Biostatistics (2015), **16**, 4, *pp.* 713–726
doi:10.1093/biostatistics/kxv019
Advance Access publication on May 14, 2015

Estimating the dynamics and dependencies of accumulating mutations with applications to HIV drug resistance

HESAM MONTAZERI

Department of Biosystems Science and Engineering, ETH Zurich, Basel 4058, Switzerland and SIB Swiss Institute of Bioinformatics, Basel 4058, Switzerland

HULDRYCH F. GÜNTHARD, WAN-LIN YANG, ROGER KOUYOS

Division of Infectious Diseases and Hospital Epidemiology, University Hospital Zurich, University of Zurich, Zurich 8091, Switzerland
Institute of Medical Virology, University of Zurich, Zurich 8057, Switzerland

NIKO BEERENWINKEL*, THE SWISS HIV COHORT STUDY

Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland and SIB Swiss Institute of Bioinformatics, Basel, Switzerland

niko.beerenwinkel@bsse.ethz.ch

SUMMARY

We introduce a new model called the observed time conjunctive Bayesian network (OT-CBN) that describes the accumulation of genetic events (mutations) under partial temporal ordering constraints. Unlike other CBN models, the OT-CBN model uses sampling time points of genotypes in addition to genotypes themselves to estimate model parameters. We developed an expectation–maximization algorithm to obtain approximate maximum likelihood estimates by accounting for this additional information. In a simulation study, we show that the OT-CBN model outperforms the continuous time CBN (CT-CBN) (Beerenwinkel and Sullivant, 2009. Markov models for accumulating mutations. *Biometrika* **96**(3), 645–661), which does not take into account individual sampling times for parameter estimation. We also show superiority of the OT-CBN model on several datasets of HIV drug resistance mutations extracted from the Swiss HIV Cohort Study database.

Keywords: Conjunctive Bayesian networks; Expectation–maximization algorithm; Genetic progression; HIV drug resistance; Maximum likelihood estimation.

1. INTRODUCTION

HIV drug resistance development is a consequence of viral evolution. This evolutionary process is characterized mainly by the accumulation of resistance mutations, i.e. mutations that confer a selective advantage

*To whom correspondence should be addressed.

under the selective pressure of antiviral drugs. Models of HIV viral evolution have been shown to improve the prediction of therapy response (Beerenwinkel, Sing *and others*, 2005; Deforche *and others*, 2008; Altmann *and others*, 2009; Prosperi *and others*, 2009; Beerenwinkel *and others*, 2013). Models of the accumulation of beneficial mutations have also been used successfully in other biological applications such as the somatic evolution of cancer (Rahnenführer *and others*, 2005; Gerstung *and others*, 2009). Different statistical models have been proposed for modeling the accumulation of mutations such as mutagenetic trees (Desper *and others*, 1999), mutagenetic trees with hidden inner nodes (von Heydebreck *and others*, 2004), mixture models of mutagenetic trees (Beerenwinkel, Rahnenführer *and others*, 2005), and conjunctive Bayesian networks (CBNs) (Beerenwinkel *and others*, 2006). In these models, the order in which mutations accumulate are subject to some constraints. Consequently, evolution follows only a subset of all possible mutational pathways from the wild type, the genotype carrying no mutation, to the fully resistant genotype, the genotype carrying all resistance mutations. It has been suggested that only very few mutational paths are needed to explain the molecular evolution of drug resistance (Weinreich *and others*, 2006; Poelwijk *and others*, 2007; Lozovsky *and others*, 2009).

CBNs are specialized Bayesian networks defined by a partially ordered set (poset) of mutations (Beerenwinkel *and others*, 2006). The partial order specifies the ordering constraints in which mutations can happen. In contrast to general Bayesian networks, in CBNs, a mutation can only occur if all its predecessor mutations have already occurred. The number of free parameters in the model is equal to the number of mutations, in contrast to general Bayesian networks, where the number of parameters can grow exponentially in the number of mutations. Hence, CBNs generally do not suffer from non-identifiability problems and there exist computationally efficient inference algorithms for these models (Beerenwinkel and Sullivant, 2009; Gerstung *and others*, 2009; Sakoparnig and Beerenwinkel, 2012). CBNs can be discrete or continuous in time. The continuous time CBN (CT-CBN) represents a waiting time process for the accumulation of mutations (Beerenwinkel and Sullivant, 2009). In this model, the occurrence time of each mutation is assumed to be exponentially distributed with a specific rate of evolution for each mutation. The waiting time process of a mutation starts only when all its predecessor mutations have already occurred.

Given that the occurrence times of mutations are available for each observation, the maximum likelihood (ML) estimates of the evolutionary rates and of the poset structure of the CT-CBN model are given in Beerenwinkel and Sullivant (2009). However, in practice, occurrence times of individual mutations are not observed, but usually we can only measure which mutations have occurred before a certain sampling time t_s. In some applications, sampling times themselves are difficult to measure. For instance, in cancer progression, due to the fact that the start of the tumor evolutionary process is not known, measurement of the relative sampling time is generally impossible. For such applications, the CT-CBN model assumes the unknown sampling times are themselves random and drawn from an independent exponential distribution, $T_s \sim \exp(\lambda_s)$. Omitting sampling times of individual genotypes in the estimation process results in less accurate estimates of evolutionary rates and of poset structures, especially when the sampling times are not exponentially distributed or not independent of the mutational process.

In other applications, the sampling times for individual genotypes are observable. For instance, the evolutionary escape process of HIV starts from the onset of therapy and the HIV genome is typically determined after virological failure, i.e. after the total viral load exceeding a certain threshold. Hence, the sampling time is simply the time from the start of the therapy to genotyping. These data are available in the Swiss HIV Cohort Study (SHCS) database, a large observational cohort in Switzerland with integrated genotypic drug resistance test results (von Wyl *and others*, 2007; Schoeni-Affolter *and others*, 2010). Here, we propose a new model called observed time CBN (OT-CBN) which performs parameter estimation not only based on observed genotypes but also on the corresponding sampling times. In particular, we develop an expectation–maximization (EM) algorithm (Dempster *and others*, 1977) for approximating ML estimates. The EM algorithm takes into account this additional temporal information and performs estimation of evolutionary rates using the input data $D = (g_i, t_{s,i})_{i=1,\ldots,N}$ where genotype g_i is observed

at sampling time $t_{s,i}$ for the ith observation and N is the number of observations. Simulation studies and application to several HIV datasets from the SHCS show that using the sampling times significantly improves accuracy of parameter estimation and model selection.

This paper is organized as follows. In Section 2, we formally introduce the OT-CBN. In Section 2, we present the EM algorithm for approximating the ML estimates of the model. Section 3 reports performance measures of the OT-CBN and CT-CBN models in different simulation settings. In Section 4, the applications of the OT-CBN model to several datasets of HIV drug resistance mutations are presented. Finally, we close with conclusions in Section 5.

2. Observed time conjuctive bayesian networks

In this section, we introduce the OT-CBN and derive some of its properties. The OT-CBN model, like other CBN models, is a non-ergodic continuous time Markov chain model on the distributive lattice of a partially ordered set (poset) of events. In our model, the poset P is a set of genetic events (i.e. mutations) with ground set $[n] = \{1, \ldots, n\}$ and a transitive relation $\preceq$ which specifies constraints on the order in which events can occur. We use the terms "event" and "mutation" interchangeably in this paper. The relation $i \prec j$ implies that event i must happen before event j. The state space of the OT-CBN model is $J(P) \times \mathcal{R}_+$. Each observation consists of a genotype $g \in J(P)$ and its sampling time $t \in \mathcal{R}_+$ where $J(P)$ is the distributive lattice of order ideals of P. An order ideal is a subset of events $g \subseteq P$ for which $i \in g$ whenever $i \prec j$ and $j \in g$. The order ideals in $J(P)$ correspond to the genotypes (or mutational patterns, or subsets of mutations) compatible with the poset P. The wild type is denoted by $\emptyset \in J(P)$ and is defined as the genotype carrying no mutation. For each event $i \in P$, we define an exponentially distributed random variable $Z_i \sim \text{Exp}(\lambda_i)$ and a random variable T_i as

$$T_i = \max_{j \in \text{pa}(i)} T_j + Z_i, \tag{2.1}$$

where $\text{pa}(i)$ is the set of parents of event i in the poset P. The random variable T_i represents the occurrence time of event i by assuming that no events have been observed at time zero. The random variable T_i is not observed. The density function of the random vector $T = (T_1, \ldots, T_n)$ is

$$f_{P,\Lambda}(t) = \prod_{i=1}^{n} f_{\lambda_i}\left(t_i - \max_{j \in \text{pa}(i)} t_j\right), \tag{2.2}$$

where $\Lambda_n = (\lambda_1, \ldots, \lambda_n) \in \mathcal{R}_+^n$, and the density function f_{λ_i} is the univariate exponential probability density function. The density function $f_{P,\Lambda}$ is zero if there exists an event $i \in P$ such that $t_i < \max_{j \in pa(i)} t_j$. A genotype with occurrence times t is said to be *compatible* with the poset P if $f_{P,\Lambda}(t) > 0$, or equivalently if $t_i \geqslant \max_{j \in \text{pa}(i)} t_j$ for all $i \in P$, i.e. if no mutation occurred before any of its predecessor mutations in the poset.

Example 2.1 As a running example, we consider the poset P defined on the set $\{1, 2, 3, 4\}$ with the relations $1 \prec 3, 2 \prec 4, 2 \prec 3$, displayed in Figure 1(a). The corresponding genotype lattice $J(P)$ consists of eight genotypes compatible with the poset (Figure 1(b)). The random variables T_i and Z_i are defined as $Z_i \sim \text{Exp}(\lambda_i),\ i = 1, \ldots, 4$, $T_1 = Z_1,\ T_2 = Z_2$, $T_3 = Z_3 + \max(T_1, T_2)$, and $T_4 = Z_4 + T_2$. For $t_1, t_2 \geqslant 0,\ t_3 \geqslant \max(t_1, t_2),\ t_4 \geqslant t_2$, the joint exponential probability density function is given by

$$f(t) = \lambda_1\lambda_2\lambda_3\lambda_4 \exp(-\lambda_1 t_1) \exp(-\lambda_2 t_2) \exp[-\lambda_3\{t_3 - \max(t_1, t_2)\}] \exp\{-\lambda_4(t_4 - t_2)\},$$

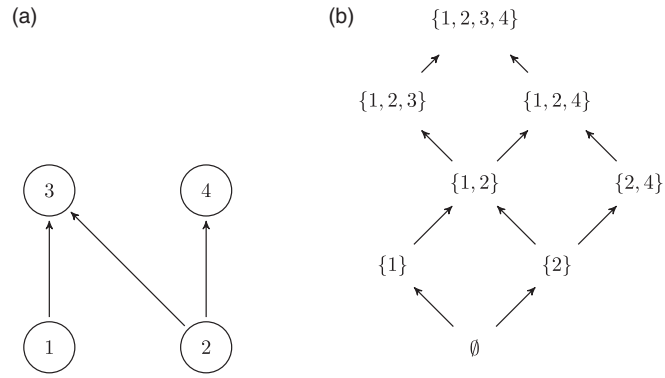

Fig. 1. The poset P, consisting of four elements subject to the relations $1 \prec 3$, $2 \prec 3$, and $2 \prec 4$, is shown in (a). Each vertex labeled $i \in P$ represents the random variable T_i that describes the waiting time for event i. The corresponding genotype lattice $J(P)$, consisting of eight genotypes compatible with the poset P, is shown in (b).

and otherwise the density is zero. Figure 1(a) and 1(b) of supplementary material available at *Biostatistics* online illustrate two realizations of the random variables T_i and Z_i for the chains $1 \prec 2 \prec 3 \prec s \prec 4$ and $2 \prec 1 \prec s \prec 4 \prec 3$, respectively, where s is the sampling event.

Let $g \in J(P)$ be a genotype. The set $\mathrm{Exit}(g) = \{j \in P \mid j \notin g, g \cup j \in J(P)\}$ is the set of events that are not in g but that can happen next. We define for any subset $A \subseteq P$, $\lambda_A = \sum_{j \in A} \lambda_j$, and $T_A = \{T_j \mid j \in A\}$. The complement of a genotype g is denoted by $\bar{g} = P \setminus g$. We define the poset Q_g by adding a new event s for the sampling time t_s to the poset P, i.e. $Q_g = P \cup s$. The poset Q_g has extra relations imposed by the observed genotype g. In addition to the relations in the poset P, it consists of the relations $i \prec s$ for all $i \in g$ and $s \prec i$ for all $i \notin g$. The poset Q_g is said to refine the poset P by the genotype g. Figure 2(a) and (c) of supplementary material available at *Biostatistics* online show the refinements of the poset P, defined in Example 2.1, for the genotypes {1, 2} and {1, 2, 3}, respectively.

A chain in $J(P)$ is a collection of subsets $C_0, \ldots, C_k \in J(P)$ with $C_i \subsetneqq C_{i+1}$ for all i. A chain with maximum length is called a maximal chain. Maximal chains in $J(P)$ have length $n + 1$ with $n = |P|$. We denote by $\mathcal{C}(J(P))$ the set of all maximal chains in $J(P)$. For any maximal chain, $C = (C_0, \ldots, C_n)$, we have $C_0 = \emptyset$ and $C_n = P$. For every maximal chain C in $J(Q_g)$, in addition it holds that $C_{|g|} = g$. Equivalently, a chain is denoted by the sequence of corresponding events, $C_1 \backslash C_0 \to C_2 \backslash C_1 \to \cdots \to C_n \backslash C_{n-1}$.

EXAMPLE 2.2 Figure 2(b) and (d) of supplementary material available at *Biostatistics* online show the genotype lattices refined by the genotypes {1, 2} and {1, 2, 3} for the poset P of Example 2.1, respectively. For the genotype lattice $J(P)$, there are five maximal chains $1 \to 2 \to 3 \to 4$, $1 \to 2 \to 4 \to 3$, $2 \to 1 \to 3 \to 4$, $2 \to 1 \to 4 \to 3$, and $2 \to 4 \to 1 \to 3$, while the refined genotype lattice $J(Q_{\{1,2,3\}})$ consists of only two maximal chains $1 \to 2 \to 3 \to 4$ and $2 \to 1 \to 3 \to 4$. The chain $2 \to 1 \to 3 \to 4$ is denoted alternatively by $C = (\emptyset, \{2\}, \{2, 1\}, \{2, 1, 3\}, \{2, 1, 3, 4\})$.

Let T be a random vector of mutation occurrence times and g be an observed genotype at sampling time t_s. The random variable T is said to be compatible with a poset P, denoted by $T \vdash P$, if $T_i \leqslant T_j$ for every $i \prec j$ in P. We use the notation $(T, t_s) \vdash Q_g$ when the random vector (T, t_s) is compatible with the refined poset Q_g.

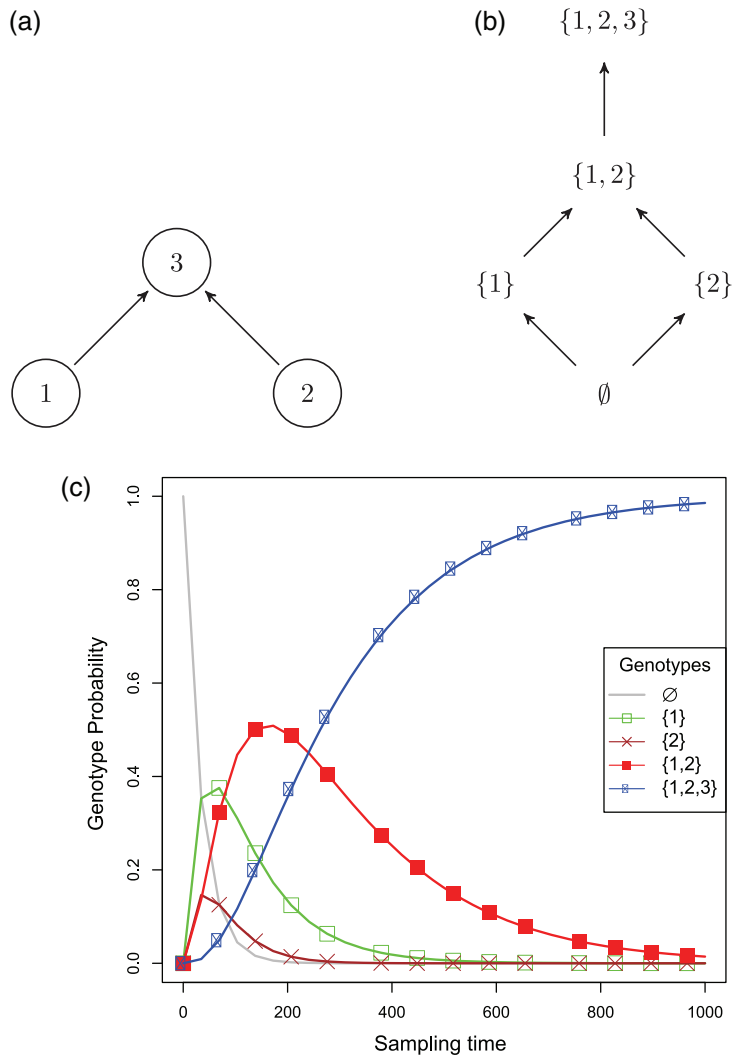

Fig. 2. The poset with three events and the relations $1 \prec 3$ and $2 \prec 3$ (Example 2.4) is shown in (a). The corresponding genotype lattice is shown in (b). Genotype probabilities for all compatible genotypes versus sampling time are shown in (c). Theorem 2.3 is used to compute the genotype probabilities.

THEOREM 2.3 Let the random vector T, as defined in (2.1), be compatible with the poset P, i.e. $T \vdash P$. Then the probability that T is compatible with the refined poset Q_g at sampling time t_s is

$$\text{Prob}\{(T, t_s) \vdash Q_g\} = \prod_{i=1}^{n} \lambda_i \sum_{C \in \mathcal{C}(J(Q_g))} \Phi_s\{(\lambda_{\text{Exit}(C_0)}, \ldots, \lambda_{\text{Exit}(C_{n-1})}), l\},$$

where $l = |g|$ is the number of mutations in genotype g and

$$\Phi_s\{(\lambda_1, \ldots, \lambda_n), l\} = \int_{u_1=0}^{t_s} \int_{u_2=0}^{t_s-u_1} \cdots \int_{u_l=0}^{t_s-(u_1+\cdots+u_{l-1})} \int_{u_{l+1}=t_s-(u_1+\cdots+u_l)}^{\infty} \cdots \int_{u_n=0}^{\infty} \prod_{i=1}^{n} \exp(-\lambda_i u_i)\, \mathrm{d}u. \tag{2.3}$$

Proof. The proof is given in supplementary material available at *Biostatistics* online. □

EXAMPLE 2.4 Let P be the poset defined on the ground set $\{1, 2, 3\}$ with the relations $1 \prec 2$ and $1 \prec 3$ (Figure 2(a)). The possible genotypes are $\emptyset$, $\{1\}$, $\{2\}$, $\{1, 2\}$, $\{1, 2, 3\}$. The corresponding genotype lattice of this poset, $J(P)$, is displayed in Figure 2(b). The probabilities of observing different genotypes are shown in Figure 2(c). Theorem 2.3 is used to compute these probabilities. The wild type is the dominant genotype in the very beginning of the evolutionary process (small sampling times). For higher sampling times, the fully resistant genotype becomes dominant.

The estimation of the OT-CBN model consists of two parts: learning the ML poset and estimating the ML rate parameters for a given poset. We will discuss the poset learning in Section 4. In the rest of this section, we present a method to perform parameter estimation for a given poset using observed genotypes and sampling times. Since occurrence time for each mutation, t_i, is a latent variable in the joint density function (2.2), it is not possible to compute the density (2.2) from observed data. Hence, we use the EM algorithm to find approximate ML estimates of the rate parameters λ. In the *E-step* of the EM algorithm, for each mutation $i \in P$, we need to compute the expected value, $\mathrm{E}[T_i - \max_{j\in \mathrm{pa}(i)} T_j | (T, t_s) \vdash Q_g]$, for each pair of observed genotype and sampling time, (g, t_s). These expected values are the expected sufficient statistics for λ, therefore in the *M-step* of the EM algorithm, the new estimate of λ_i is $N/\sum_{k=1}^{N} \mathrm{E}[T_i - \max_{j\in \mathrm{pa}(i)} T_j \mid (T, t_{s,k}) \vdash Q_{g_k}]$ where N is the number of observations. Theorem 2.5 can be used to compute this expectation. Similar to the technique used in Theorem 2.3, the integration required for computing the expectation is decomposed into sum of integrals over simpler regions. Each of these integrals is then computed by the recursive formula given in Proposition 2 of supplementary material available at *Biostatistics* online, Appendix C.

THEOREM 2.5 The expected value of $T_i - \max_{j\in \mathrm{pa}(i)} T_j$ given $(T, t_s) \vdash Q_g$ is

$$\mathrm{E}\left[T_i - \max_{j\in \mathrm{pa}(i)} T_j \,\middle|\, (T, t_s) \vdash Q_g \right] = \frac{\prod_{i=1}^{n} \lambda_i}{\mathrm{Prob}\{(T, t_s) \vdash Q_g\}} \sum_{C\in \mathcal{C}(J(Q_g))} \sum_{k=1}^{n} \iota(i, C_{k-1}) \zeta_s\{(\lambda_{\mathrm{Exit}(C_0)}, \ldots, \lambda_{\mathrm{Exit}(C_{n-1})}), k, l\},$$

where $\iota(i, \mathcal{C}_k) = 1$ if $i \notin \mathcal{C}_k$ and $\mathrm{pa}(i) \subseteq \mathcal{C}_k$, and 0 otherwise, and the function ζ_s is defined as

$$\zeta_s\{(\lambda_1, \ldots, \lambda_n\}, k, l) = \int_{u_1=0}^{t_s} \int_{u_2=0}^{t_s-u_1} \cdots \int_{u_l=0}^{t_s-(u_1+\cdots+u_{l-1})} \int_{u_{l+1}=t_s-(u_1+\cdots+u_l)}^{\infty} \cdots \int_{u_n=0}^{\infty} u_k \prod_{i=1}^{n} \exp(-\lambda_i u_i)\, \mathrm{d}u. \tag{2.4}$$

Proof. The proof is given in supplementary material available at *Biostatistics* online. □

In summary, we developed a new EM algorithm for the estimation of the evolutionary rates for the given poset structure. In the E-step, the conditional expected values of random variable $T_i - \max_{j\in \mathrm{pa}(i)} T_j$ (or simply Z_i) is computed for each observation and each mutation $i \in P$. In the M-step, these expected values are then used to compute the MLEs for the rate parameters λ. The EM algorithm is implemented in the R language and the code is available at: http://www.cbg.ethz.ch/software/otcbn.

3. Simulation study

We studied the performance of the OT-CBN and CT-CBN models in simulation experiments. Estimated rate parameters ($\hat{\lambda}_e^{\mathrm{OT}}$ and $\hat{\lambda}_e^{\mathrm{CT}}$) of two models were compared with the true rate parameter for each event e of a given poset. The mean absolute error (MAE), $|\hat{\lambda}_e - \lambda_e|$, was computed for each event. The relative MAE difference, $\Delta\mathrm{MAE}_e = (\mathrm{MAE}_e^{\mathrm{CT}} - \mathrm{MAE}_e^{\mathrm{OT}})/\mathrm{MAE}_e^{\mathrm{CT}}$, is reported for each event e of the considered posets for different simulation settings (see Tables 1–3 of supplementary material available at *Biostatistics* online).

In our simulation experiments, we studied 6 different posets: two empty posets and two linear posets each with four and six events, the poset shown in Figure 1(a) and the poset depicted in Figure 3 of supplementary material available at *Biostatistics* online. We investigated different sampling time distributions, namely sampling at constant time c for all samples (the density function is $\delta(c)$ where $\delta(\cdot)$ is the Dirac delta function), exponential sampling time distribution $T_s \sim \exp(\lambda_s)$, normal sampling time distribution $N(\mu_s, \sigma_s)$, and a distribution in which sampling time depends on the mutational process. The dependent sampling time distribution is a mixture of the distribution of $\max(T_1, T_2)$ (the waiting time to the occurrence of both the first and second mutation) and a uniformly distributed component to introduce variability into the simulated datasets. The weight of the uniform component is 0.05. The parameters of the different sampling distributions except the dependent distribution are selected such that the expected value of the sampling time is equal to one. Hence, all the parameters c, λ_s, and μ_s are equal to one and the standard error of the normal distribution is set to $\sigma_s = \mu_s/10$. For each poset, the rate parameter of each mutation, λ_e, was generated by drawing uniform random numbers between $\frac{1}{2}$ and 2 (between two-fold slower and faster than average sampling time).

We drew N pairs of genotypes and corresponding sampling times, $(g_i, t_{s,i})$ for $i = 1, \ldots, N$, for each parameter setting. The OT-CBN and CT-CBN models were estimated and compared using the simulated

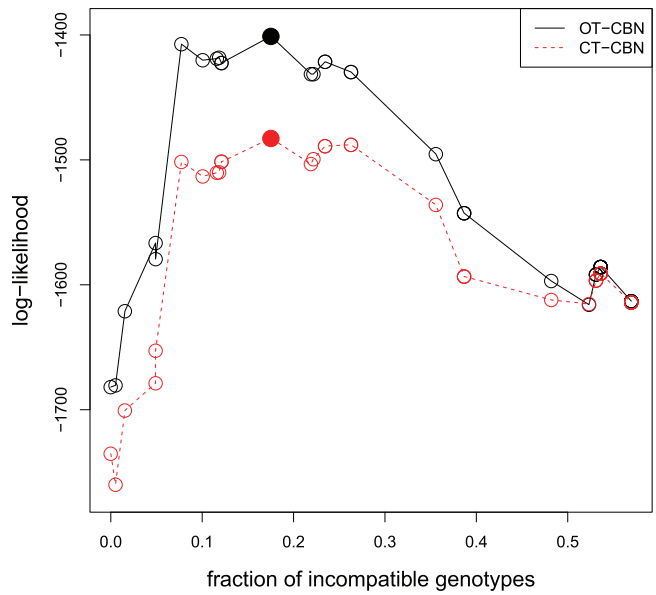

Fig. 3. ML estimation of the OT-CBN and CT-CBN models for the HIV dataset ZDV+3TC-all. The maximum log-likelihoods are computed for different fractions of incompatible genotypes, $1 - \alpha$. We generated this plot by estimating maximal posets P_ϵ for all possible values of ϵ in the extended model. Filled circles correspond to the ML posets of the OT-CBN and CT-CBN models (the poset is shown in Figure 4(d) for the OT-CBN model).

data. Tables 1–3 of supplementary material available at *Biostatistics* online show the results for the empty posets, the linear posets, and two posets shown in Figure 1(a) and Figure 3 of supplementary material available at *Biostatistics* online, respectively. Wilcoxon signed rank tests (paired tests) of the relative MAE of the OT-CBN model against the CT-CBN model were performed for all events. The p-values are corrected for multiple testing using the Benjamini–Hochberg method and significant ones are represented by adding an asterisk next to the relative MAE in Tables 1–3 of supplementary material available at *Biostatistics* online. For constant, normal, and dependent sampling distributions, the OT-CBN outperforms the CT-CBN in 87 comparisons out of a total of ninety comparisons (83 of them are significant). The CT-CBN model assumes that sampling times are distributed exponentially and hence its performance is better for exponential sampling distribution. Even in this case, the OT-CBN still outperforms the CT-CBN in 24 comparisons (only one of them is significant) and the CT-CBN is better in the other six comparisons (only two of them are significant). Tables 4–6 of supplementary material available at *Biostatistics* online show the similar results when the relative mean squared error is used as the performance measure. No major difference has been observed between the results of two performance measures.

It is evident from the simulation results that the OT-CBN is more accurate than the CT-CBN model in the estimation of the rate parameters for different posets and sampling time distributions. The superiority of the OT-CBN model is due to the fact that firstly this model takes into account not only observed genotypes but also observed sampling times for parameter estimation, and secondly as opposed to the CT-CBN model, the OT-CBN model makes no assumptions on the distribution of the sampling time and its independence of the mutational process.

4. HIV GENETIC DATA

In this section, we analyze five datasets obtained from the SHCS database (Schoeni-Affolter *and others*, 2010) (Table 7 of supplementary material available at *Biostatistics* online). We want to model the accumulation of resistance mutations in the reverse transcriptase gene of the HIV viral genome under monotherapy with zidovudine (ZDV) or combination therapy with zidovudine plus lamivudine (ZDV+3TC) or any therapy consisting of ZDV (denoted by ZDV+$*$). Resistance mutations for each therapy are selected from Johnson *and others* (2013) and reported in Table 7 of supplementary material available at *Biostatistics* online. The mutation 67N, for instance, indicates amino acid asparagine (N) has been observed at position 67 of the reverse transcriptase gene of the HIV viral genome. The datasets ZDV-fl and ZDV+3TC-fl only contain genotypes for first-line (fl) therapies, while the other datasets consist of genotypes for both first-line and salvage therapies. For all datasets, the sampling time of each individual genotype is observed. By applying the one-sample Kolmogorov–Smirnov test, we found that sampling time distributions deviate significantly from the exponential distribution, for all datasets. Figure 5 of supplementary material available at *Biostatistics* online shows to what extent each dataset deviates from the exponential assumption. Hence, one of the assumptions of the CT-CBN model is violated and we expect the OT-CBN model to be more accurate than the CT-CBN model for these datasets.

In real-world applications, the true dependency structure among mutations, i.e. the poset, is not usually known and one has to learn it from observed data. It has been shown that the ML poset is the largest poset that is compatible with all observed genotypes (Beerenwinkel and Sullivant, 2009). However, in practice, real-world datasets are not perfect and observations are subject to noise. Hence, since the ML poset only consists of relations that are compatible with all observations, the ML poset will be very sparse for most real-world datasets. The problem of noisy genotypes was addressed in Beerenwinkel *and others* (2007, 2011) and Beerenwinkel and Sullivant (2009). In this paper, we follow the approach outlined in Beerenwinkel and Sullivant (2009). Let P_ϵ be the maximal poset which consists of all relations which are violated by at most a fraction ϵ of all genotypes. The rate parameters λ of the poset P_ϵ are estimated

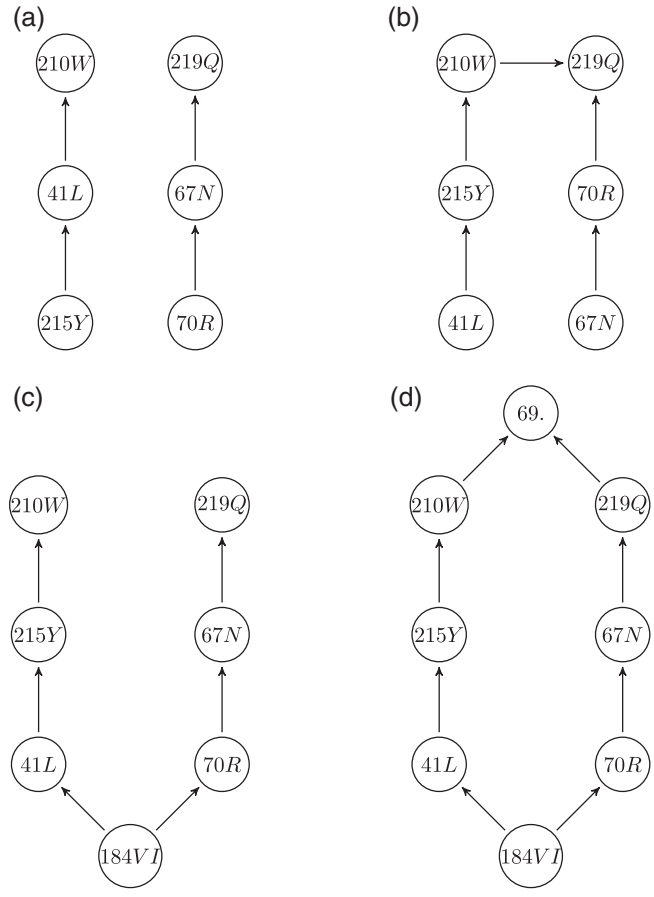

Fig. 4. The ML posets learned from the datasets explained in Table 7 of supplementary material available at *Biostatistics* online. The poset (a) is the ML estimate obtained from both datasets ZDV-fl or ZDV-all. The poset (b)–(d) were learned from the datasets ZDV+*-all, ZDV+3TC-fl, and ZDV+3TC-all, respectively. These posets explain from 78% to 95% of the input datasets (Table 7 of supplementary material available at *Biostatistics* online).

only by using the compatible genotypes. The incompatible genotypes are assumed to be generated with uniform probability $q_\epsilon = 1/(2^n - |J(P_\epsilon)|)$. Then, the probability that a random vector T, defined in (2.1), is compatible with the extended model P_ϵ is defined as

$$\text{Prob}\{(T, t_s) \vdash Q_g \mid \lambda, \alpha\} = \begin{cases} \alpha \text{Prob}\{(T, t_s) \vdash Q_g \mid \lambda\} & \text{if } Q_g \text{ refines } P_\epsilon, \\ (1-\alpha) q_\epsilon & \text{otherwise,} \end{cases}$$

where the parameter α denotes the fraction of genotypes compatible with the poset P_ϵ.

In Figure 3, the CT-CBN and OT-CBN estimates for different values of α are compared for the dataset ZDV+3TC-all. For all values of α, the likelihood of the OT-CBN model is larger than the likelihood of the CT-CBN model. The ML poset of the OT-CBN model explains 82% of the observations (Table 7 of supplementary material available at *Biostatistics* online, Figure 4(d)). Similarly, for all other datasets except ZDV+3TC-fl, the OT-CBN model outperforms the CT-CBN model (Figure 4 of supplementary material available at *Biostatistics* online). For ZDV+3TC-fl (Figure 4(c) of supplementary material available at

Table 1. *Comparison of the RMSE of the OT-CBN and CT-CBN models computed on the test data of each cross-validation fold for the HIV datasets*

Dataset	$\Delta_{\mathrm{RMSE}(\hat{t}_s)}$	p-value
ZDV-all	0.034	0.0020
ZDV-fl	0.035	0.0039
ZDV+3TC-all	0.057	0.0020
ZDV+3TC-fl	0.040	0.0273
ZDV+*-all	0.055	0.0020

The relative RMSE difference is defined as $\Delta_{\mathrm{RMSE}(\hat{t}_s)} = \{\mathrm{RMSE}(\hat{t_s}^{\mathrm{CT}}) - \mathrm{RMSE}(\hat{t_s}^{\mathrm{OT}})\}/\mathrm{RMSE}(\hat{t_s}^{\mathrm{CT}})$. The p-values are computed from Wilcoxon signed ranked tests of the RMSE of the OT-CBN model versus the CT-CBN model, based on ten-fold cross-validation.

Biostatistics online), this result might be due to the fact that this dataset has the smallest number of observations and hence least statistical power. All ML posets learned from the different datasets consist of two main components (Figure 4). The first component consists of the mutations 41L, 210W, and 215Y and the second one consists of the mutations 67N, 70R, and 219Q. This is in agreement with previous studies where two distinct evolutionary routes, called the 215-41 pathway and the 70-219 pathway, were suggested for the development of resistance to HIV under ZDV (Hanna *and others*, 2000; Beerenwinkel *and others*, 2005a).

In addition to comparing the likelihoods of two models on the HIV datasets, we also compared the accuracy of sampling time estimation for a given genotype for both models. For real-world datasets, the true rate parameters λ are not known and it is not possible to assess directly the accuracy of the estimated rate parameters. However, we can compare two models by comparing the observed sampling times and the estimated sampling times. The estimation of sampling times depends on the rate parameters and the poset. We can compute the ML estimates of the rate parameters and the poset using the training data of each cross-validation fold. Then, the sampling time for a given genotype is estimated by the expected sampling time $E(T_s \mid g, P, \lambda)$. In order to compute this expectation, we apply Bayes rule to obtain the posterior distribution of the sampling time,

$$P(T_s = t_s \mid g, P, \lambda) = \frac{P(g \mid T_s = t_s, P, \lambda) P(T_s = t_s \mid P, \lambda)}{\int_{t=0}^{\infty} P(g \mid T_s = t, P, \lambda) P(T_s = t \mid P, \lambda)\, \mathrm{d}t},$$

where the probability $P(g \mid T_s = t_s, P, \lambda)$ is computed by Theorem 2.3 (alternatively denoted by $\mathrm{Prob}\{(T, t_s) \vdash Q_g\}$ in the theorem) and $P(T_s = t_s \mid P, \lambda)$ is a prior distribution for sampling times. In the cross-validation setting, we approximated the prior distribution by a non-parametric kernel density estimation of training data of each cross-validation fold. The relative root-mean-square errors (RMSE) difference $\Delta_{\mathrm{RMSE}(\hat{t}_s)} = \{\mathrm{RMSE}(\hat{t_s}^{CT}) - \mathrm{RMSE}(\hat{t_s}^{OT})\}/\mathrm{RMSE}(\hat{t_s}^{CT})$ and the p-values obtained from Wilcoxon signed rank sum tests of the RMSE of the OT-CBN model versus the CT-CBN model are reported for all the considered datasets in Table 1. For all the HIV datasets, RMSE of the estimated sampling times of the OT-CBN model are significantly smaller than those of the CT-CBN model.

5. CONCLUSIONS

Timed CBNs have been used to analyze the timeline for the accumulation of mutations under partial temporal orders among mutations. In these models, mutations are assumed to happen after exponentially distributed waiting times. The waiting process for a mutation starts after all its predecessor mutations have

already occurred. In previous work, Beerenwinkel and Sullivant (2009) assumed that the sampling time for a given genotype is an unknown exponentially distributed random variable and is independent of the mutational process. Here, in the OT-CBN model, the sampling times are observed and no assumptions on the distribution of the sampling times and its dependency on the mutational process have been made. We developed an EM algorithm for estimation of the evolutionary rates for a given poset based on observed genotypes and corresponding sampling times. We compared the OT-CBN to the CT-CBN model on simulated data as well as real-world data from multiple genotypic HIV drug resistance datasets. In the simulation study, we investigated different sampling time distributions. The OT-CBN model is accurate in recovering true parameters for different distributions of sampling times while the CT-CBN was unable to recover true parameters for distributions that are non-exponential or dependent on the mutational process. For the HIV datasets, the OT-CBN was better than the CT-CBN model in terms of likelihood and sampling time estimation. We conclude that the superiority of the OT-CBN model is the result of less restrictive assumptions of the model as well as taking into account the individual sampling time points for parameter estimation.

We provide an analytical expression for the expected sampling time of a genotype, $\mathrm{E}(T_s \mid g)$. This quantity is closely related to the genetic progression score (GPS), defined as the expected waiting time of the genotype, $\mathrm{E}(\max_{e \in g} T_e)$ (Rahnenführer *and others*, 2005). The GPS was previously computed from simulations for a large number of samples obtained from mixture models of timed oncogenetic trees, and it was shown to be a medically relevant prognostic factor for glioblastoma and prostate cancer (Rahnenführer *and others*, 2005). Like the GPS, we expect the expected sampling time of the genotype to be an informative predictor of HIV treatment outcome. Since the observation time is a result of treatment failure, it may even be a better predictor of treatment success. Another interesting quantity, computed in this paper in Theorem 2.3, is the probability of observing a genotype at a certain sampling time t_s, $P(g \mid T_s = t_s, P, \lambda)$. The probabilities of different genotypes over time (see Example 2.4 and Figure 2) are of particular interest for clinicians. This additional information enables clinicians to get more insights about the underlying genotypic information of a patient without performing an actual genotypic resistance testing, which can be helpful in HIV therapy selection particularly in resource-limited countries where genotypic resistance testing may not be available (Prosperi *and others*, 2010; Revell *and others*, 2013).

The OT-CBN model has a number of limitations. Firstly, the use of cross-sectional data for the estimation of the evolutionary rates relies on the assumption of independent observations. The estimation of the rates can be improved by using longitudinal genotypic data where each patient might have more than one genotype at different time points. Further research is required to develop a new CBN model based on longitudinal data (see Beerenwinkel and Drton, 2007). Secondly, in the HIV application, we found that sampling times cannot be explained solely by the patient's genotype. In other words, we found that the observed sampling times for each genotype are highly variable. This suggests other relevant information of patients such as demographic variables and clinical outcomes have to be taken into account to explain this variability. Further research is needed to fit the CBN model not only based on the observed genotypes and sampling times but also based on other covariates specific to a given patient. This means the evolutionary rates and possibly even the network structure of the CBN model would become patient-specific. Hence, the extended CBN model would be able to describe the dynamics and dependencies of accumulating mutations for each individual patient. This extension could be particularly useful for therapy selection of HIV or in the case of cancer patients where treatment choices are highly personalized. Finally, the OT-CBN model does not work when some of the sampling times are missing. However, there are several ways to overcome this issue. One way is to perform imputation and replace the missing sampling times with sampling times of similar genotypes. A more sophisticated method would be to introduce a more general likelihood function for this case, i.e. $\prod_{(g,t_s)\in\mathcal{O}} \Pr\{(T, t_s) \vdash \mathcal{Q}_g\} \prod_{g\in\mathcal{M}} \Pr\{T \vdash \mathcal{Q}_g\}$ where $\mathcal{O}$ is the set of observed genotype-sampling time pairs and $\mathcal{M}$ is the set of genotypes for which the corresponding sampling times are missing. The first product is the observed likelihood of the OT-CBN model and the second

one is the observed likelihood of the CT-CBN model. Hence, a combined CT-CBN and OT-CBN approach is possible to overcome missing sampling time issues.

SUPPLEMENTARY MATERIAL

Supplementary material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

We thank the patients who participate in the SHCS; the physicians and study nurses for excellent patient care; the resistance laboratories for high-quality genotypic drug resistance testing; SmartGene, Zug, Switzerland, for technical support; Brigitte Remy, RN, Martin Rickenbach, MD, Franziska Schöni-Affolter, MD, and Yannick Vallet, MSc from the SHCS Data Center in Lausanne for data management; and Daniéle Perraudin and Mirjam Minichiello for administrative assistance. The members of the SHCS are: Aubert V., Battegay M., Bernasconi E., Böni J., Bucher H. C., Burton-Jeangros C., Calmy A., Cavassini M., Dollenmaier G., Egger M., Elzi L., Fehr J., Fellay J., Furrer H. (Chairman of the Clinical and Laboratory Committee), Fux C. A., Gorgievski M., Günthard H. (President of the SHCS), Haerry D. (deputy of "Positive Council"), Hasse B., Hirsch H. H., Hoffmann M., Hösli I., Kahlert C., Kaiser L., Keiser O., Klimkait T., Kouyos R., Kovari H., Ledergerber B., Martinetti G., Martinez de Tejada B., Metzner K., Müller N., Nadal D., Nicca D., Pantaleo G., Rauch A. (Chairman of the Scientific Board), Regenass S., Rickenbach M. (Head of Data Center), Rudin C. (Chairman of the Mother and Child Substudy), Schöni-Affolter F., Schmid P., Schüpbach J., Speck R., Tarr P., Telenti A., Trkola A., Vernazza P., Weber R., Yerly S. *Conflict of Interest*: H.F.G. has been an adviser and/or consultant for the following companies: GlaxoSmithKline, Abbott, Gilead, Novartis, Boehringer Ingelheim, Roche, Tibotec, Pfizer, and Bristol-Myers Squibb, and has received unrestricted research and educational grants from Roche, Abbott, Bristol-Myers Squibb, Gilead, Astra-Zeneca, GlaxoSmithKline, and Merck Sharp and Dohme (all money went to institution). He also is a DSMB member for EUROSIDA and Merck.

FUNDING

This work was supported by the Swiss HIV Cohort Study. The Swiss HIV Cohort Study is supported by the Swiss National Science Foundation [SNF grant #33CS30-134277]. The project was further supported by the SHCS projects #470, 528, 569, 629 (to N.B. and H.F.G), the SHCS Research Foundation, the Swiss National Science Foundation [grant #324730-130865 to H.F.G.], the European Community's Seventh Framework Program [grant FP7/ 2007–2013], under the Collaborative HIV and Anti-HIV Drug-resistance Network [CHAIN; grant 223131, to H.F.G.], by the Yvonne-Jacob foundation and by a further research grant of the Union Bank of Switzerland, in the name of an anonymous donor to H.F.G., an unrestricted research grant from Gilead, Switzerland to the SHCS research foundation, and by the University of Zurich's Clinical research Priority Program (CRPP) "Viral infectious diseases: Zurich Primary HIV Infection Study" (to H.F.G.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

ALTMANN, A., DÄUMER, M., BEERENWINKEL, N., PERES, Y., SCHÜLTER, E., BCH, J., RHEE, S.-Y., SNNERBORG, A., FESSEL, W. J., SHAFER, R. W., ZAZZI, M., KAISER, R. *and others* (2009). Predicting the response to combination antiretroviral therapy: retrospective validation of geno2pheno-THEO on a large clinical database. *The Journal of Infectious Diseases* **199**(7), 999–1006.

Beerenwinkel, N., Däumer, M., Sing, T., Rahnenfuhrer, J., Lengauer, T., Selbig, J., Hoffmann, D. and Kaiser, R. (2005). Estimating HIV evolutionary pathways and the genetic barrier to drug resistance. *The Journal of Infectious Diseases* **191**(11), 1953–1960.

Beerenwinkel, N. and Drton, M. (2007). A mutagenetic tree hidden Markov model for longitudinal clonal HIV sequence data. *Biostatistics* **8**(1), 53–71.

Beerenwinkel, N., Eriksson, N. and Sturmfels, B. (2006). Evolution on distributive lattices. *Journal of Theoretical Biology* **242**(2), 409–420.

Beerenwinkel, N., Eriksson, N. and Sturmfels, B. (2007). Conjunctive Bayesian networks. *Bernoulli* **13**(4), 893–909.

Beerenwinkel, N., Knupfer, P. and Tresch, A. (2011). Learning monotonic genotype-phenotype maps. *Statistical Applications in Genetics and Molecular Biology* **10**(1), 1–27.

Beerenwinkel, N., Montazeri, H., Schuhmacher, H., Knupfer, P., von Wyl, V., Furrer, H., Battegay, M., Hirschel, B., Cavassini, M., Vernazza, P. *and others* (2013). The individualized genetic barrier predicts treatment response in a large cohort of hiv-1 infected patients. *PLoS Computational Biology* **9**(8), e1003203.

Beerenwinkel, N., Rahnenführer, J., Kaiser, R., Hoffmann, D., Selbig, J. and Lengauer, T. (2005). Mtreemix: a software package for learning and using mixture models of mutagenetic trees. *Bioinformatics* **21**(9), 2106–2107.

Beerenwinkel, N., Sing, T., Lengauer, T., Rahnenführer, J., Roomp, K., Savenkov, I., Fischer, R., Hoffmann, D., Selbig, J., Korn, K. *and others* (2005). Computational methods for the design of effective therapies against drug resistant HIV strains. *Bioinformatics* **21**(21), 3943–3950.

Beerenwinkel, N. and Sullivant, S. (2009). Markov models for accumulating mutations. *Biometrika* **96**(3), 645–661.

Deforche, K., Cozzi-Lepri, A., Theys, K., Clotet, B., Camacho, R. J., Kjaer, J., Van Laethem, K., Phillips, A., Moreau, Y. and Lundgren, J. D. *and others* (2008). Modelled in vivo HIV fitness under drug selective pressure and estimated genetic barrier towards resistance are predictive for virological response. *Antiviral Therapy* **13**(3), 399–407.

Dempster, A., Laird, N. and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussions). *Journal of the Royal Statistical Society: Series B* **39**, 1–38.

Desper, R., Jiang, F., Kallioniemi, O. P., Moch, H., Papadimitriou, C. H. and Schäffer, A. A. (1999). Inferring tree models for oncogenesis from comparative genome hybridization data. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* **6**(1), 37–51.

Gerstung, M., Baudis, M., Moch, H. and Beerenwinkel, N. (2009). Quantifying cancer progression with conjunctive bayesian networks. *Bioinformatics* **25**(21), 2809–2815.

Hanna, G. J., Johnson, V. A., Kuritzkes, D. R., Richman, D. D., Brown, A. J. L., Savara, A. V., Hazelwood, J. D. and Richard, T. D. (2000). Patterns of resistance mutations selected by treatment of human immunodeficiency virus type 1 infection with zidovudine, didanosine, and nevirapine. *The Journal of Infectious Diseases* **181**(3), 904–911.

Johnson, V. A., Calvez, V., Gunthard, H. F., Paredes, R., Pillay, D., Shafer, R. W., Wensing, A. M. and Richman, D. D. (2013). Update of the drug resistance mutations in hiv-1: March 2013. *Topics in Antiviral Medicine* **21**(1), 6–14.

Lozovsky, E. R., Chookajorn, T., Brown, K. M., Imwong, M., Shaw, P. J., Kamchonwongpaisan, S., Neafsey, D. E., Weinreich, D. M. and Hartl, D. L. (2009). Stepwise acquisition of pyrimethamine resistance in the malaria parasite. *Proceedings of the National Academy of Sciences of the United States of America* **106**(29), 12025–12030.

Poelwijk, F. J., Kiviet, D. J., Weinreich, D. M. and Tans, S. J. (2007). Empirical fitness landscapes reveal accessible evolutionary paths. *Nature* **445**(7126), 383–386.

PROSPERI, M. C. F., D'AUTILIA, R., INCARDONA, F., DE LUCA, A., ZAZZI, M. AND ULIVI, G. (2009). Stochastic modelling of genotypic drug-resistance for human immunodeficiency virus towards long-term combination therapy optimization. *Bioinformatics* **25**(8), 1040–1047.

PROSPERI, M. C. F., ROSEN-ZVI, M., ALTMANN, A., ZAZZI, M., DI GIAMBENEDETTO, S., KAISER, R., SCHLTER, E., STRUCK, D., SLOOT, P., VAN DE VIJVER, D. A., VANDAMME, A-M. AND SÖNNERBORG, A. for the EuResist *and others* (2010). Antiretroviral therapy optimisation without genotype resistance testing: A perspective on treatment history based models. *PLoS One* **5**(10), e13753.

RAHNENFÜHRER, J., BEERENWINKEL, N., SCHULZ, W. A., HARTMANN, C., VON DEIMLING, A., WULLICH, B. AND LENGAUER, T. (2005). Estimating cancer survival and clinical outcome based on genetic tumor progression scores. *Bioinformatics (Oxford, England)* **21**(10), 2438–2446.

REVELL, A. D., WANG, D., WOOD, R., MORROW, C., TEMPELMAN, H., HAMERS, R. L., ALVAREZ-URIA, G., STREINU-CERCEL, A., ENE, L., WENSING, A. M. J. *and others* (2013). Computational models can predict response to HIV therapy without a genotype and may reduce treatment failure in different resource-limited settings. *Journal of Antimicrobial Chemotherapy* **68**, 1406–1414.

SAKOPARNIG, T. AND BEERENWINKEL, N. (2012). Efficient sampling for bayesian inference of conjunctive bayesian networks. *Bioinformatics* **28**(18), 2318–2324.

SCHOENI-AFFOLTER, F., LEDERGERBER, B., RICKENBACH, M., RUDIN, C., GÜNTHARD, H. F., TELENTI, A., FURRER, H., YERLY, S. AND FRANCIOLI, P. (2010). Cohort profile: the Swiss HIV Cohort Study. *International Journal of Epidemiology* **39**(5), 1179–1189.

VON HEYDEBRECK, A., GUNAWAN, B. AND FÜZESI, L. (2004). Maximum likelihood estimation of oncogenetic tree models. *Biostatistics (Oxford, England)* **5**(4), 545–556.

VON WYL, V., YERLY, S., BÖNI, J., BRGISSER, P., KLIMKAIT, T., BATTEGAY, M., FURRER, H., TELENTI, A., HIRSCHEL, B., VERNAZZA, P. L. *and others* (2007). Emergence of HIV-1 drug resistance in previously untreated patients initiating combination antiretroviral treatment: a comparison of different regimen types. *Archives of Internal Medicine* **167**(16), 1782–1790.

WEINREICH, D. M., DELANEY, N. F., DEPRISTO, M. A. AND HARTL, D. L. (2006). Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312**(5770), 111–114.

[*Received October 1, 2014; revised March 12, 2015; accepted for publication March 13, 2015*]