**JOURNAL OF COMPUTATIONAL BIOLOGY**
**Volume 12, Number 6, 2005**

**Pp. 584–598**

# Learning Multiple Evolutionary Pathways from Cross-Sectional Data

NIKO BEERENWINKEL,[1] JÖRG RAHNENFÜHRER,[1] MARTIN DÄUMER,[2]
DANIEL HOFFMANN,[3] ROLF KAISER,[2] JOACHIM SELBIG,[4]
and THOMAS LENGAUER[1]

## ABSTRACT

**We introduce a mixture model of trees to describe evolutionary processes that are characterized by the ordered accumulation of permanent genetic changes. The basic building block of the model is a directed weighted tree that generates a probability distribution on the set of all patterns of genetic events. We present an EM-like algorithm for learning a mixture model of $K$ trees and show how to determine $K$ with a maximum likelihood approach. As a case study, we consider the accumulation of mutations in the HIV-1 reverse transcriptase that are associated with drug resistance. The fitted model is statistically validated as a density estimator, and the stability of the model topology is analyzed. We obtain a generative probabilistic model for the development of drug resistance in HIV that agrees with biological knowledge. Further applications and extensions of the model are discussed.**

**Key words:** mixture models, tree models, Bayesian networks, EM algorithm, HIV drug resistance, mutational pathways.

## 1. INTRODUCTION

DESPITE THE INTRODUCTION OF 18 DIFFERENT DRUGS that inhibit replication of human immunodeficiency virus type 1 (HIV-1), therapeutic success is still limited. A major factor contributing to therapy failure even of modern combination therapies (highly active antiretroviral therapy, HAART) is the virus' ability to escape from drug pressure by developing drug resistance (Perrin and Telenti, 1998; Vandamme *et al.*, 1999). This escape mechanism is based on HIV's high rates of replication and mutation. Residual viral reproduction under therapy allows for generating genetic variants that have a selective advantage under drug pressure. Eventually, some of these mutants replicate as well or nearly as well as a wild type virus and thus lead to viral rebound.

Considerable work has been carried out on characterizing the relationship between genetic changes in the viral drug targets and phenotypic drug resistance. Many single mutations have been linked to resistance against one or more drugs (Shafer, 2000). Mutational patterns conferring resistance have been identified

[1]Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, D-66123 Saarbrücken, Germany.
[2]Institute of Virology, University of Cologne, Fürst-Pückler-Str. 56, D-50935 Köln, Germany.
[3]Center of Advanced European Studies and Research, Ludwig-Erhard-Allee 2, D-53175 Bonn, Germany.
[4]Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, D-14476 Golm, Germany.

by several statistical and machine learning methods (Beerenwinkel *et al.*, 2001a, 2002; Sevin *et al.*, 2000). Some computational approaches make use of this (either data-derived or expert) knowledge in order to find optimal drug combinations based on sequence information on the viral drug targets (Beerenwinkel *et al.*, 2003; Lathrop and Pazzani, 1999). In particular, it has been shown that predictions of clinical response can benefit from exploring possible mutational pathways of the virus.

However, much less is known about *how* resistance-associated mutations accumulate. Some mutations are known to occur preferentially in clusters (Beerenwinkel, 2001b; Gonzales *et al.*, 2003; Wu *et al.*, 2003), but the order of accumulation is usually unknown. Only a few studies based on longitudinal (time series) data have revealed directed dependencies between mutations (Boucher *et al.*, 1992; Molla *et al.*, 1996). This type of analysis is not practical for many different drugs or even drug combinations, because large longitudinal samples from patients under the same therapeutic regimen are difficult to obtain. As an alternative, we propose a method for estimating mutational pathways from cross-sectional data (i.e., data from different patients at different time points), which are much more abundant.

We develop the technique in a general setting and consider the development of HIV-1 drug resistance to the nucleoside reverse transcriptase inhibitor zidovudine as a test case.

### *1.1. Zidovudine resistance*

Zidovudine (AZT) was approved for clinical use in 1987 as the first anti-HIV drug. Soon after approval, mutations in HIV-1 reverse transcriptase (RT) were found that decrease susceptibility to the drug and develop within a few month of therapy (Larder *et al.*, 1989; Larder and Kemp, 1989). The most common RT mutations ("classical zidovudine mutations") that develop under zidovudine therapy are M41L,[1] D67N, K70R, L210W, T215F/Y, and K219E/Q. Other mutations such as the multinucleoside resistance mutations V75I, F77L, F116Y, Q151M, and the two amino acids insertion after position 69 are less frequent and typically occur under prolonged combination therapies containing two or more nucleoside RT inhibitors. K70R and T215F/Y are generally the first mutations to occur causing 4- to 8- and 10- to 16-fold zidovudine resistance *in vitro*, respectively (Boucher *et al.*, 1992; Larder, 1994). The double mutant 41L+215F/Y already causes 60- to 70-fold resistance. 41L, 215F/Y and 210W tend to occur together (215-41 pathway) as well as 70R and 219E/Q (70-219 pathway). Substitution M41L may also appear first, albeit less frequently. However, in contrast to the 70R+215F/Y double mutant, the 41L+70R co-occurrence is hardly ever observed. This discrepancy is explained by a replication defect of the 41L+70R intermediate (Jeeninga *et al.*, 2001).

In general, the evolution of drug resistance is driven by several factors including codon usage bias (Keulen *et al.*, 1996), random genetic drift, and natural selection (viral fitness). Under therapy, the viral population is exposed to a strong selective pressure. Mutations almost always arise one at a time, and each single advantageous mutation must be fixed into the population. In consequence, relatively few evolutionary pathways lead from the wild type to a highly resistant, well replicating mutant as compared to the large number of possible mutational patterns (Hall, 2002).

### *1.2. Outline*

We describe the evolution of drug resistance as the accumulation of permanent genetic changes. Our model of the stochastic evolutionary process aims at identifying directed dependencies between mutational events. The basic building block of the model is a directed tree. Vertices of the tree represent binary random variables, each indicating the occurrence of an event (mutation). Edge weights represent conditional probabilities between events with the constraint that a child event does not occur whenever the parent event has not occurred. These restricted Bayesian tree models have been pioneered by Desper *et al.* (1999) in the context of oncogenesis. In Section 3, we recall basic model properties and an efficient combinatorial algorithm for tree reconstruction from observed (cross-sectional) patterns of events.

The tree models provide a detailed and interpretable description of the process of accumulating genetic changes. They represent a considerable improvement over independence or linear path models. Nevertheless, we will see that the special tree structure fits only certain subgroups of the data. We interpret this

[1]We use the syntax $a\,x\,b$ (or simply $x\,b$) to denote amino acid substitutions in RT, where $a$ is the amino acid in the reference strain HXB2 at position $x$ and $b$ the mutated residue.

shortcoming as indicating that the data has been generated by more than one (tree-like) process. Therefore, we introduce the broader class of mixture models of trees. Ideally, we would like to identify multiple evolutionary processes acting on the same gene (or genome), each process in one specialized component of the mixture model. In particular, we will introduce a "noise component" that includes all otherwise unexplained samples.

After reviewing related work in Section 2, we define the class of mixture models in Section 4 and present an EM-like algorithm for learning structure and parameters of the model from data. In Section 5, we illustrate how model selection (choosing the optimal number of trees) can be performed in a maximum likelihood (ML) fashion. We compare the mixture model for the development of zidovudine resistance with biological knowledge. In Section 6, we present cross-validation and bootstrap methods for the validation of our model. Section 7 discusses further applications and extensions of the method.

## 2. RELATED WORK

Chow and Liu (1968) have used unrestricted Bayesian tree models to approximate multivariate discrete probability distributions. They show that solving the maximum weight spanning tree problem in the complete graph with edges between features (events) weighted by their mutual information provides an ML tree estimate. Maximum weight branchings have been proposed in a similar setting (Heckerman *et al.*, 1995). Our mixture models are similar in spirit to the work of Meilă and Jordan (2000), who apply an EM algorithm to generalize the Chow–Liu algorithm to mixtures of undirected trees. Friedman *et al.* (1997) have extended the Chow–Liu procedure for classification tasks.

Related graph models have been developed for oncogenesis, where chromosomal losses and gains are considered as events. The distance matrix between events $u$ and $v$ defined by

$$-2 \log \Pr(u, v) + \log \Pr(u) + \log \Pr(v)$$

has been used as input for distance-based phylogeny methods like neighbor-joining (Desper *et al.*, 2000). The resulting phylogenetic tree represents events as leaves of the tree and groups closely related events together. Internal vertices are considered "hidden events" and do not have a direct interpretation.

A similar approach uses an ML estimation procedure for tree fitting (von Heydebreck *et al.*, 2004). A closed formula for the ML parameters of a tree is derived, while searching for the ML topology is done heuristically.

Finally, generalizing from tree models, directed acyclic graph (DAG) models have been proposed (Radmacher *et al.*, 2001; Simon *et al.*, 2000). Here, vertices represent subsets of the set of events, and an edge $\{u\} \to \{u, v\}$ represents the probability that $u$ occurs first and $v$ occurs second. Edges for larger subsets are defined similarly. For limited subset size, ML estimation of model parameters is feasible.

## 3. MUTAGENETIC TREES

### 3.1. Data representation

We consider $\ell$ different events $\{1, \ldots, \ell\}$, including a special "null event" that has initially occurred in all samples. A pattern $x_i$ of events is represented by a row vector of indicator variables of length $\ell$:

$$x_i = (x_{i1}, \ldots, x_{i\ell}),$$

$$x_{ij} = \begin{cases} 1, & \text{if event } j \text{ has occurred in sample } i \\ 0, & \text{else.} \end{cases}$$

Thus, a set of $N$ observed patterns is represented by the binary matrix

$$X = (x_{ij})_{\substack{1 \le i \le N \\ 1 \le j \le \ell}}.$$

We denote by $\Omega = 2^{\{1,\ldots,\ell\}}$ the set of all possible patterns of length $\ell$.

For the case of HIV-1 zidovudine resistance, we consider a set of $N = 364$ samples derived from previously untreated patients under zidovudine mono-therapy as available from the Stanford HIV Drug Resistance Database (Rhee *et al.*, 2003). No resistance-associated mutations other than the six classical zidovudine mutations are present in this dataset. Thus, the set of $\ell = 7$ events comprises 41L, 67N, 70R, 210W, 215F/Y, and 219E/Q, plus the initial null event characterized by M41, D67, K70, L210, T215, and K219 and referred to as the "wild type." The dataset contains 35 different mutational patterns. The null pattern representing the wild type is observed in 115 samples.

We reconstruct dependencies between events from the joint probabilities between all pairs of events, which can be estimated reliably from moderately large datasets.

### 3.2. Definition

To describe the ordered accumulation of genetic changes, we consider directed trees over the set of events, where each edge is weighted with the conditional probability of the child event given that the parent event has occurred. Formally, a *mutagenetic tree*[2] $\mathcal{T} = (V, E, r, p)$ consists of a set of vertices $V = 1, ..., \ell$ representing events, a set of edges $E$, a special vertex $r \in V$, and a map $p : E \to [0, 1]$ such that

- $(V, E)$ is a branching, i.e., a digraph whose underlying undirected graph is a forest, and each vertex has at most one entering edge,
- the vertex $r$ represents the null event and has no entering edge,
- for all edges $e = (u, v) \in E$,
  —$p(e) = \Pr(v|u)$ is the conditional probability of event $v$ given that event $u$ has occurred,
  —$p(e) > 0$ (if $p(e) = 0$, we can delete $e$ from $E$),
  —$p(e) < 1$ if $e$ leaves the root (if $\Pr(v|r) = 1$, events $v$ and $r$ can be merged).

Note that a mutagenetic tree can have more than one connected component. However, most of the time we will be concerned only with the arborescence (connected branching) containing the special root vertex $r$. Figure 1 shows a mutagenetic tree for the development of resistance to zidovudine.

A mutagenetic tree induces a probability distribution on the set $\Omega$ of all possible mutational patterns as follows. Draw each edge independently from $E$ with probability $p(e)$. Then the set of vertices that are reachable from the root is the outcome of the experiment.

### 3.3. Tree reconstruction

Desper *et al.* (1999) have shown how to reconstruct the mutagenetic tree from all pairwise joint probabilities of events. Consider the complete digraph $G = (V, V \times V, w)$ on the set of vertices $V$ corresponding to the events with weights

$$w(u, v) = \log \Pr(u, v) - \log(\Pr(u) + \Pr(v)) - \log \Pr(v),$$

where $\Pr(u)$ denotes the marginal probability of event $u$ and $\Pr(u, v)$ the joint probability of events $u$ and $v$. Then the mutagenetic tree is the branching in $G$ that maximizes the sum of its edge weights. The maximum weight branching can be computed in $O(|V||E|)$ time by Edmonds' branching algorithm (Chu and Liu, 1965; Edmonds, 1967; Karp, 1971; Tarjan, 1977).

In practice, we do not know the joint probabilities of events, but have to estimate them from the data. For sufficiently many samples, the above algorithm will reconstruct the correct mutagenetic tree with high probability (see Desper *et al.* [1999] for proofs and a quantitative version of this statement).

Finally, if the observed sample does not come from a distribution generated by a mutagenetic tree, we hope that the reconstructed tree captures many of the strong dependencies (causality flows) between events.

The weight function $w$ scores edges $e = (r, v)$ leaving the root with $w(e) = -\log(1 + \Pr(v))$. The scoring implies that less frequent events are favored as initial vertices in the tree. This behavior appears to

[2]We follow the notation of Desper *et al.* (1999) who call these tree models in the context of oncogenesis *oncogenetic trees*.

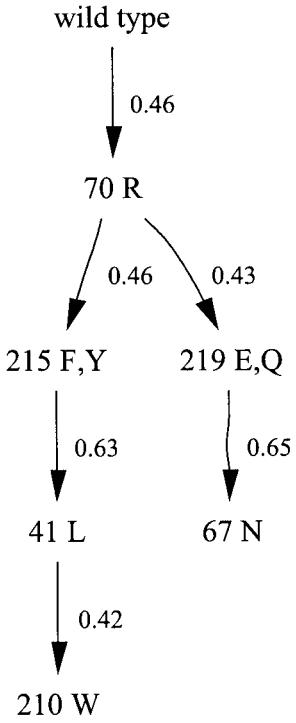


**FIG. 1.** Mutagenetic tree for the development of zidovudine resistance. Nodes are labeled with resistance-associated mutations in the HIV-1 reverse transcriptase; edge labels represent conditional probabilities between mutational events.

be undesirable for some applications. As an alternative, we propose the special weights $w(r, v) = \log \Pr(v)$ for edges leaving the root vertex. This modified scoring function is used for the zidovudine data.

### *3.4. Likelihood computation*

Given a mutagenetic tree $\mathcal{T} = (V, E, r, p)$, the likelihood of a pattern $x$ of events is the probability that $\mathcal{T}$ generates $x$: $L(x|\mathcal{T}) = \Pr(x|\mathcal{T})$. Let $S \subseteq V$ be the set of events specified by $x$. If there is a subset $E' \subseteq E$ such that $S$ is exactly the set of all vertices reachable from $r$ in the subtree $(V, E')$, then $x$ can be generated by $\mathcal{T}$, and the likelihood is given by

$$L(x|\mathcal{T}) = \prod_{e \in E'} p(e) \cdot \prod_{e \in (S \times V \setminus S)} (1 - p(e)) .$$

If there is no such edge subset, the topology of $\mathcal{T}$ does not allow for generating $x$, and hence $L(x|\mathcal{T}) = 0$. Note that the set $E'$ is well defined, because $(V, E)$ is a tree. For example, for the mutagenetic tree displayed in Fig. 1 and the pattern defined by the mutations 70R and 219Q, we find

$$L(x|\mathcal{T}) = 0.46 \cdot 0.43 \cdot (1 - 0.46) \cdot (1 - 0.65) = 0.037,$$

whereas the pattern composed of 215Y and 41L can not be generated by that tree.

The likelihood computation can be done efficiently by traversing the mutagenetic tree in a breadth-first search starting from $r$. Note that connected components of $\mathcal{T}$ other than the arborescence rooted at $r$ do not affect the likelihood of a pattern.

We call $\mathcal{T}$ a star, if all edges $e \in E$ leave the root vertex $r$. A star models events as being inpedendent of each other. In terms of the likelihood a star is characterized as follows.

**Lemma 1.** *A mutagenetic tree is a star, if and only if all $2^\ell$ possible patterns of events have positive likelihood.*

**Proof.** If $\mathcal{T}$ is a star,

$$L(x|\mathcal{T}) = \prod_{\{u|x_u=1\}} \Pr(u) \cdot \prod_{\{u|x_u=0\}} (1 - \Pr(u)) > 0,$$

since $\Pr(u) = \Pr(u|r) \in (0, 1)$ by definition.

If $\mathcal{T}$ is not a star, there is at least one edge $(u, v)$ with $u \neq r$, and any pattern with $x_u = 0$ and $x_v = 1$ has likelihood zero. ■

The lemma implies that for noisy real world data the assumption of a tree topology will generally be too strict in the likelihood sense. Moreover, the estimated mutagenetic tree for the development of zidovudine resistance (Fig. 1) does not capture all of the known pathways. Indeed, the tree topology implies that M41L and T215F/Y can occur only *after* K70R despite the fact that the 215-41 pathway is also observed in the absence of K70R. Consequently, one third of the observed mutational patterns have likelihood zero in the estimated mutagenetic tree. In general, the tree reconstruction is not an ML procedure, and the maximum branchings tend to describe only part of the data satisfyingly.

To overcome these limitations, we consider the broader class of mixture models of mutagenetic trees.

## 4. MIXTURE MODELS

### *4.1. Definition*

Suppose that $Y_1, \ldots, Y_K$ are multivariate discrete random variables with range $\Omega$ that are distributed according to mutagenetic trees

$$\mathcal{T}_k = (V, E_k, r, p_k), \quad k = 1, \ldots, K,$$

respectively. Let $\triangle_1, \ldots, \triangle_K \in \{0, 1\}$ be binary random variables with $\Pr(\triangle_k = 1) = \alpha_k$. We call the model

$$\mathcal{M} = \sum_{k=1}^{K} \alpha_k \mathcal{T}_k \quad \text{with} \quad \alpha_k \in [0, 1] \quad \text{and} \quad \sum_{k=1}^{K} \alpha_k = 1$$

that generates the random variable $Y = \sum_{k=1}^{K} \triangle_k Y_k$, a *K-mutagenetic trees mixture model*.

Thus, the likelihood of a pattern of events $x$ in the mixture model is

$$L(x|\mathcal{M}) = \sum_{k=1}^{K} \alpha_k L(x|\mathcal{T}_k).$$

Throughout, we will consider mixture models that have a special structure in the first mutagenetic tree $\mathcal{T}_1$. We assume that, in addition to different pathways of accumulation of events, there is a certain probability $\beta$ of any event occurring spontaneously independent of all other events. Thus, $\mathcal{T}_1$ is a star with $p(e) = \beta$ for all $e \in E_1$. Tree $\mathcal{T}_1$ can be regarded as the noise component of the model. By Lemma 1, including a star in the mixture model ensures that all patterns of events have positive likelihood.

### *4.2. EM-like learning algorithm*

Given the number of trees $K$, we want to reconstruct a $K$-mutagenetic trees mixture model from observed patterns $X$. This task would be easy, if we knew for each pattern of events from which component of the model it has been generated: We would apply $K$ times the reconstruction technique for a single tree based on the pair probabilities estimated from the respective samples. However, this information is missing and we have to estimate it from the data, too. This procedure results in an algorithm similar to an EM algorithm (Dempster *et al.,* 1977).

Our goal is to find mutagenetic trees $\mathcal{T}_1, \ldots, \mathcal{T}_K$ and mixture parameters $\alpha_1, \ldots, \alpha_k$ that maximize the log-likelihood of the data, which can be written as

$$\sum_{i=1}^{N} \log \sum_{k=1}^{K} \alpha_k L(x_i|\mathcal{T}_k),$$

if the $x_i$ are independent. The *responsibility* of model component $k$ for sample $x_i$ is defined as

$$\gamma_{ik} = \Pr(\triangle_k = 1|\mathcal{M}, x_i).$$

Let $N_k = \sum_{i=1}^{N} \gamma_{ik}$ be the weighted number of samples generated by $\mathcal{T}_k$. In an iterative fashion, we estimate $\gamma$ (E step) and $\mathcal{M}$ (M step) from the data.

Given an estimate of $\mathcal{M} = \sum_{k=1}^{K} \alpha_k \mathcal{T}_k$, we can estimate $\gamma$ by

$$\gamma_{ik} = \frac{\alpha_k L(x_i|\mathcal{T}_k)}{\sum_{m=1}^{K} \alpha_m L(x_i|\mathcal{T}_m)}.$$

Given an estimate of $\gamma$, we update $\mathcal{M}$ as follows. For the noise component ($k = 1$), we let $\mathcal{T}_1$ be a star and estimate $\beta$ as the rate of occurrence of any event in this component,

$$\beta = \frac{1}{\ell N_1} \sum_{j=1}^{\ell} \sum_{i=1}^{N} \gamma_{i1} x_{ij}.$$

For $k \geq 2$, we first estimate all joint probabilities between pairs of events within the $k$-th component:

$$p_k(u, v) = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_{ik} x_{iu} x_{iv}.$$

Next, we reconstruct $\mathcal{T}_k$ from $p_k$ by solving the maximum weight branching problem as described in Section 3.3. Edges with $p_k(u, v) < 0.01$ are previously deleted from the complete graph in order to avoid weekly connected components within one mutagenetic tree. Finally, the mixture parameters are updated by the equation

$$\alpha_k = \frac{N_k}{N} = \frac{1}{N} \sum_{i=1}^{N} \gamma_{ik}.$$

We iterate the E step and the M step until the log-likelihood function does not increase any more.

To run the algorithm, we need initial values for the responsibilities $\gamma_{ik}$. The starting solution can be picked at random, but in general this strategy will yield poor results. The two common approaches to overcome this problem are either to sample many random starting solutions, or to identify a single promising initial solution. To limit computational costs, we decided for the latter approach and perform an ordinary $k$-means clustering with $k = K - 1$ on the set of patterns using squared Euclidean distance as dissimilarity measure[3] (Hastie *et al.*, 2001). From the $k$-means clusters, we derive the initial responsibilities

$$\gamma_{ik} = \begin{cases} \dfrac{1}{2}, & \text{if sample } x_i \text{ belongs to cluster } k-1 \\ \dfrac{1}{2(K-1)}, & \text{else.} \end{cases}$$

[3]The starting solution for the $k$-means algorithm, i.e., the set of initial cluster centers, is a random subset of the data of size $k$. In all experiments, we have chosen the best $k$-means clustering (the one minimizing the within-cluster point scatter) obtained from 100 random starting solutions.

These soft assignments provide a good starting solution, but do not have too strong an effect on the final solution.

The algorithm for learning a $K$-mutagenetic trees mixture model is summarized in Fig. 2. It is differing from a true EM algorithm in the fact that the tree reconstruction step does not provide an ML estimate. Thus, unlike with a true EM algorithm, our EM-like algorithm is not guaranteed to converge to a local maximum of the log-likelihood function. Nevertheless, we have not observed such deviating behavior on any real world dataset so far.

INPUT:

- Patterns of events $X = (x_{ij})_{\substack{1 \le i \le N \\ 1 \le j \le \ell}}$
- Number of mutagenetic trees $K \geq 2$

OUTPUT:

- $K$-mutagenetic trees mixture model $\sum_{k=1}^{K} \alpha_k \mathcal{T}_k$

PROCEDURE:

1. Guess initial responsibilities:
   (a) Run $(K-1)$-means clustering algorithm
   (b) Set responsibilities
   $$\gamma_{ik} = \begin{cases} \frac{1}{2}, & \text{if } x_i \text{ is in cluster } k-1 \\ \frac{1}{2(K-1)}, & \text{else.} \end{cases}$$
2. *M-like step.* Update model parameters:
   Set $N_k = \sum_{i=1}^{N} \gamma_{ik}$ for all $k = 1, \ldots, K$.
   Let $\mathcal{T}_1$ be a star with edge weights
   $$\beta = \frac{1}{\ell N_1} \sum_{j=1}^{\ell} \sum_{i=1}^{N} \gamma_{i1} x_{ij}.$$
   For $k = 2, \ldots, K$:
   (a) For all pairs of events $(u, v)$, $1 \leq u, v \leq \ell$, estimate their joint probabilities
   $$p_k(u, v) = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_{ik} x_{iu} x_{iv}.$$
   (b) Compute the maximum weight branching $\mathcal{T}_k$ from the complete digraph with weights $w$ derived from $p_k$.
   (c) Compute the mixture parameter $\alpha_k = \frac{N_k}{N}$.
3. *E step.* Compute responsibilities:
   $$\gamma_{ik} = \frac{\alpha_k L(x_i | \mathcal{T}_k)}{\sum_{m=1}^{K} \alpha_m L(x_i | \mathcal{T}_m)}.$$
4. Iterate steps 2 and 3 until convergence.

**FIG. 2.** EM-like algorithm for learning a $K$-mutagenetic trees mixture model from data.

## 5. MODEL SELECTION

There is still one model parameter that we do not know, namely, the number of mutagenetic trees $K$ in the mixture model. Our learning algorithm is efficient enough to perform model selection in a cross-validation setting. For the zidovudine resistance data, we have used 10-fold cross-validation in order to estimate the likelihood on unseen data for values of $K$ between 1 and 10. Figure 3 shows the estimated likelihood as a function of $K$. We propose to apply the one-standard-error rule, i.e., to pick the most parsimonious model within one standard error of the maximum mean log-likelihood (Hastie *et al.*, 2001). This yields $K = 3$, and the 3-mutagenetic trees mixture model for the development of zidovudine resistance in the HIV-1 RT is shown in Fig. 4.

This model assigns 19% of the data to the noise component.[4] These data are not necessarily free of any dependencies between events, but within the model class they are best explained by the independence assumption. Forty-seven percent of the data are estimated to be generated by a linear path model that involves the 70-219 pathway followed by 67N and the 215-41 pathway. The remaining 34% of the data are assigned to a mutagenetic tree with initial event 215F/Y followed by either the 70-219 pathway or—with greater probability—the remainder of the 215-41 pathway, namely 41L and 210W.

In conclusion, the mixture model of mutagenetic trees captures all major established facts about the development of zidovudine resistance under zidovudine mono-therapy. Moreover, it provides a quantitative, generative probabilistic model of the accumulation of resistance-associated mutations.

## 6. VALIDATION

After comparing the estimated mixture model with biological knowledge, we now turn to quantitative approaches for model validation. We will derive measures of confidence regarding the mixture model as both a density estimator and a way of learning structural dependencies between events.

### *6.1. Goodness of fit*

We want to quantify how closely a trained mixture model reproduces the empirical probability distribution on $\Omega = 2^{\{1,\dots,\ell\}}$. To compare two discrete probability distributions, we use the following distance measures on the probability vectors $p, q \in [0,1]^{2^\ell}$.

a) Cosine distance: $1 - \cos\angle(p,q) = 1 - \frac{\langle p,q\rangle}{\|p\|_2\,\|q\|_2}$

b) $L_1$ distance: $\|p-q\|_1 = \sum_{i=1}^{2^\ell} |p_i - q_i|$

c) $L_2$ distance: $\|p-q\|_2 = \sqrt{\sum_{i=1}^{2^\ell}(p_i-q_i)^2}$

d) Kullback–Leibler distance (relative entropy): $\mathbb{E}_p\left[\frac{p}{q}\right] = \sum_{i=1}^{2^\ell} p_i \log_2 \frac{p_i}{q_i}$

Using cross-validation we calculate for each partition of the data into training and test set the distances between the distributions induced by the test data and by

(1) the training data,
(2) the 3-mutagenetic trees mixture model estimated from the training data,
(3) the single mutagenetic tree model estimated from the training data,
(4) the null model (a single star model with nonuniform edge weights) estimated from the training data.

Test and training data give rise to empirical distributions obtained from the observed histograms, while the model distributions are computed as described in Sections 2.4 and 3.1. We compare the mixture model

[4] In the sense of the mixture model: in general, each sample is distributed over several components (Section 6.1).

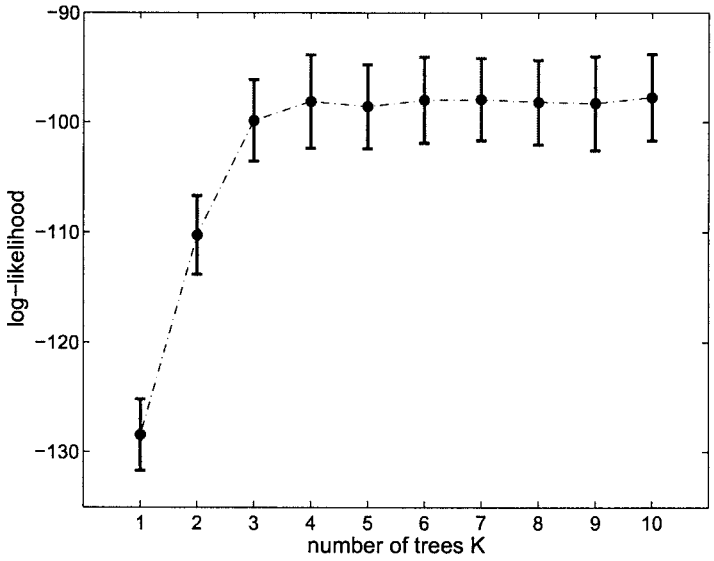


**FIG. 3.** Ten-fold cross-validation log-likelihood estimates for $K$-mutagenetic trees mixture models as a function of $K$.

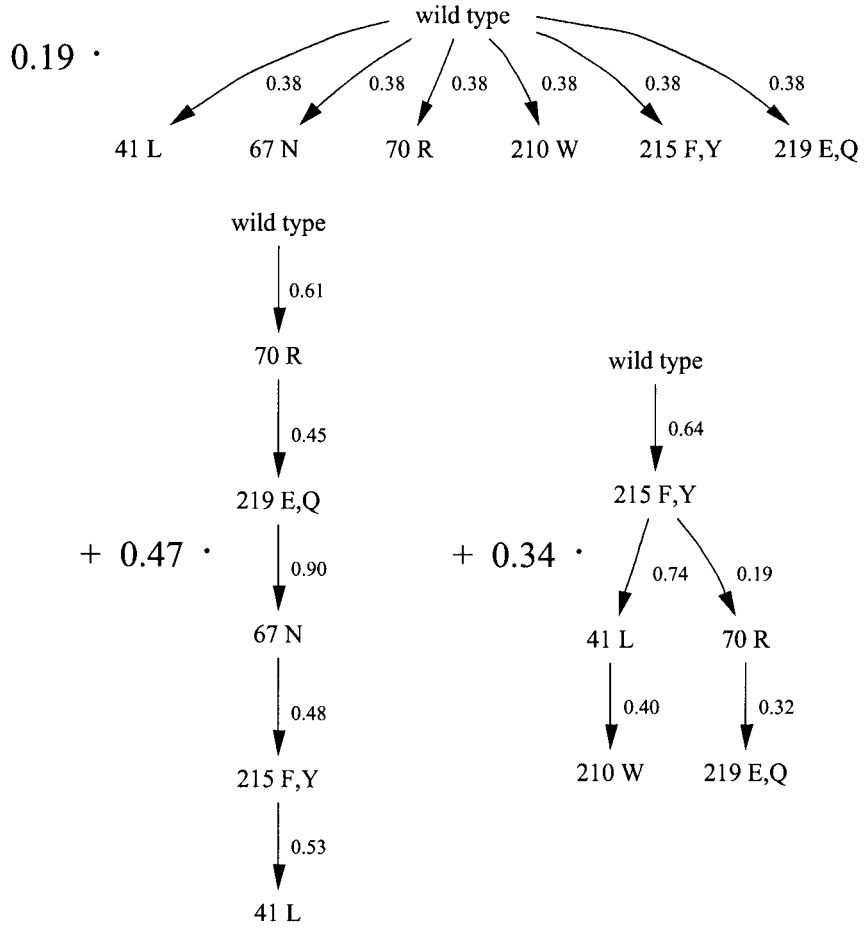


**FIG. 4.** Three-mutagenetic trees mixture model $\mathcal{M} = 0.19\,\mathcal{T}_1 + 0.47\,\mathcal{T}_2 + 0.34\,\mathcal{T}_3$ for the development of zidovudine resistance. Each tree is preceded by its weight $\alpha_k$. The upper tree $\mathcal{T}_1$ is a star and represents the noise component. In the second tree, we have omitted the connected component consisting of the single vertex with label 210W, and similarly 67N in the third tree.

with a single mutagenetic tree model and with a star model representing the null hypothesis of independence of events. Distance (1) measures only the effect of finite sampling, whereas the distances (2)–(4) include losses that are due to imperfect model assumptions and/or parameter estimates.

For the Kullback–Leibler distance, we replace probabilities $\hat{q}_i$ that are estimated to be zero by the value $1/(2n)$, where $n$ is the fraction of samples used for estimating the distribution (either empirically or by training a model). Thus, we effectively use a pseudocount of $1/2$, a common strategy in estimating the Kullback–Leibler distance.

Figure 5 shows the distributions of all distances for 100 runs of 10-fold cross-validation each. For all distance measures, the mixture model distribution closely resembles the empirical test data distribution. In contrast, both the single tree model and the null model provide inferior density fits. The fact that the single tree misses an entire mutational pathway becomes most evident in the $L_2$ measure. Frequently observed patterns from the unconsidered pathway have likelihood zero in this tree and give rise to large (quadratic) contributions to the $L_2$ norm.

In the same cross-validation runs, we have determined the percentage of samples that remain unexplained by the nontrivial components of the mixture model. The mean percentage of samples with likelihood zero in all but the noise component was 13%. Thus, the mixture model maps 87% of the observed patterns onto the other identified mutagenetic trees. For the optimal model on the full data (Fig. 4), it happens to be the case that the only pattern that can be generated by both nontrivial trees is the null pattern. We report in Table 1 the distribution of samples among the trees in detail.

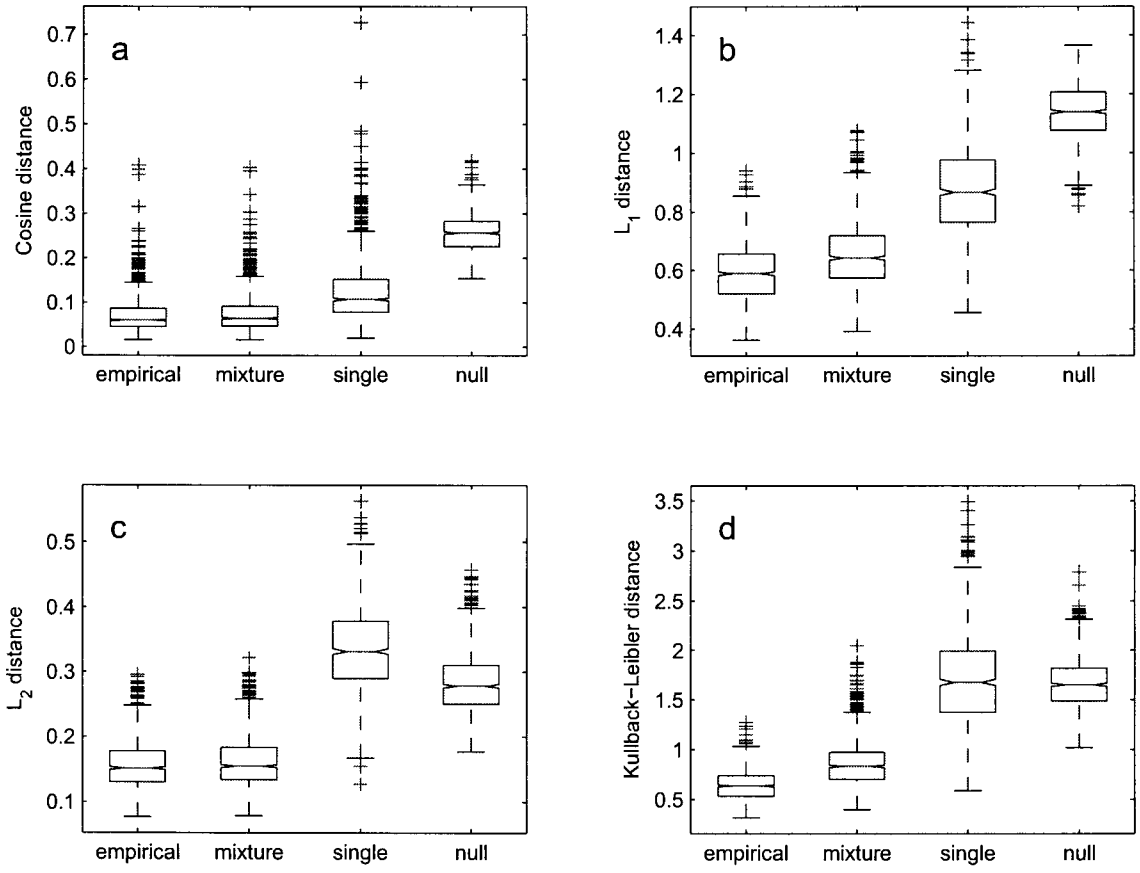


**FIG. 5.** Box-plot of distribution distances obtained from 100 runs of 10-fold cross-validation: (**a**) cosine distance, (**b**) $L_1$ distance, (**c**) $L_2$ distance, (**d**) Kullback–Leibler distance. Each subfigure shows from left to right: distances between distributions of the test data and the training data (empirical, first column), the mixture model (mixture, second column), the single tree model (single, third column), the null model (null, fourth column). The null model refers to the independence assumption of events and is a single star with nonuniform edge weights.

TABLE 1. DISTRIBUTION OF SAMPLES AMONG THE COMPONENTS OF THE MIXTURE MODEL SHOWN IN FIG. 4[a]

| *Likelihood* | | | *Sample subset* | |
|---|---|---|---|---|
| $L(x\|\mathcal{T}_1)$ | $L(x\|\mathcal{T}_2)$ | $L(x\|\mathcal{T}_3)$ | *Fraction* | *Description* |
| $> 0$ | $> 0$ | $> 0$ | 31.6% | Null patterns |
| $> 0$ | $> 0$ | $= 0$ | 30.2% | 70-219 pathway |
| $> 0$ | $= 0$ | $> 0$ | 25.0% | 215-41 pathway |
| $> 0$ | $= 0$ | $= 0$ | 13.2% | "Noise" |

[a]The fraction of samples refers to the full dataset of 364 samples.

### 6.2. *Tree stability*

We have already interpreted the topology of the mutagenetic trees in detail. Here we use the bootstrap (Efron and Tibshirani, 1998) in order to obtain an estimate of the dependence of the topology on sampling effects. Since there is no standard way of comparing two mixture models, we confine ourselves to analyzing the stability of each single mutagenetic tree. For a mixture model, we fix the responsibilities $\gamma$ obtained from the EM-algorithm. For tree $\mathcal{T}_k$, we resample with replacement each pattern of events $x_i$ with probability $\gamma_{ik}$. From the bootstrap sample of size $N$, a mutagenetic tree is reconstructed. As a test statistic, we consider the relative count of each edge $e \in E_k$ in the bootstrap trees.

In Fig. 6, we report for the zidovudine data these estimates based on 10,000 bootstrap samples. The two mutagenetic trees are the two nontrivial components of the mixture model displayed in Fig. 4. We find strong support for 70R as initial event and for the dependencies 215F/Y $\rightarrow$ 41L $\rightarrow$ 210W and 70R $\rightarrow$ 219E/Q $\rightarrow$ 67N. The latter edge suggests that mutation D67N may be best explained as a late event in the 70-219 pathway.

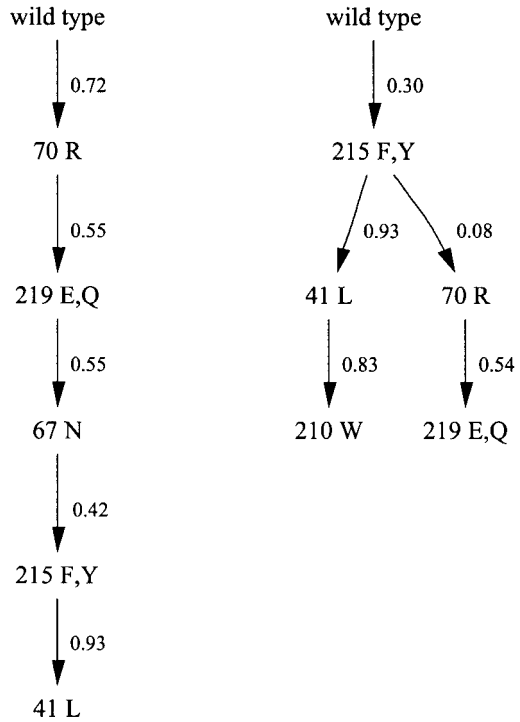


**FIG. 6.** Bootstrap analysis of tree stability. Edge weights represent relative counts in 10,000 bootstrap samples. The two mutagenetic trees are the two nontrivial components of the mixture model in Fig. 4.

## 7. DISCUSSION

We have presented mixture models of mutagenetic trees for modeling evolutionary processes that can be described as an accumulation of permanent genetic changes along multiple pathways. Application to the development of zidovudine resistance in the HIV-1 RT has shown that, compared to single mutagenetic trees, the class of mixture models provides both better density estimation of observed patterns of events and biologically more plausible models.

It will be interesting to compare mutagenetic trees mixture models to other statistical models, in particular to the mixture model of undirected trees (Meilă and Jordan, 2000) and to the ML tree with hidden variables (von Heydebreck *et al.*, 2004). Mutagenetic trees are particularly useful for estimating mutational pathways, because the directed edges induce an order on the set of mutational patterns. Thus, they may serve for generating hypotheses about evolutionary processes. The software used in this paper is freely available for noncommercial purposes as the "mtreemix" package at *www.mtreemix.bioinf.mpi-sb.mpg.de/*. The zidovudine dataset used for evaluation can also be downloaded from this site.

### 7.1. Applications

Further applications of the model include the accumulation of mutations associated with resistance to other antiretroviral drugs in protease and RT, the current drug targets of HAART. Together, these models may be helpful in designing treatment protocols that avoid the accumulation of cross-resistance conferring mutations so that clinicians do not run out of therapeutic options too early. However, since monotherapies are obsolete, mutational pathways under combination therapy must be identified. It remains to be investigated whether combination therapy pathways are a function of the pathways for the single drugs forming the combination. If such a relationship cannot be found, we may face another data scarceness problem induced by the large number of possible drug combinations.

The generative probabilistic tree models can be used for extensive simulations of sequence populations under the selective pressure of different drugs within or among hosts. Indeed, drawing a random sample from a $K$-mutagenetic trees mixture model $\mathcal{M} = \sum_{k=1}^{K} \alpha_k \mathcal{T}_k$ is straightforward. We first draw a uniform random number and decide according to the mixture parameters $\alpha = (\alpha_1, \ldots, \alpha_K)$ which mutagenetic tree to use. In the selected tree, we draw each edge $e \in E$ independently with probability $p(e)$. The sample consists of all events that are reachable from $r$ in the induced subgraph. Such sequence simulations are useful for studying the effectiveness of different drug sequencing strategies and thus for the *in silico* design of optimal drug use patterns (Phillips *et al.*, 2003).

As density estimators, the mixture models can also be used for classification. A common classification problem in the context of HIV drug resistance is to separate susceptible from resistant strains. With mixture models $\mathcal{M}_{sus}$ trained on the susceptible and $\mathcal{M}_{res}$ trained on the resistant subset, we consider the likelihood ratio

$$\frac{L(x|\mathcal{M}_{sus})}{L(x|\mathcal{M}_{res})}$$

to decide whether a pattern $x$ is more likely to originate from the susceptible or the resistant subpopulation. Analysis of the model $\mathcal{M}_{res}$ may reveal pathways leading to resistance independent of the applied drug pressure, including cross-resistance pathways induced by other drugs.

### 7.2. Model-based clustering

The classical EM algorithm for learning Gaussian mixture models can be regarded as a soft version of $K$-means clustering (Hastie *et al.*, 2001). For each model component, the EM algorithm assigns a responsibility to each sample, whereas the $K$-means clustering algorithm assigns each sample to exactly one of the $K$ clusters. Likewise, we can easily modify our $K$-mutagenetic trees mixture model learning algorithm (Fig. 2) to obtain a clustering algorithm. It suffices to store cluster assignments instead of responsibilities in the matrix $\gamma$:

$$\gamma_{ik} = \begin{cases} 1, & \text{if sample } x_i \text{ is in cluster } k \\ 0, & \text{else,} \end{cases}$$

and change the E step to

$$\gamma_{ik} = \begin{cases} 1, & \text{if } k = \arg\max_{1 \le m \le K} L(x_i | \mathcal{T}_m) \\ 0, & \text{else.} \end{cases}$$

This model-based clustering is useful in situations where pathways are known or suspected to be mutually exclusive.

### *7.3. Full ML estimation*

We would arrive at a true EM algorithm if we estimated the mutagenetic trees $\mathcal{T}_k$ (Fig. 2, step 2(b)) in an ML fashion. However, finding the ML topology of a mutagenetic tree appears difficult in absence of a construction rule and in view of the large number of possible trees. Thus, heuristic search methods need to be applied, such as those used in ML phylogeny estimation (Felsenstein, 1981).

## ACKNOWLEDGMENTS

Part of this research has been funded by Deutsche Forschungsgemeinschaft under Grant No. HO 1582/1-3 (N.B.) and by BMBF Grant No. 031U117 (J.R.). We are grateful to Robert Shafer for his assistance in preparing the zidovudine dataset. This work has been performed in the context of the BioSapiens Network of Excellence (EU contract no. LSHG-CT-2003-503265).

## REFERENCES

Beerenwinkel, N., *et al.* 2003. Methods for optimizing antiviral combination therapies. *Proc. 11th Int. Conf. on Intellig. Syst. for Mol. Biol. (ISMB '03), Bioinformatics* 19, i16–i25.

Beerenwinkel, N., *et al.* 2001a. Geno2pheno: Interpreting genotypic HIV drug resistance tests. *IEEE Intellig. Syst.* 16, 35–41.

Beerenwinkel, N., *et al.* 2001b. Identifying drug resistance-associated patterns in HIV genotypes. *Proc. German Conf. on Bioinformatics*, Braunschweig, 126–130.

Beerenwinkel, N., *et al.* 2002. Diversity and complexity of HIV-1 drug resistance: A bioinformatics approach to predicting phenotype from genotype. *Proc. Natl. Acad. Sci. USA* 99(12), 8271–8276.

Boucher, C., *et al.* 1992. Ordered appearance of zidovudine resistance mutations during treatment of 18 human immunodeficiency virus-positive subjects. *J. Infect. Dis.* 165, 105–110.

Chu, Y., and Liu, T. 1965. On the shortest arborescense of a directed graph. *Sci. Sinica* 14, 1396–1400.

Chow, C., and Liu, C. 1968. Approximating discrete probability distributions with dependence trees. *IEEE Trans. Inform. Theory* 14(3), 462–467.

Dempster, A., *et al.* 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussions). *J. R. Statist. Soc. B* 39, 1–38.

Desper, R., *et al.* 1999. Inferring tree models for oncogenesis from comparative genome hybridization data. *J. Comp. Biol.* 6(1), 37–51.

Desper, R., *et al.* 2000. Distance-based reconstruction of tree models for oncogenesis. *J. Comp. Biol.* 7(6), 789–803.

Edmonds, J. 1967. Optimum branchings. *J. Res. Nat. Bur. Stand.* 71B, 233–240.

Efron, B., and Tibshirani, R. 1998. *An Introduction to the Bootstrap*, Monographs on Statistics and Applied Probability, Chapman and Hall/CRC, London.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences. *J. Mol. Evol.* 17, 368–376.

Friedman, N., *et al.* 1997. Bayesian network classifiers. *Machine Learning* 29(2–3), 131–163.

Gonzales, M., *et al.* 2003. Extended spectrum of HIV-1 reverse transcriptase mutations in patients receiving multiple nucleoside analog inhibitors. *AIDS* 17, 791–799.

Hall, B. 2002. Predicting evolution by *in vitro* evolution requires determining evolutionary pathways. *Antimicrob. Agents and Chemother.* 46(9), 3035–3038.

Hastie, T., *et al.* 2001. *The Elements of Statistical Learning*, Springer, New York.

Heckerman, D., *et al.* 1995. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning* 20, 197–243.

Jeeninga, R., *et al.* 2001. Evolution of AZT resistance in HIV-1: The 41-70 intermediate that is not observed *in vivo* has a replication defect. *Virology* 283, 294–305.

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

Keulen, W., *et al.* 1996. Nucleotide substitution patterns can predict the requirements for drug-resistance of HIV-1 proteins. *Antiviral Res.* 31, 45–57.

Larder, B. 1994. Interactions between drug resistance mutations in human immunodeficiency virus type 1 reverse transcriptase. *J. Gen. Virol.* 75, 951–957.

Larder, B., *et al.* 1989. HIV with reduced sensitivity to zidovudine (AZT) isolated during prolonged therapy. *Science* 243(4899), 1731–1734.

Larder, B., and Kemp, S. 1989. Multiple mutations in HIV-1 reverse transcriptase confer high-level resistance to zidovudine (AZT). *Science* 246(4934), 1155–1158.

Lathrop, R., and Pazzani, M. 1999. Combinatorial optimization in rapidly mutating drug-resistant viruses. *J. Comb. Opt.* 3, 301–320.

Meilă, M., and Jordan, M. 2000. Learning with mixtures of trees. *J. Machine Learning Res.* 1, 1–48.

Molla, A., *et al.* 1996. Ordered accumulation of mutations in HIV protease confers resistance to ritonavir. *Nat. Med.* 2(7), 760–766.

Perrin, L., and Telenti, A. 1998. HIV treatment failure: Testing for HIV resistance in clinical practice. *Science* 280, 1871–1873.

Phillips, A.N., *et al.* 2003. Use of a stochastic model to develop understanding of the impact of different patterns of antiretroviral drug use on resistance development. *AIDS* 17, 1009–1016.

Radmacher, M., *et al.* 2001. Graph models of oncogenesis with an application to melanoma. *J. Theor. Biol.* 212, 535–548.

Rhee, S.-Y., *et al.* 2003. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucl. Acids Res.* 31(1), 298–303.

Sevin, A., *et al.* 2000. Methods for investigation of the relationship between drug-susceptibility phenotype and human immunodeficiency virus type 1 genotype with applications to AIDS clinical trials group 333. *J. Infect. Dis.* 182, 59–67.

Shafer, R., *et al.* 2000. The genetic basis of HIV-1 resistance to reverse transcriptase and protease inhibitors. *AIDS Rev.* 2, 211–228.

Simon, R., *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.

Tarjan, R. 1977. Finding optimum branchings. *Networks* 7, 25–35.

Vandamme, A., *et al.* 1999. Managing resistance to anti HIV drugs: An important consideration for effective disease management. *Drugs* 57, 337–361.

von Heydebreck, A., *et al.* 2004. Maximum likelihood estimation of oncogenetic tree models. *Biostatistics* 5(4), 545–556.

Wu, T., *et al.* 2003. Mutation patterns and structural correlates in human immunodeficiency virus type 1 protease following different protease inhibitor treatments. *J. Virol.* 77(8), 4836–4847.

Address correspondence to:
*Niko Beerenwinkel*
*Department of Mathematics*
*University of California at Berkeley*
*898 Evans Hall*
*Berkeley, CA 94720-3840*

*E-mail:* niko@math.berkeley.edu