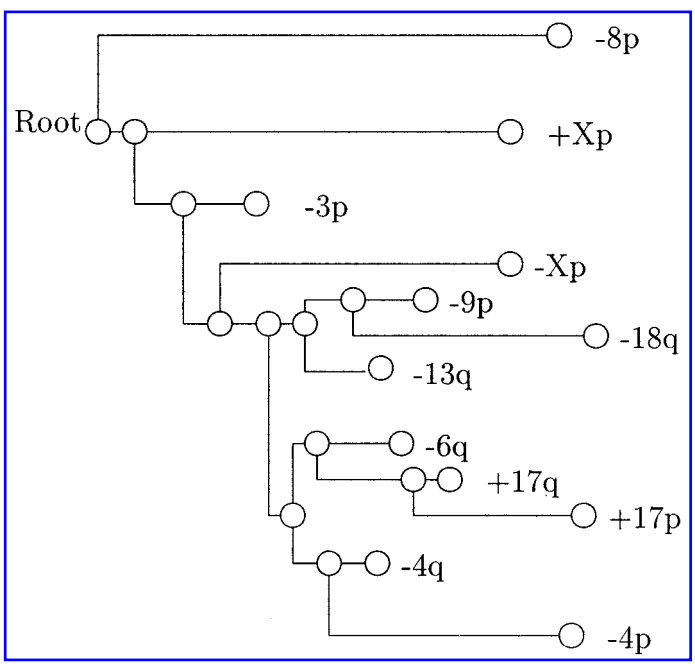

FIG. 2. Distance-based tree for RCC data.

close to each other). Third, a linear program was used upon these topologies to find optimal edge-lengths. The linear program returned some zero-length edges, so nearest-neighbor interchanges (NNIs) were tested across each nonpendant zero-length edge, with edge-lengths re-optimized. Fourth, edge-lengths were converted back to edge-probabilities, and the Fitch tree generated with $p = 2$ was observed to be measurably superior at matching leaf-to-leaf probability weights. This tree is shown in Fig. 2, with edge-lengths drawn to scale as horizontal distances. Vertical distances are included solely to make the picture clearer.

The distance-based tree is consistent with the established theory that a loss on 3p is an early important event for clear cell renal carcinoma and suggests that it is not causatively associated with specific other gains or losses. Our model predicts that a loss on 4q is an important early event for clear cell renal carcinomas and that, even though the loss on 3p is more common, the loss on 4q correlates more highly with the occurrence of other events. The subtrees in our model also predict that there may be at least two subclasses of RCC loosely associated with −4q: one subclass is marked by the events −6q, +17q, +17p, and the other by the events −9p, −13q, −18q. The event −4p is closer to the first subclass, but may represent a third subclass since it appears in its own subtree or may be closer to independent as indicated by the long branch to it. The extremely long branches to −8p and +Xp and the placement to the outside suggest that these two events are more likely to be late effects than early causes and that these two events are not associated with any particular subclass of RCC.

Using the method of Desper *et al.* (1999) we constructed another tree on the same set of events based on a maximum branching in a weighted graph. See Fig. 3. Remarkably, both trees share many relevant properties including:

1. −3p and −4q are important early events near the root.
2. Although −3p is close to the root, −4q sits more centrally in relation to the main body of the tree.
3. +17q is tightly linked with and precedes +17p in a side branch closely related to −6q.
4. Both trees have subtrees with −13q, −9p, and −18q.
5. Both show a close relationship between −4q and −4p.
6. In both trees, −8p is essentially independent from all the other events.

The only significant difference lies with the placement of the relatively rare event +Xp.

In comparing the trees, it is important to keep in mind the different methods of data analysis which lead to the tree construction. The branching in Fig. 3 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 which are incident to the root, such as −3p and −8p, can indicate that the event shares no strong correlations with any other events, as appears to be the case for −8p, or that the event may correlate, but not as much as another event, as appears to be the case for −3p with respect to −4q.

In contrast, the tree-fitting method considers all pairwise correlations simultaneously. 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.

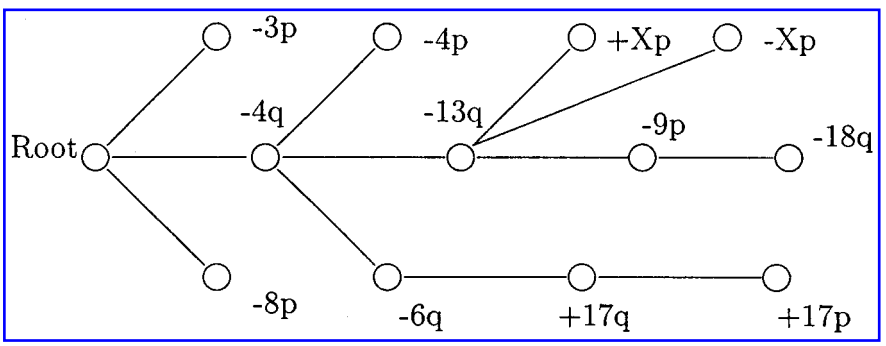

FIG. 3. Maximum-weight branching for RCC data.

5. DISCUSSION AND FURTHER WORK

Identifying early genetic changes in tumors and markers of genetic heterogeneity are important problems in cancer genetics. Comparative genomic hybridization (CGH) is a powerful laboratory technique for identifying genetic changes in tumors. Mathematical modeling of CGH data should help interpret the results. Most of the previous analyses of CGH data have simply counted the frequencies of different events and noticed in various ways that some events occurred much more frequently than would be expected at random (see Forozan *et al.* [1997] and references therein). The collection of large sample sets, such as the renal cancer set used herein, affords the opportunity to analyze co-occurrences of multiple events looking for more complex and comprehensive models of the genetic changes that mark tumor development.

In this paper, we continued our investigation of tree models of tumor progression started in Desper *et al.* (1999). We showed how to transform the probabilistic trees into distance-based phylogenetic trees. We defined the problem of inferring a distance-based oncogenetic tree from CGH data, and we proved that it is possible to infer a tree that is provably not too far in variational distance from the optimal tree in the L_∞ norm. However, other phylogenetic tree inference methods, without provable bounds, are used in practice, and we suggest use of these methods on real data.

Comparing how we use the branching and distance-based tree models to make predictions illustrates some of the strengths of the new 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, distances are measured less precisely by numbers of edges. A CNA that is adjacent to the root in the branching, but has no children (such as -8p in the renal cancer branching tree), might reflect that this CNA is not well correlated with any other CNA rather than that it is an early event. 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, while the branching tree only shows the best correlated pairs. One advantage of the branching trees is that edges between CNAs lead to direct predictions of cause-and-effect relationships.

In practice, as in the renal cancer example above, we construct both types of trees. We look for similarities in which events are near the root to predict the early CNAs. We look for similar clustering in subtrees to predict which events tend to occur together and may mark genetically homogeneous subsets.

Both the distance-based method and the branching-tree method are encoded in software that is freely available by sending e-mail to desper@ncbi.nlm.nih.gov or schaffer@helix.nih.gov. The software includes an implementation of the method of Brodeur *et al.* (1982) to select the nonrandom events. For the branching tree method, our software is self-contained. For the distance-based method, our software goes as far as producing a distance matrix. In practice, we feed the distance matrix into various components of PHYLIP (Felsenstein, 1989) and/or other locally produced software to select the best distance-based tree. The use of external software can be a complication in the short term since there are several choices for how to produce the best distance-based tree and nontrivial user intervention is needed. In the long run, reducing tumor progression modeling to phylogenetic tree inference should be an advantage, since we would expect new and better methods to be developed for tree inference due to its many uses.

We illustrated our distance-based oncogenetic tree methods on a large CGH data set for renal cell carcinoma (RCC), and we also compared distance-based and branching trees. The distance-based tree is consistent with the established theory that -3p is an important early event in RCC. The distance-based tree supports the prediction we made from the branching tree (Desper *et al.*, 1999) that -4q is another important early event in RCC. Both the distance-based tree and the branching tree suggest that there may be two classes of RCC in which -4q occurs: one class marked by -6q, $+17$q, $+17$p, and the other class marked by -9q, -13p, -18q. The distance-based tree clarifies that -8p is largely independent of other events, which is consistent with the branching tree, but is not the only possible interpretation of the branching tree.

Not nearly enough is known experimentally about tumor progression and CGH measurements to validate our models. It is extremely hard to figure out the early events in the laboratory, so our tree models should be very helpful to cancer geneticists. Since Desper *et al.* (1999) was published and this paper was submitted for publication, our tree modeling methods have been used in several studies, including one on ovarian cancer (Simon *et al.*, 2000), and one on breast cancer (Kainu *et al.*, 2000), which have already been

published. In the latter study, the tree models provided one line of evidence that the long-sought third gene for hereditary susceptibility to breast cancer may be on chromosome 13.

Future theoretical work will include development of maximum likelihood methods to compare trees of various topologies. Also, the underlying mathematical models are sufficiently general to be applied to non-CGH data, and we are applying the methods described to some large breakpoint data sets, as well as other CGH data sets. The essential purpose of our mathematical modeling is to suggest directions for experimental follow-up to better understand the genetic changes that occur in human cancer cells.

ACKNOWLEDGMENTS

Special thanks go to Martin Farach-Colton for his assistance as Richard Desper's Ph.D. advisor at Rutgers University. His work at Rutgers University was supported in part by NSF grant 94-12594. Mr. Desper's work was done partly while he was a summer student intern at N.I.H. The work of Feng Jiang and Holger Moch was supported in part by grant number 31-50752.97 from the Swiss National Science Foundation. The work of Christos H. Papadimitriou was supported in part by NSF grant CCR-9626361. Thanks to Richard Simon for suggesting the method of Brodeur *et al.* (1982) and to Greg Schuler for referring us to the tables in Morton (1991).

REFERENCES

Agarwala, R., Bafna, V., Farach, M., Paterson, M., Thorup, M. 1999. On the approximability of numerical taxonomy (fitting distances by tree metrics). *SIAM J. Comp.* 28, 1073–1085.

Ambainis, A., Desper, R., Farach, M., and Kannan, S. 1997. Nearly tight bounds on the learnability of evolution, *in Proc. 38th Symp. Found. of Comp. Sci.* 524–533.

Barthélemy, J.-P., and Guénoche, A. 1991. *Trees and Proximity Representations*, Wiley, New York.

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

Buetow, K.H., Edmonson, M.N., Shen, F.M., Chen, G.C., London, W.T., and McGlynn, K.A. 1998. Identification of molecular heterogeneity in HCC using STRPs and tree-building algorithms. *Am. J. Hum. Genet.* 63, A336, (abst).

Cavender, J.A. 1978. Taxonomy with confidence. *Math. Biosci.* 40, 271–280.

Cohen, J., and Farach, M. 1997. Numerical taxonomy on data: Experimental results. *J. Comp. Biol.* 4, 547–558.

Day, W.H.E. 1987. Computational complexity of inferring phylogenies from dissimilarity matrices. *Bull. Math. Biol.* 49, 461–467.

Desper, R., Jiang, F., Kallioniemi, O.-P., Moch, H., Papadimitriou, C.H., Schäffer, A.A. 1999. Inferring tree models for oncogenesis from comparative genomic hybridization data. *J. Comp. Biol.* 6, 37–51.

Erlandsson, R. 1998. Molecular genetics of renal cell carcinoma. *Cancer Genet. Cytogenet.* 104, 1–18.

Farach, M., and Kannan, S. 1996. Efficient algorithms for inverting evolution. *Proc. 28th ACM Symp. Theory Comput.*, 230–236.

Farris, J.S. 1973. A probability model for inferring evolutionary trees. *Syst. Zool.* 22, 250–256.

Fearon, E., and Vogelstein, B. 1990. A genetic model for colorectal tumorigenesis. *Cell* 61, 759–776.

Felsenstein, J. 1989. PHYLIP—Phylogeny inference package (ver. 3.2). *Cladistics* 5, 164–166. See also http://evolution.genetics.washington.edu/phylip.html.

Fitch, W.M., and Margoliash, E. 1967. Construction of phylogenetic trees. *Science* 155, 279–284.

Forozan, F., Karhu, R., Kononen, J., Kallioniemi, A., and Kallioniemi, O.-P. 1997. Genome screening by comparative genome hybridization. *Trends Genet.* 13, 405–409.

Gnarra, J.R., Tory, K., Weng, Y., *et al.* 1994. Mutations of the VHL tumour suppressor gene in renal carcinoma. *Nature Genet.* 7, 85–90.

Jiang, F., Richter, J., Schraml, P., *et al.* 1998. Chromosomal imbalances in papillary renal cell carcinoma: Genetic differences between histological subtypes. *Am. J. Pathol.* 153, 1467–1473.

Kainu, T., Juo, S-H.H., Desper, R., *et al.* 2000. Somatic deletions in hereditary breast cancers implicate 13q21 as a putative novel breast cancer susceptibility locus. *Proc. Nat. Acad. Sci. USA* 97, 9603–9608.

Kallioniemi, A., Kallioniemi, O.P., Sudar, D., *et al.* 1992. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* 258, 818–821.

Karp, R.M. 1971. A simple derivation of Edmonds' algorithm for optimum branching. *Networks* 1, 265–272.

Kuukasjärvi, T., Karhu, R., Tanner, M., *et al.* 1997. Genetic heterogeneity and clonal evolution underlying development of asynchronous metastasis in human breast cancer. *Cancer Res.* 57, 1597–1604.

Latif, F., Tory, K., Gnarra, J., *et al.* 1993. Identification of the von Hippel-Lindau disease tumor suppressor gene. *Science* 260, 1317–1320.

Moch, H., Presti, J.C., Sauter, G., *et al.* 1996. Genetic aberrations detected by comparative genomic hybridization are associated with clinical outcome in renal cell carcinoma. *Cancer Res.* 56, 27–30.

Moch, H., Schraml, P., Bubendorf, L., *et al.* 1998. Intratumoral heterogeneity of von Hippel-Lindau gene deletions in renal cell carcinoma detected by fluorescence in situ hybridization. *Cancer Res.* 58, 2304–2309.

Morton, N.E. 1991. Parameters of the human genome. *Proc. Nat. Acad. Sci. USA* 88, 7474–7476.

Motzer, R.J., Bander, N.H., and Nanus, D.M. 1996. Renal-cell carcinoma. *N. Engl. J. Med.* 335, 865–875.

Newton, M.A., Wu, S.-Q., and Reznikoff, C.A. 1994. Assessing the significance of chromosome-loss data: Where are suppressor genes for bladder cancer? *Stat. Med.* 13, 839–858.

Neyman, J. 1971. Molecular studies of evolution: A source of novel statistical problems, *in Statistical Decision Theory and Related Topics*, 1–27, Academic Press, New York.

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

Saitou, N., and Nei, M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–424.

Simon, R., Desper, R., and Papadimitriou, C.H., *et al.* 2000. Chromosome abnormalities in ovarian adenocarcinoma III: Using breakpoint data to infer and test mathematical models for oncogenesis. *Genes, Chromosomes and Cancer* 28, 106–120.

Sokal, R.R., and Rohlf, F.J. 1995. *Biometry.* W.H. Freeman, New York.

Swofford, D.L., and Olsen, G.J. 1990. Phylogeny reconstruction, *in Molecular Systematics*, 411–501. Sinauer Associates, Sunderland, MA.

Thrash-Bingham, C.A., Salazar, H., Freed, J.J., Greenberg, R.E., and Tartof, K.D. 1995. Genomic alterations and instabilities in renal cell carcinomas and their relationship to tumor pathology. *Cancer Res.* 55, 6189–6195.

Vogelstein, B., Fearon, E., Hamilton, S., *et al.* 1988. Genetic alterations during colorectal tumor development. *N. Engl. J. Med.* 319, 525–532.

Waterman, M.S., Smith, T.F., Singh, M., and Beyer, W.A. 1977. Additive evolutionary trees. *J. Theor. Biol.* 64, 199–213.

Address correspondence to:
Alejandro Schäffer
NCBI/NIH
Building 38A, Room 8N 805
8600 Rockville Pike
Bethesda, MD 20894

E-mail: schaffer@helix.nih.gov

This article has been cited by:

1. Darawalee Wangsa, Salim Akhter Chowdhury, Michael Ryott, E. Michael Gertz, Göran Elmberger, Gert Auer, Elisabeth Åvall Lundqvist, Stefan Küffer, Philipp Ströbel, Alejandro A. Schäffer, Russell Schwartz, Eva Munck-Wikland, Thomas Ried, Kerstin Heselmeyer-Haddad. 2016. Phylogenetic analysis of multiple FISH markers in oral tongue squamous cell carcinoma suggests that a diverse distribution of copy number changes is associated with poor prognosis. *International Journal of Cancer* **138**, 98-109. [CrossRef]
2. Ramon Diaz-Uriarte. 2015. Identifying restrictions in the order of accumulation of mutations during tumor progression: effects of passengers, evolutionary models, and sampling. *BMC Bioinformatics* **16**. . [CrossRef]
3. Philipp M. Altrock, Lin L. Liu, Franziska Michor. 2015. The mathematics of cancer: integrating quantitative models. *Nature Reviews Cancer* **15**, 730-745. [CrossRef]
4. Daniele Ramazzotti, Giulio Caravagna, Loes Olde Loohuis, Alex Graudenzi, Ilya Korsunsky, Giancarlo Mauri, Marco Antoniotti, Bud Mishra. 2015. CAPRI: efficient inference of cancer progression models from cross-sectional data. *Bioinformatics* **31**, 3016-3026. [CrossRef]
5. Salim Akhter Chowdhury, E. Michael Gertz, Darawalee Wangsa, Kerstin Heselmeyer-Haddad, Thomas Ried, Alejandro A. Schäffer, Russell Schwartz. 2015. Inferring models of multiscale copy number evolution for single-tumor phylogenetics. *Bioinformatics* **31**, i258-i267. [CrossRef]
6. Raphael Benjamin J., Vandin Fabio. 2015. Simultaneous Inference of Cancer Pathways and Tumor Progression from Cross-Sectional Mutation Data. *Journal of Computational Biology* **22**:6, 510-527. [Abstract] [Full Text HTML] [Full Text PDF] [Full Text PDF with Links]
7. Paola Lecca, Nicola Casiraghi, Francesca Demichelis. 2015. Defining order and timing of mutations during cancer progression: the TO-DAG probabilistic graphical model. *Frontiers in Genetics* **6**. . [CrossRef]
8. N. Beerenwinkel, R. F. Schwarz, M. Gerstung, F. Markowetz. 2015. Cancer Evolution: Mathematical Models and Computational Inference. *Systematic Biology* **64**, e1-e25. [CrossRef]
9. Hao Wu, Lin Gao, Nikola Kasabov. 2015. Inference of cancer progression from somatic mutation data. *IFAC-PapersOnLine* **48**, 234-238. [CrossRef]
10. Loes Olde Loohuis, Andreas Witzel, Bud Mishra. 2014. Improving Detection of Driver Genes: Power-Law Null Model of Copy Number Variation in Cancer. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **11**, 1260-1263. [CrossRef]
11. Loes Olde Loohuis, Andreas Witzel, Bud Mishra. 2014. Cancer hybrid automata: Model, beliefs and therapy. *Information and Computation* **236**, 68-86. [CrossRef]
12. E. Purdom, C. Ho, C. S. Grasso, M. J. Quist, R. J. Cho, P. Spellman. 2013. Methods and challenges in timing chromosomal abnormalities within cancer samples. *Bioinformatics* . [CrossRef]
13. F. Strino, F. Parisi, M. Micsinai, Y. Kluger. 2013. TrAp: a tree approach for fingerprinting subclonal tumor composition. *Nucleic Acids Research* **41**, e165-e165. [CrossRef]
14. T. Sakoparnig, N. Beerenwinkel. 2012. Efficient sampling for Bayesian inference of conjunctive Bayesian networks. *Bioinformatics* **28**, 2318-2324. [CrossRef]
15. Katrin Hainke, Jörg Rahnenführer, Roland Fried. 2012. Cumulative disease progression models for cross-sectional data: A review and comparison. *Biometrical Journal* **54**:10.1002/bimj.v54.5, 617-640. [CrossRef]
16. Xiaobo Li, Jian Chen, Bingjian Lü, Sihua Peng, Richard Desper, Maode Lai. 2011. –8p12–23 and +20q Are Predictors of Subtypes and Metastatic Pathways in Colorectal Cancer: Construction of Tree Models Using Comparative Genomic Hybridization Data. *OMICS: A Journal of Integrative Biology* **15**:1-2, 37-47. [Abstract] [Full Text HTML] [Full Text PDF] [Full Text PDF with Links] [Supplemental Material]

17. Michal Ozery-Flato, Chaim Linhart, Luba Trakhtenbrot, Shai Izraeli, Ron Shamir. 2011. Large-scale analysis of chromosomal aberrations in cancer karyotypes reveals two distinct paths to aneuploidy. *Genome Biology* **12**, R61. [CrossRef]

18. Thomas Longerich, Michael Martin Mueller, Kai Breuhahn, Peter Schirmacher, Axel Benner, Christiane Heiss. 2011. Oncogenetic tree modeling of human hepatocarcinogenesis. *International Journal of Cancer* n/a-n/a. [CrossRef]

19. Nitin Kumar, Hubert Rehrauer, Haoyang Cai, Michael Baudis. 2011. CDCOCA: A statistical method to define complexity dependence of co-occuring chromosomal aberrations. *BMC Medical Genomics* **4**, 21. [CrossRef]

20. H. Lilljebjorn, C. Soneson, A. Andersson, J. Heldrup, M. Behrendtz, N. Kawamata, S. Ogawa, H. P. Koeffler, F. Mitelman, B. Johansson, M. Fontes, T. Fioretos. 2010. The correlation pattern of acquired copy number changes in 164 ETV6/RUNX1-positive childhood acute lymphoblastic leukemias. *Human Molecular Genetics* **19**, 3150-3158. [CrossRef]

21. M. Gerstung, M. Baudis, H. Moch, N. Beerenwinkel. 2009. Quantifying cancer progression with conjunctive Bayesian networks. *Bioinformatics* **25**, 2809-2815. [CrossRef]

22. J. Liu, N. Bandyopadhyay, S. Ranka, M. Baudis, T. Kahveci. 2009. Inferring progression models for CGH data. *Bioinformatics* **25**, 2208-2215. [CrossRef]

23. Xiao-bo Li. 2009. Mathematical modeling of carcinogenesis based on chromosome aberration data. *Chinese Journal of Cancer Research* **21**, 240-246. [CrossRef]

24. Martha L. Slattery, Karen Curtin, Roger K. Wolff, Kenneth M. Boucher, Carol Sweeney, Sandra Edwards, Bette J. Caan, Wade Samowitz. 2009. A Comparison of Colon and Rectal Somatic DNA Alterations. *Diseases of the Colon & Rectum* **52**, 1304-1311. [CrossRef]

25. Swapnali Pathare, Alejandro A. Schäffer, Niko Beerenwinkel, Manoj Mahimkar. 2009. Construction of oncogenetic tree models reveals multiple pathways of oral cancer progression. *International Journal of Cancer* **124**:10.1002/ijc.v124:12, 2864-2871. [CrossRef]

26. Yutaka Midorikawa, Shogo Yamamoto, Shingo Tsuji, Naoko Kamimura, Shumpei Ishikawa, Hisaki Igarashi, Masatoshi Makuuchi, Norihiro Kokudo, Haruhiko Sugimura, Hiroyuki Aburatani. 2009. Allelic imbalances and homozygous deletion on 8p23.2 for stepwise progression of hepatocarcinogenesis. *Hepatology* **49**:10.1002/hep.v49:2, 513-522. [CrossRef]

27. Carol Sweeney, Kenneth M. Boucher, Wade S. Samowitz, Roger K. Wolff, Hans Albertsen, Karen Curtin, Bette J. Caan, Martha L. Slattery. 2009. Oncogenetic tree model of somatic mutations and DNA methylation in colon tumors. *Genes, Chromosomes and Cancer* **48**:10.1002/gcc.v48:1, 1-9. [CrossRef]

28. Xiao-Bo LI. 2008. Exploration of carcinogenesis based on tree models using CGH data. *Hereditas (Beijing)* **30**, 407-412. [CrossRef]

29. B Gunawan, A von Heydebreck, B Sander, H-J Schulten, F Haller, C Langer, T Armbrust, M Bollmann, S Gašparov, D Kovač, L Füzesi. 2007. An oncogenetic tree model in gastrointestinal stromal tumours (GISTs) identifies different pathways of cytogenetic evolution with prognostic implications. *The Journal of Pathology* **211**:10.1002/path.v211:4, 463-470. [CrossRef]

30. Lei Chen, Carola Nordlander, Afrouz Behboudi, Björn Olsson, Karin Klinga Levan. 2007. Deriving evolutionary tree models of the oncogenesis of endometrial adenocarcinoma. *International Journal of Cancer* **120**:10.1002/ijc.v120:2, 292-296. [CrossRef]

31. Marcus Hjelm, Mattias Höglund, Jens Lagergren. 2006. New Probabilistic Network Models and Algorithms for Oncogenesis. *Journal of Computational Biology* **13**:4, 853-865. [Abstract] [Full Text PDF] [Full Text PDF with Links]

32. Friedrich W. Cremer, Jelena Bila, Isabelle Buck, Mutlu Kartal, Dirk Hose, Carina Ittrich, Axel Benner, Marc S. Raab, Ann-Cathrin Theil, Marion Moos, Hartmut Goldschmidt, Claus R. Bartram, Anna Jauch. 2005. Delineation of distinct subgroups of multiple myeloma and a model

for clonal evolution based on interphase cytogenetics. *Genes, Chromosomes and Cancer* **44**:10.1002/gcc.v44:2, 194-203. [CrossRef]

33. Niko Beerenwinkel, Jörg Rahnenführer, Martin Däumer, Daniel Hoffmann, Rolf Kaiser, Joachim Selbig, Thomas Lengauer. 2005. Learning Multiple Evolutionary Pathways from Cross-Sectional Data. *Journal of Computational Biology* **12**:6, 584-598. [Abstract] [Full Text PDF] [Full Text PDF with Links]
34. Mattias H�glund, Attila Frigyesi, Torbj�rn S�ll, David Gisselsson, Felix Mitelman. 2005. Statistical behavior of complex cancer karyotypes. *Genes, Chromosomes and Cancer* **42**:10.1002/gcc.v42:4, 327-341. [CrossRef]
35. Zhongxi Huang, Richard Desper, Alejandro A. Sch�ffer, Zhihua Yin, Xin Li, Kaitai Yao. 2004. Construction of tree models for pathogenesis of nasopharyngeal carcinoma. *Genes, Chromosomes and Cancer* **40**:10.1002/gcc.v40:4, 307-315. [CrossRef]
36. Michael A Newton. 2002. Discovering Combinations of Genomic Aberrations Associated With Cancer. *Journal of the American Statistical Association* **97**, 931-942. [CrossRef]
37. Qiang Huang, Guo Pei Yu, Steven A. McCormick, Juan Mo, Bhakti Datta, Manoj Mahimkar, Philip Lazarus, Alejandro A. Sch�ffer, Richard Desper, Stimson P. Schantz. 2002. Genetic differences detected by comparative genomic hybridization in head and neck squamous cell carcinomas from different tumor sites: construction of oncogenetic trees for tumor progression. *Genes, Chromosomes and Cancer* **34**:10.1002/gcc.v34:2, 224-233. [CrossRef]
38. MICHAEL D RADMACHER, RICHARD SIMON, RICHARD DESPER, RAYMOND TAETLE, ALEJANDRO A SCHÄFFER, MARK A NELSON. 2001. Graph Models of Oncogenesis with an Application to Melanoma. *Journal of Theoretical Biology* **212**, 535-548. [CrossRef]
39. Pei Wang, Young Kim, J. Pollack, R. TibshiraniBoosted PRIM with application to searching for oncogenic pathway of lung cancer 573-578. [CrossRef]
40. H.K. GillOncogenetics tree models - an estimation 426. [CrossRef]