Statistical Applications in Genetics and Molecular Biology

Volume 10, *Issue* 1 2011 *Article* 3

Learning Monotonic Genotype-Phenotype Maps

Niko Beerenwinkel, *ETH Zürich*
Patrick Knupfer, *ETH Zürich*
Achim Tresch, *Ludwig-Maximilians-Universität München*

Recommended Citation:
Beerenwinkel, Niko; Knupfer, Patrick; and Tresch, Achim (2011) "Learning Monotonic Genotype-Phenotype Maps," *Statistical Applications in Genetics and Molecular Biology*: Vol. 10: Iss. 1, Article 3.
DOI: 10.2202/1544-6115.1603

Learning Monotonic Genotype-Phenotype Maps

Niko Beerenwinkel, Patrick Knupfer, and Achim Tresch

Abstract

Evolutionary escape of pathogens from the selective pressure of immune responses and from medical interventions is driven by the accumulation of mutations. We introduce a statistical model for jointly estimating the dynamics and dependencies among genetic alterations and the associated phenotypic changes. The model integrates conjunctive Bayesian networks, which define a partial order on the occurrences of genetic events, with isotonic regression. The resulting genotype-phenotype map is non-decreasing in the lattice of genotypes. It describes evolutionary escape as a directed process following a phenotypic gradient, such as a monotonic fitness landscape. We present efficient algorithms for parameter estimation and model selection. The model is validated using simulated data and applied to HIV drug resistance data. We find that the effect of many resistance mutations is non-linear and depends on the genetic background in which they occur.

KEYWORDS: genotype-phenotype map, conjunctive Bayesian networks, HIV drug resistance, isotonic regression

Author Notes: A.T. was supported by an LMUexcellent guest professorship. N.B. was partially supported by the Swiss National Science Foundation under grant no. CR32I2_127017.

1 Introduction

Most pathogens, including viruses, bacteria, eukaryotic parasites, and cancer cells, have a tendency to escape from selective pressure that is meant to control them. Rapid evolutionary change of the pathogen population facilitates escape from natural immune responses and from medical interventions such as chemotherapy. A quantitative understanding of evolutionary escape is at the heart of designing effective vaccines and treatment strategies.

The escape dynamics are governed by the space of possible genotypes that is accessible to the pathogen population, by the fitness landscape over these genotypes, and by additional population genetics parameters, such as population size and mutation rate (Iwasa, Michor, and Nowak, 2003). Here, we focus on the structure of the genotype space and the fitness landscape defined on it. We develop a statistical framework to estimate this fitness landscape from observed data subject to order and monotonicity constraints.

Constraints on the order in which mutations reach fixation in a population are common to many biological systems (Weinreich, Delaney, Depristo, and Hartl, 2006, Poelwijk, Kiviet, Weinreich, and Tans, 2007, Lozovsky, Chookajorn, Brown, Imwong, Shaw, Kamchonwongpaisan, Neafsey, Weinreich, and Hartl, 2009). We represent these constraints by a partial order among mutational events. The genotype space is the lattice of order ideals of this poset (Figure 1). For the fitness landscape, we assume that evolution proceeds in a directed fashion following an evolutionary gradient. We require that whenever a genotype g precedes another genotype h, their fitness is non-decreasing, $\phi(g) \leq \phi(h)$. This assumption appears reasonable in the situations indicated above, where the pathogen is under strong selective pressure and can avoid extinction only by accumulating advantageous mutations.

In the present paper, our goal is to jointly estimate both the underlying mutational order constraints and the fitness landscape from observed genotype-phenotype data. Estimating a fitness landscape amounts to learning a mapping that assigns each genotype a non-negative fitness value, or more generally, a phenotype. Because of the monotonicity assumption that we make, the regression problem is constraint and known as isotonic regression.

The two tasks of estimating mutational dependencies and of estimating a fitness landscape have been addressed separately before. Regressing phenotype on genotype is a recurrent task, because understanding the genotype-phenotype map is a central question in biology (Sevin, DeGruttola, Nijhuis, Schapiro, Foulkes, Para, and Boucher, 2000, Reidys and Stadler, 2002, Beerenwinkel, Schmidt, Walter, Kaiser, Lengauer, Hoffmann, Korn, and Selbig, 2002, Beerenwinkel, Däumer, Oette, Korn, Hoffmann, Kaiser, Lengauer, Selbig, and Walter, 2003a, Wang and Larder, 2003, Draghici and Potter, 2003, Segal, Barbour, and Grant, 2004, Rabi-

nowitz, Myers, Banjevic, Chan, Sweetkind-Singer, Haberer, McCann, and Wolkowicz, 2006, Rhee, Taylor, Wadhera, Ben-Hur, Brutlag, and Shafer, 2006).

Estimating dependencies among mutations is also a question of general interest in molecular biology and genetics. Several statistical models have been proposed for this purpose, including Bayesian networks (Klingler and Brutlag, 1994, Deforche, Silander, Camacho, Grossman, Soares, Laethem, Kantor, Moreau, and Vandamme, 2006, Poon, Lewis, Pond, and Frost, 2007) and dependency networks (Carlson, Brumme, Rousseau, Brumme, Matthews, Kadie, Mullins, Walker, Harrigan, Goulder, and Heckerman, 2008). Order constraints represent a specific type of dependency and a specialized Bayesian network model, called conjuctive Bayesian network (CBN), has been proposed that uses a partial order to represent these constraints (Beerenwinkel, Eriksson, and Sturmfels, 2006, 2007, Beerenwinkel and Sullivant, 2009, Gerstung, Baudis, Moch, and Beerenwinkel, 2009).

Here, we introduce a more general statistical model based on a partially ordered set and on isotonic regression to describe constraint and directed evolution in genotype space. We present algorithms for estimating both the poset structure and the isotonic regression function from observed data. The resulting genotype-phenotype map is optimal in the likelihood sense subject to order constraints and monotonicity. The algorithms have been implemented in the R package `icbn`, available at `www.cbg.ethz.ch/software/icbn`.

The model is applied to a dataset of mutational patterns in the genome of HIV and the corresponding levels of phenotypic drug resistance of the respective viruses. We want to learn mutational order constraints that apply to the evolutionary escape of HIV from drug pressure and, at the same time, the genotype-phenotype map which assigns a resistance phenotype to each genotype and is non-decreasing in the induced genotype space.

In Section 2, we present a self-contained introduction of CBNs following Beerenwinkel et al. (2007), but with some simplifications and advancements. Section 3 is devoted to isotonic regression. In Section 4, we combine the two models to obtain the isotonic CBN (I-CBN) model, which is further developed into the noisy I-CBN (NI-CBN) to handle measurement noise in Section 5. Section 6 reports performance measures of the inference algorithms based on simulated data, and in Section 7, the application of the NI-CBN model to HIV drug resistance data is presented.

2 Conjunctive Bayesian networks

We consider a fixed finite set of genetic events $\mathscr{E}$, and assume that genetic changes are irreversible. To model the accumulation of these mutations, we define the CBN

as a triple $(\mathscr{E}, \prec, \theta)$, where "$\prec$" is a partial order on $\mathscr{E}$, and $\theta = (\theta_e)_{e\in\mathscr{E}} \in [0,1]^{\mathscr{E}}$ is a set of parameters. A relation $e_1 \prec e_2$ between two distinct events is interpreted as event e_2 requiring event e_1 to have happened before. A relation $e_1 \prec e_2$ is called a cover relation, if for all $e' \in \mathscr{E}$ with $e_1 \prec e' \prec e_2$, either $e' = e_1$ or $e' = e_2$.

A subset g of events is called a genotype. The set of all possible genotypes, denoted $\mathscr{G}$, is the power set of $\mathscr{E}$, which is identified in a natural way with the set of all binary strings of length $|\mathscr{E}|$ by assigning $g \subset \mathscr{E}$ to $(g_e)_{e\in\mathscr{E}}$ with $g_e = 1$ if $e \in g$, and $g_e = 0$ otherwise. With subset inclusion, $\mathscr{G}$ forms a distributive lattice. We say that a genotype $g \subseteq \mathscr{E}$ and a relation $e_1 \prec e_2$ are compatible, if $(e_2 \in g) \Rightarrow (e_1 \in g)$ holds. This definition extends to sets of genotypes and to sets of relations in the obvious way. The state space $G(\mathscr{E}, \prec)$ of the CBN model is defined as the set of all genotypes that are compatible with $(\mathscr{E}, \prec)$, The elements of $G(\mathscr{E}, \prec)$ are the order ideals of the poset $(\mathscr{E}, \prec)$, where an order ideal is a subset $g \subseteq \mathscr{E}$ that is closed downwards, i.e., if $e_2 \in g$ and $e_1 \prec e_2$, then $e_1 \in g$. Conversely, given any set of genotypes $G \subseteq \mathscr{G}$, let $(\mathscr{E}, \prec_G)$ be the set of all events compatible with G. Then $(\mathscr{E}, \prec_G)$ forms a poset, which is the unique largest poset compatible with G. For the empty poset with no relation, we have $G(\mathscr{E}, \prec_{\text{empty}}) = \mathscr{G}$ and $(\mathscr{E}, \prec_{\mathscr{G}}) = (\mathscr{E}, \prec_{\text{empty}})$. We refer to the genotype $g = \emptyset$ as the wild type, and to $g = \mathscr{E}$ as the completely mutated type.

For a genotype g, we denote by $\text{Exit}_{\prec}(g)$ the set of all events that have not yet occurred in g but could happen next. An event $e \in \mathscr{E}$ might happen next if and only if e is minimal in $\mathscr{E} \setminus g$ with respect to the partial order. For $e \in \mathscr{E}$, let θ_e be the conditional probability that the event e has occurred given that all of its predecessor events have already occurred. The CBN defines the following probability distribution for the discrete random variable X with state space $G(\mathscr{E}, \prec)$

$$\Pr(X = g \mid \mathscr{E}, \prec, \theta) = \prod_{e\in g} \theta_e \cdot \prod_{e\in \text{Exit}_{\prec}(g)} (1-\theta_e) \tag{1}$$

We write $\text{CBN}(\mathscr{E}, \prec, \theta)$ for this statistical model. The probability of observing $g \in G(\mathscr{E}, \prec)$ is the probability that all the events in g have happened times the probability that none of the events that could happen next has occurred.

CBNs are Bayesian network models and they can also be defined as graphical models as follows. Consider the graph H with vertex set $\mathscr{E}$ and edges $e_1 \to e_2$ for all cover relations $e_1 \prec e_2$. The CBN model is the directed graphical model defined by H and the probability tables

$$\tau^e = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1-\theta_e & \theta_e \end{pmatrix}$$

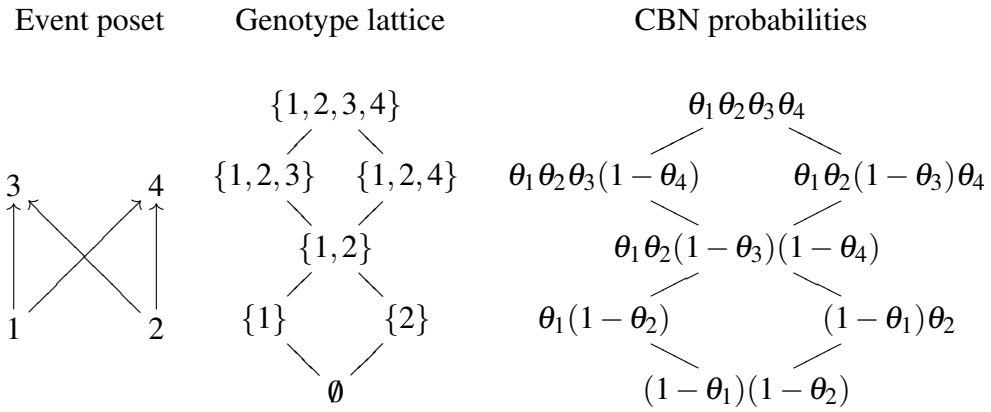

Figure 1: Conjunctive Bayesian network (CBN) model. Shown is the event poset (left), the induced genotype lattice (center), and the genotype probabilities (right) of the CBN model introduced in Example 1. In the event poset, each directed edge $e_1 \to e_2$ stands for a relation $e_1 \prec e_2$.

The entries of τ^e are the conditional probabilities $\tau^e_{a,b} = \Pr(X_e = b \mid X_{\mathrm{pa}(e)} = a)$, for all $a \in \{0,1\}^{\mathrm{pa}(e)}$ and $b \in \{0,1\}$, where $\mathrm{pa}(e)$ denotes the parents of e in H, $\mathbf{1} = (1,\ldots,1)$, and $\tau^e_{\mathbf{1},1} = \Pr(X_e = 1 \mid X_{\mathrm{pa}(e)} = \mathbf{1}) = \theta_e$. The joint distribution of X factorizes as

$$\begin{aligned}\Pr(X = g \mid H, \tau) \;&=\; \prod_{e\in\mathscr{E}} \Pr(X_e = g_e \mid X_{\mathrm{pa}(e)} = g_{\mathrm{pa}(e)}) \;=\; \prod_{e\in\mathscr{E}} \tau^e_{g_{\mathrm{pa}(e)},g_e} \\ &=\; \prod_{\substack{e\in g \\ \mathrm{pa}(e)=\mathbf{1}}} \theta_e \prod_{\substack{e\notin g \\ \mathrm{pa}(e)=\mathbf{1}}} (1-\theta_e) \prod_{\substack{e\notin g \\ \mathrm{pa}(e)\neq\mathbf{1}}} 1 \prod_{\substack{e\in g \\ \mathrm{pa}(e)\neq\mathbf{1}}} 0 \;=\; \Pr(X = g \mid \mathscr{E}, \prec, \theta)\end{aligned}$$

because the index sets of the first, second, and last product are, respectively, g, $\mathrm{Exit}_\prec(g)$, and the empty set, for all $g \in G(\mathscr{E}, \prec)$.

Example 1. *Let* $\mathscr{E} = \{1,2,3,4\}$ *with the relations* $1 \prec 3$, $1 \prec 4$, $2 \prec 3$ *and* $2 \prec 4$. *The lattice of order ideals of this poset consists of the seven genotypes* $G(\mathscr{E}, \prec) = \{\emptyset, \{1\}, \{2\}, \{1,2\}, \{1,2,3\}, \{1,2,4\}, \{1,2,3,4\}\}$ *(Figure 1). The CBN model* $(\mathscr{E}, \prec, \theta)$ *is given by the probabilities*

$$\begin{aligned}
\Pr(\emptyset) &= (1-\theta_1)(1-\theta_2) \\
\Pr(\{1\}) &= \theta_1(1-\theta_2) \\
\Pr(\{2\}) &= (1-\theta_1)\theta_2 \\
\Pr(\{1,2\}) &= \theta_1\theta_2(1-\theta_3)(1-\theta_4) \\
\Pr(\{1,2,3\}) &= \theta_1\theta_2\theta_3(1-\theta_4) \\
\Pr(\{1,2,4\}) &= \theta_1\theta_2(1-\theta_3)\theta_4 \\
\Pr(\{1,2,3,4\}) &= \theta_1\theta_2\theta_3\theta_4
\end{aligned}$$

In the remainder of this section, we recall maximum likelihood (ML) parameter estimation and model selection for CBNs from (Beerenwinkel et al., 2007). Let $(\mathscr{E}, \prec, \theta)$ be a CBN model. The data for this model is a count vector $n = (n_g) \in \mathbb{N}^{\mathscr{G}}$, where n_g is the number of observations of genotype g. We assume throughout the paper that each event $e \in \mathscr{E}$ has been observed in at least one genotype, i.e., $\sum_{g: e \in g} n_g > 0$. The log-likelihood function of the CBN model is

$$\ell_X(\theta) = \sum_{g \in G} n_g \left[\sum_{e \in \mathscr{E}} \log(\theta_e) + \sum_{e \in \mathrm{Exit}_\prec(g)} \log(1-\theta_e) \right] \tag{2}$$

Proposition 1. *Let $(\mathscr{E}, \prec)$ be a fixed poset and $n \in \mathbb{N}^{\mathscr{G}}$ an observed set of genotypes. The ML parameters of the CBN model $(\mathscr{E}, \prec, \theta)$ are given by*

$$\hat{\theta}_e = \frac{\sum_{g: e \in g} n_g}{\sum_{g: \mathrm{below}_\prec(e) \subseteq g} n_g}, \quad \textit{for all } e \in \mathscr{E},$$

where $\mathrm{below}_\prec(e) = \{e' \in \mathscr{E} \mid e' \neq e \text{ and } e' \prec e\}$ *is the set of events strictly below e.*

Proof. See (Beerenwinkel et al., 2007, Prop. 2). □

We say that a set of genotypes $G \subset \mathscr{G}$ separates the events, if for any two distinct elements $e_1, e_2 \in \mathscr{E}$, there exists a genotype $g \in G$ and $i \in \{1,2\}$ such that $g \cap \{e_1, e_2\} = \{e_i\}$. It is easy to see that for $G \subset \mathscr{G}$, the relation $\prec_G$ on $\mathscr{E}$ is reflexive and transitive. Furthermore, if G separates the events, then $\prec_G$ is a partial order on $\mathscr{E}$. The support of a data set $n \in \mathbb{N}^{\mathscr{G}}$ is defined as the set of genotypes that have actually been observed, $\mathrm{supp}(n) = \{g \in \mathscr{G} \mid n_g > 0\}$. If $\mathrm{supp}(n)$ does not separate the events, then there exist events that are always observed in common. The observation of several of those events does not provide additional information. Hence non-separable events may be mapped to one event. The following result has been reported in (Beerenwinkel et al., 2007, Thm. 5). Here, we present a new and simplified proof.

Theorem 1. *Let $n \in \mathbb{N}^{\mathscr{G}}$ be a set of observed genotypes. If* $\mathrm{supp}(n)$ *separates the events, then the ML CBN model is $(\mathscr{E}, \prec_{\mathrm{supp}(n)}, \hat{\theta})$, with $\hat{\theta}$ defined as in Proposition 1 for the partial order $\prec_{\mathrm{supp}(n)}$.*

Proof. Recall that $(\mathscr{E}, \prec_{\mathrm{supp}(n)})$ is the unique largest poset compatible with $\mathrm{supp}(n)$. For any event poset $(\mathscr{E}, \prec)$ that is not compatible with $\mathrm{supp}(n)$, the likelihood function $L_X(\theta) = \Pr(n \mid \mathscr{E}, \prec, \theta)$ is identical zero. Thus, it is sufficient to show that if $\prec_1$ and $\prec_2$ are two partial orders on $\mathscr{E}$ that are compatible with $\mathrm{supp}(n)$ and $\prec_2$ is larger than $\prec_1$ (i.e., for all $e, e' \in \mathscr{E}$, $e \prec_1 e'$ implies $e \prec_2 e'$), then the likelihood is non-decreasing, $\Pr(n \mid \mathscr{E}, \prec_1) \leq \Pr(n \mid \mathscr{E}, \prec_2)$.

Let $g \in \mathscr{G}$ be a genotype. If $\prec_2$ is larger than $\prec_1$, then

$$\min_{\prec_2} \mathscr{E} \setminus g = \mathrm{Exit}_{\prec_2}(g) \subseteq \mathrm{Exit}_{\prec_1}(g) = \min_{\prec_1} \mathscr{E} \setminus g$$

To see this, suppose that $e \in \mathscr{E} \setminus g$ is not $\prec_1$-minimal. Then there is an element $d \in \mathscr{E} \setminus g$ with $d \prec_1 e$. But this implies $d \prec_2 e$ and hence e is not $\prec_2$-minimal either.

For any genotype compatible with $\prec_{\mathrm{supp}(n)}$ (and hence also with $\prec_1$ and $\prec_2$), we find

$$\begin{aligned} \Pr(X = g \mid \mathscr{E}, \prec_1, \theta) &= \prod_{e \in g} \theta_e \cdot \prod_{e \in \mathrm{Exit}_{\prec_1}(g)} (1 - \theta_e) \\ &\leq \prod_{e \in g} \theta_e \cdot \prod_{e \in \mathrm{Exit}_{\prec_2}(g)} (1 - \theta_e) = \Pr(X = g \mid \mathscr{E}, \prec_2, \theta) \end{aligned}$$

We assume that genotype observations are independent, hence

$$\begin{aligned} \Pr(n \mid \mathscr{E}, \prec_1, \theta) &= \prod_{g \in \mathrm{supp}(n)} \Pr(X = g \mid \mathscr{E}, \prec_1, \theta)^{n_g} \\ &\leq \prod_{g \in \mathrm{supp}(n)} \Pr(X = g \mid \mathscr{E}, \prec_2, \theta)^{n_g} = \Pr(n \mid \mathscr{E}, \prec_2, \theta) \end{aligned}$$

$(\mathscr{E}, \prec_{\mathrm{supp}(n)})$ is a partial order, because $\mathrm{supp}(n)$ separates the events. By definition, no compatible poset can contain more relations than $(\mathscr{E}, \prec_{\mathrm{supp}(n)})$. Thus

$$\Pr(n \mid \mathscr{E}, \prec, \theta) \leq \Pr(n \mid \mathscr{E}, \prec_{\mathrm{supp}(n)}, \theta) \leq \Pr(n \mid \mathscr{E}, \prec_{\mathrm{supp}(n)}, \hat{\theta})$$

for any partial order $\prec$ and any parameter vector θ. □

3 Isotonic regression

In this section, we fix a given poset $(\mathscr{E}, \prec)$ with genotype lattice $G = G(\mathscr{E}, \prec)$. We assume that the evolutionary process on G, i.e., the partially ordered accumulation

of mutations, follows a certain one-dimensional real-valued phenotype in a monotonic fashion. We require that the genotype-phenotype map $\phi : G \to \mathbb{R}$ satisfies for all $g_1, g_2 \in G$,

$$g_1 \subseteq g_2 \;\Rightarrow\; \phi(g_1) \leq \phi(g_2)$$

Our goal is to estimate the unknown monotonic function ϕ from observed genotype-phenotype pairs $(g, y) \in G \times \mathbb{R}$.

We assume that the conditional phenotypes $Y \mid X = g$ are independent normal random variables with unknown means μ_g and common unknown variance σ^2,

$$Y \mid X = g \;\sim\; \text{Norm}(\mu_g, \sigma^2), \quad \text{for all } g \in G$$

Let $y_g = \{y_{g,1}, \ldots, y_{g,n_g}\}$ be the phenotypes observed with genotype g. For a given dataset $(y_g)_{g \in G}$, the conditional log-likelihood is

$$\ell_{Y|X=g}(\mu, \sigma) = -\frac{N}{2}\log(2\pi) - N\log(\sigma) - \frac{1}{2\sigma^2}\sum_{g \in G}\sum_{j=1}^{n_g}(y_{g,j} - \mu_g)^2 \tag{3}$$

where $N = \sum_{g \in G} n_g$ is the total size of the data.

We estimate the parameters $\mu = (\mu_g)_{g \in G}$ and σ^2 from the data using ML subject to the monotonicity constraints

$$g_1 \subseteq g_2 \;\Rightarrow\; \mu_{g_1} \leq \mu_{g_2}, \quad \text{for all } g_1, g_2 \in G \tag{4}$$

This problem is known as the isotonic regression problem and its solution has the following structure. Let $\bar{y}_g = (1/n_g)\sum_{j=1}^{n_g} y_{g,j}$ denote the average phenotype observed with genotype g. For fixed σ, the ML estimates (MLEs) of μ are found by minimizing the sum of squares

$$\sum_{g \in G}\sum_{j=1}^{n_g}(y_{g,j} - \mu_g)^2 \;=\; \sum_{g \in G}\left[\sum_{j=1}^{n_g}(y_{g,j} - \bar{y}_g)^2 + n_g(\bar{y}_g - \mu_g)^2\right]$$

subject to the constraints (4), i.e., by solving

$$\begin{aligned} \min_{\mu} \quad & \sum_{g \in G}(\bar{y}_g - \mu_g)^2 n_g \\ \text{s. t.} \quad & \mu_{g_1} \leq \mu_{g_2} \text{ for all } g_1 \subseteq g_2 \text{ in } G \end{aligned} \tag{5}$$

The optimization problem (5) is a convex quadratic programming problem with a unique local solution $\hat{\mu}$ which is also the global minimum. Several algorithms have been proposed for solving this constraint least squares problem (Barlow, Bartholomew, Bremner, and Brunk, 1972). In our applications, we use the R package `isotone`

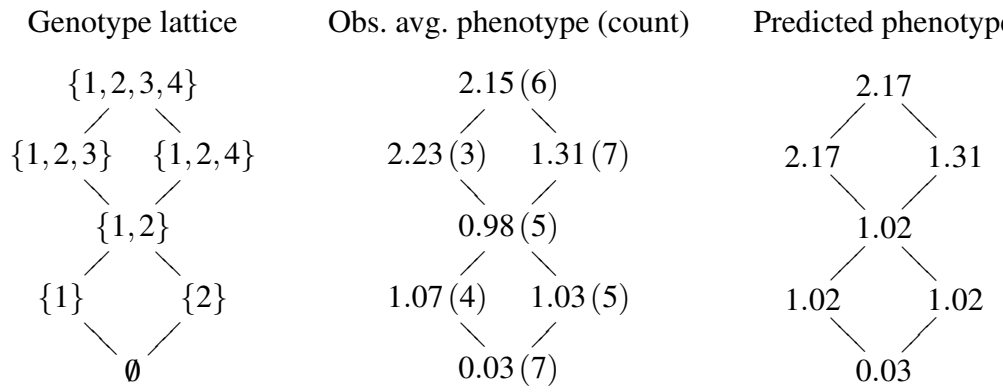

Figure 2: Isotonic regression on a genotype lattice. The genotype space (left) is the lattice of order ideals of the event poset shown in Figure 1. A total of 37 phenotypic measurements are summarized by their respective means and counts in the center diagram. The solution of the isotonic regression problem (5) is shown on the right, i.e., the estimated phenotypes $\hat{\mu}_g$. See Example 2 for more details.

which implements a solution based on a convex programming formulation with linear constraints and employs an active set algorithm (de Leeuw, Hornik, and Mair, 2009). The MLE of σ^2 is then

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{g \in G} \sum_{j=1}^{n_g} (y_{g,j} - \hat{\mu}_g)^2$$

Example 2. *For the genotype lattice of Example 1 and Figure 1, we consider the phenotype data summarized in the center diagram of Figure 2 by the average phenotypes $\bar{y}_g$ and, in parenthesis, the genotype counts n_g. The MLEs of μ are found by solving the optimization problem (5). The solution is displayed on the right of Figure 2 and it has the following block structure:*

$$\begin{aligned} \hat{\mu}_{\emptyset} &= 0.03 \\ \hat{\mu}_{\{1\}} = \hat{\mu}_{\{2\}} = \hat{\mu}_{\{1,2\}} &= 1.02 \\ \hat{\mu}_{\{1,2,4\}} &= 1.31 \\ \hat{\mu}_{\{1,2,3\}} = \hat{\mu}_{\{1,2,3,4\}} &= 2.17 \end{aligned}$$

The MLE of σ can not be computed from the average phenotypes $\bar{y}_g$, but only from the full data $\{y_{g,j}\}$ not shown in this example.

The estimated genotype-phenotype map is monotonic along any mutational pathway $g_1 \subset \cdots \subset g_k$ in G, and it has two additional properties that are important

in biological applications. First, the mapping is non-linear in the events. It allows for different phenotypic effects of the same genetic event, depending on the genetic context of the mutation. Second, the block structure implies that neighboring genotypes often have the same phenotype. In other words, blocks represent neutral mutational networks with respect to the considered phenotype.

4 Isotonic conjunctive Bayesian network model

We think of the observed genotype-phenotype pairs as intermediate steps of a non-reversible evolutionary process that is subject to partial order constraints and directed by a non-decreasing phenotype. For a fixed poset $(\mathscr{E}, \prec)$ with induced genotype lattice $G = G(\mathscr{E}, \prec)$, we define the joint distribution of genotype-phenotype pairs (X, Y) by the hierarchical model

$$\begin{aligned} X &\sim \mathrm{CBN}(\mathscr{E}, \prec, \theta) \\ Y \mid X = g &\sim \mathrm{Norm}(\mu_g, \sigma^2), \quad g \in G \end{aligned}$$

with $\mu_{g_1} \leq \mu_{g_2}$ whenever $g_1 \subseteq g_2$ in G. We call this model the Isotonic Conjunctive Bayesian Network (I-CBN) model. For a dataset $(n_g, y_g)_{g \in G}$, the log-likelihood function of the I-CBN model is the sum of the CBN log-likelihood (2) and the isotonic regression log-likelihood (3), $\ell_{X,Y}(\theta, \mu, \sigma^2) = \ell_X(\theta) + \ell_{Y|X}(\mu, \sigma^2)$. The results on ML parameter estimation and model selection for CBNs extend to I-CBNs as follows.

Proposition 2. *The ML parameters of the I-CBN model $(\mathscr{E}, \prec, \theta, \mu, \sigma)$ are given by*

$$\begin{aligned} \hat{\theta}_e &= \frac{\sum_{g: e \in g} n_g}{\sum_{g: \mathrm{below}_\prec(e) \subseteq g} n_g}, \quad \textit{for all } e \in \mathscr{E} \\ \hat{\mu} &= \min_{\mu} \sum_{g \in G} (\bar{y}_g - \mu_g)^2 n_g, \quad \textit{s.t. } \mu_{g_1} \leq \mu_{g_2} \textit{ for all } g_1 \subseteq g_2 \textit{ in } G \\ \hat{\sigma}^2 &= \frac{1}{N} \sum_{g \in G} \sum_{j=1}^{n_g} (y_{g,j} - \hat{\mu}_g)^2 \end{aligned}$$

Proof. See Proposition 1 and Section 3, and note that the partial derivatives of $\ell_{X,Y}$ are the same as those of ℓ_X and $\ell_{Y|X}$, respectively. □

Theorem 2. *Let $n \in \mathbb{N}^{\mathscr{G}}$ be a set of observed genotypes. If* $\mathrm{supp}(n)$ *separates the events, then the ML I-CBN model is $(\mathscr{E}, \prec_{\mathrm{supp}(n)}, \hat{\theta}, \hat{\mu}, \hat{\sigma}^2)$, with $\hat{\theta}$, $\hat{\mu}$, and $\hat{\sigma}^2$ defined as in Proposition 2 for the partial order $\prec_{\mathrm{supp}(n)}$.*

Proof. If $(\mathscr{E}, \prec)$ is not compatible with the data, then the likelihood function is zero. The poset $(\mathscr{E}, \prec_{\text{supp}(n)})$ is the unique maximal poset that is compatible with n. Suppose there are two different compatible posets $(\mathscr{E}, \prec_i)$, $i = 1, 2$, such that $(\mathscr{E}, \prec_2)$ is larger than $(\mathscr{E}, \prec_1)$. Then CBN$(\mathscr{E}, \prec_2)$ is more likely than CBN$(\mathscr{E}, \prec_1)$ and it suffices to shown that the isotonic regression likelihood is also non-decreasing.

The data n is compatible with both posets and we have

$$\text{supp}(n) \subseteq G(\mathscr{E}, \prec_{\text{supp}(n)}) \subseteq G(\mathscr{E}, \prec_2) \subset G(\mathscr{E}, \prec_1)$$

For any genotype $g \in G(\mathscr{E}, \prec_1) \setminus G(\mathscr{E}, \prec_2)$, we must have $n_g = 0$. Therefore the log-likelihood $\ell_{Y|X}$ does not differ whether evaluated on $G(\mathscr{E}, \prec_1)$ or $G(\mathscr{E}, \prec_2)$. □

We summarize the results of this section in the following algorithm for learning I-CBN models from data.

Algorithm 1. (Learning I-CBN models)

INPUT: *A dataset* $(n_g, y_g)_{g \in \mathscr{G}}$ *such that* supp(n) *separates the events* $\mathscr{E}$
OUTPUT: *The ML I-CBN model* $(\mathscr{E}, \prec_{\text{supp}(n)}, \hat{\theta}, \hat{\mu}, \hat{\sigma}^2)$
STEP 1: *Construct* $\prec_{\text{supp}(n)}$ *by setting, for all* $e_1, e_2 \in \mathscr{E}$, $e_1 \prec_{\text{supp}(n)} e_2$ *if and only if* $g \cap \{e_1, e_2\} \neq \{e_2\}$ *for all* $g \in \text{supp}(n)$. *Set* $G = G(\mathscr{E}, \prec)$.
STEP 2: *Compute the isotonic regression (5) to obtain the MLEs* $\hat{\mu} = (\hat{\mu}_g)_{g \in G}$.
STEP 3: *Compute the MLEs* $\hat{\sigma}^2$ *and* $\hat{\theta} = (\hat{\theta}_e)_{e \in \mathscr{E}}$ *according to Proposition 2.*
STEP 4: *Output the poset* $(\mathscr{E}, \prec_{\text{supp}(n)})$ *and the MLEs* $(\hat{\theta}, \hat{\mu}, \hat{\sigma}^2)$.

5 Error model

Algorithm 1 for learning I-CBN models using ML is appealing due its efficiency and simplicity. In practice, however, it is limited by the sensitivity of poset reconstruction (Step 1) to noise in the genotype data. A single, possibly erroneous, observed genotype containing e_2 but not e_1 is sufficient to remove the relation $e_1 \prec e_2$ from the optimal poset.

In order to account for noisy genotype observations, we extend the I-CBN model in this section. We follow the approach of Gerstung et al. (2009) and devise an error model which assumes that the true genotype Z is generated by the CBN model, but not directly observable, and that the observed genotype X is an erroneous copy of Z,

$$\Pr(X \mid Z) = \varepsilon^{d(X,Z)} (1 - \varepsilon)^{n - d(X,Z)}$$

where ε is the per-locus probability of a measurement error and d the Hamming distance between genotypes, i.e., the number of genetic events that occurred in exactly one of the two genotypes. We denote this error model by Err(Z, ε).

The model for (X,Y,Z) is defined hierarchically as

$$\begin{aligned} Z &\sim \mathrm{CBN}(\mathscr{E},\prec,\theta) \\ X \mid Z &\sim \mathrm{Err}(Z,\varepsilon) \\ Y \mid Z &\sim \mathrm{Norm}(\mu_Z,\sigma^2) \quad \text{with } \mu_{g_1} \le \mu_{g_2} \text{ for all } g_1 \subseteq g_2 \end{aligned}$$

The observed genotype X is independent of the observed phenotype Y given the true unobserved genotype Z. The noisy I-CBN (NI-CBN) model is defined as the marginalization of this model with respect to the unobserved data Z.

For fixed $(\mathscr{E},\prec)$, $G = G(\mathscr{E},\prec)$, and data $\{(x_i,y_i,z_i)\}_{i=1,\ldots,N}$, the complete-data log-likelihood of the NI-CBN model is

$$\ell_{X,Y,Z}(\theta,\varepsilon,\mu,\sigma^2) \;=\; \ell_Z(\theta) + \ell_{X|Z}(\varepsilon) + \ell_{Y|Z}(\mu,\sigma^2) \tag{6}$$

Hence the MLEs are given in Proposition 2 and by $\hat{\varepsilon} = [1/(N|\mathscr{E}|)]\sum_{i=1}^{N} d(x_i,z_i)$. The observed-data log-likelihood is

$$\ell_{X,Y}(\theta,\varepsilon,\mu,\sigma^2) \;=\; \sum_{i=1}^{N} \log \sum_{z_i \in G} \Pr(x_i,y_i,z_i)$$

In order to maximize this expression, we derive an Expectation Maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977).

The posterior of the hidden data Z given the observations (X,Y) is

$$\Pr(Z \mid X,Y) \;=\; \frac{\Pr(Z)\Pr(X \mid Z)\Pr(Y \mid Z)}{\sum_{Z'} \Pr(Z')\Pr(X \mid Z')\Pr(Y \mid Z')} \tag{7}$$

Let $\gamma_{i,g} = \Pr(Z_i = g \mid X,Y)$ denote the responsibility of genotype $g \in G$ for observation (x_i,y_i). Then, for all $g \in G$,

$$u_g = \mathbb{E}_{Z|X,Y}\left[\sum_{i=1}^{N} \delta(Z_i,g)\right] = \sum_{i=1}^{N} \gamma_{ig}$$

is the expected genotype count, where δ is the Kronecker delta function. This defines the E step.

For the M step, we estimate the model parameters by maximizing the expectation of the complete-data log-likelihood (6) with respect to the conditional

distribution (7). We obtain the following equations for updating the model parameters:

$$\begin{aligned}
\theta_e^{\text{new}} &= \frac{\sum_{g:e\in g} u_g}{\sum_{g:e\in \text{below}_{\prec}(e)\subseteq g} u_g}, \quad e \in \mathscr{E} \\
\varepsilon^{\text{new}} &= \frac{1}{N|\mathscr{E}|}\sum_{i=1}^{N}\sum_{g\in G} d(x_i,g)\,\gamma_{i,g} \\
\mu^{\text{new}} &= \min_{\mu} \sum_{g\in G} \left(\frac{1}{u_g}\sum_{i=1}^{N} y_i\gamma_{i,g} - \mu_g\right)^2 u_g \quad \text{s.t. } \mu_{g_1} \le \mu_{g_2} \text{ for all } g_1 \subseteq g_2 \text{ in } G \\
(\sigma^{\text{new}})^2 &= \frac{1}{N}\sum_{g\in G}\sum_{i=1}^{N}(y_i-\mu_g)^2\,\gamma_{i,g}
\end{aligned}$$

where the responsibilities are computed with the previous parameter estimates.

For model selection, i.e., finding the optimal poset structure, we employ simulated annealing (Kirkpatrick, Gelatt, and Vecchi, 1983), a heuristic search strategy, to find the ML poset. The poset space is sampled by modifications of relations of the current poset that result in a new poset. In each step, we allow for adding or removing a relation, or replacing two relations $e_1 \prec e_2 \prec e_3$ by $e_1 \prec e_2$ and $e_1 \prec e_3$. To speed up the procedure, we use the number of incompatible genotypes $|\mathscr{G} \setminus G(\mathscr{E}, \prec_{\text{new}})|$ as a filter to discard unpromising poset structures prior to likelihood computation (Gerstung et al., 2009).

6 Simulation study

We analyzed the performance of the simulated annealing algorithm in simulation experiments. Predicted posets were compared to the true posets in terms of the false positive rate (fpr), defined as the number of estimated false relations divided by $|\mathscr{E}|(|\mathscr{E}|-1)/2$ (the maximum number of possible relations), and the false negative rate (fnr), defined as the number of true relations not included in the estimated poset divided by the number of true relations. Using the cross-validated mean squared error (MSE) $\sum_g[\phi(g)-\hat{\phi}(g)]^2$, the NI-CBN model was compared to a baseline regression model that is linear in the events $\mathscr{E}$. We report the relative MSE difference, $\Delta_{\text{MSE}} = (\text{MSE}^{\text{linear}} - \text{MSE}^{\text{NI-CBN}})/\,\text{MSE}^{\text{linear}}$.

We analyzed six posets: two empty posets and two linear posets, each of size $|\mathscr{E}| = 4$ and 7, the poset of Example 1 shown in Figure 1, and the poset displayed in Figure 3A which was selected based on real data (see Section 7). For each poset, we investigated models with parameters $\varepsilon \in \{0.001, 0.01, 0.1\}$ and $\sigma \in \{0.1, 1\}$ by drawing N = 500 or 1000 samples. For the empty and the linear posets, the

Table 1: NI-CBN performance for empty posets. Symbols are defined in the main text. False positive rate (fpr) and false negative rate (fnr) are reported with their standard error (se). In the penultimate column, p is the p-value of a one-sided paired Wilcoxon rank sum test of the MSE of the NI-CBN model versus the linear model, based on the number of simulations given in the last column.

$\lvert\mathscr{E}\rvert$	ε	σ	N	fpr $\pm$ se	fnr $\pm$ se	Δ_{MSE}	$\log_{10} p$	runs
4	0.001	0.1	500	0	0	0.146	–17.7	100
4	0.001	0.1	1000	0	0	0.147	–17.7	100
4	0.001	1	500	0	0	0.106	–16.7	100
4	0.001	1	1000	0	0	0.139	–17.6	100
4	0.01	0.1	500	0	0	0.114	–17.7	100
4	0.01	0.1	1000	0	0	0.125	–17.7	100
4	0.01	1	500	0	0	0.121	–17.4	100
4	0.01	1	1000	0	0	0.119	–17.6	100
4	0.1	0.1	500	0	0	0.048	–17.5	100
4	0.1	0.1	1000	0	0	0.041	–17.7	100
4	0.1	1	500	0.007±0.003	0	0.029	–6.8	100
4	0.1	1	1000	0.003±0.002	0	0.029	–11.8	100
7	0.001	0.1	500	0.005±0.002	0	0.027	–6.5	50
7	0.001	0.1	1000	0.002±0.001	0	0.054	–9.3	50
7	0.001	1	500	0.040±0.003	0	0.001	–0.4	50
7	0.001	1	1000	0.017±0.003	0	0.008	–0.4	50
7	0.01	0.1	500	0.004±0.002	0	0.028	–6.0	50
7	0.01	0.1	1000	0	0	0.036	–8.7	50
7	0.01	1	500	0.040±0.003	0	0.002	–0.3	50
7	0.01	1	1000	0.018±0.003	0	-0.005	0.0	50
7	0.1	0.1	500	0.020±0.003	0	-0.008	0.0	50
7	0.1	0.1	1000	0.003±0.002	0	0.000	–0.2	50
7	0.1	1	500	0.045±0.003	0	-0.003	–0.3	50
7	0.1	1	1000	0.034±0.003	0	-0.008	–0.1	50

conditional probabilities θ were set such that all genotypes $g \in G(\mathscr{E}, \prec)$ have the same probability, $\theta_e^{\text{empty}} = 1/2$ for all $e \in \mathscr{E}$, and $\theta_i^{\text{linear}} = i/(i+1)$ for linear posets $1 \prec 2 \prec 3 \prec \cdots \prec |\mathscr{E}|$. For the poset of Example 1, equal genotype probabilities can not be achieved and θ was drawn uniformly from the interval $(0.5, 0.9)$. For the poset of Figure 3A the fitted values $\theta = (0.42, 0.40, 0.18, 0.59, 0.69, 0.87, 0.65)$ were used. The parameters μ of the NI-CBN model were generated by drawing uniform random numbers r_i, $i = 1, \ldots, |\mathscr{E}| - 1$, from the interval $(-1, 3)$, sorting

Table 2: NI-CBN performance for linear posets. Symbols are defined in the main text and in the legend of Table 1.

$\lvert\mathscr{E}\rvert$	ε	σ	N	fpr $\pm$ se	fnr $\pm$ se	Δ_{MSE}	$\log_{10} p$	runs
4	0.001	0.1	500	0	0	0.002	–4.1	100
4	0.001	0.1	1000	0	0	0.003	–14.3	100
4	0.001	1	500	0	0	0.002	–2.7	100
4	0.001	1	1000	0	0	0.003	–8.3	100
4	0.01	0.1	500	0	0.002±0.002	0.024	–17.5	100
4	0.01	0.1	1000	0	0	0.023	–17.7	100
4	0.01	1	500	0	0.003±0.003	0.024	–16.0	100
4	0.01	1	1000	0	0	0.025	–17.6	100
4	0.1	0.1	500	0.008±0.004	0.112±0.021	0.081	–17.3	100
4	0.1	0.1	1000	0.002±0.002	0.058±0.014	0.085	–17.7	100
4	0.1	1	500	0.012±0.004	0.130±0.020	0.086	–17.6	100
4	0.1	1	1000	0.002±0.002	0.073±0.017	0.084	–17.7	100
7	0.001	0.1	500	0	0.004±0.002	0.003	–12.0	100
7	0.001	0.1	1000	0	0	0.002	–14.9	100
7	0.001	1	500	0	0.010±0.003	0.004	–9.5	100
7	0.001	1	1000	0	0.003±0.003	0.003	–11.7	100
7	0.01	0.1	500	0	0.015±0.006	0.024	–17.7	100
7	0.01	0.1	1000	0	0.002±0.002	0.022	–17.7	100
7	0.01	1	500	0	0.020±0.006	0.025	–16.6	100
7	0.01	1	1000	0	0.001±0.001	0.020	–17.5	100
7	0.1	0.1	500	0.026±0.003	0.550±0.016	0.062	–17.6	100
7	0.1	0.1	1000	0.024±0.003	0.517±0.015	0.062	–17.7	100
7	0.1	1	500	0.040±0.005	0.514±0.017	0.069	–16.5	100
7	0.1	1	1000	0.039±0.004	0.488±0.016	0.067	–17.3	100

them as $-1 = r_0 < r_1 < \cdots < r_{|\mathscr{E}|-1} < r_{|\mathscr{E}|} = 3$, and setting $\mu_g = r_{|g|}$. This defines a graded fitness landscape, i.e., the fitness (or phenotype) depends only on the number of mutations (Beerenwinkel et al., 2006). The runtime for fitting each model was between one minute and two hours on a standard PC.

For the empty posets (Table 1), false negatives can not occur. False positive rates were generally small and always below 5%. For the linear posets (Table 2), false positive rates are comparably low, but the false negative rate can reach high levels, especially for high error rates $\varepsilon = 0.1$ and small sample sizes $N = 500$. Similar poset reconstruction performance was observed for the poset of Example 1 and for the poset of Figure 3A with somewhat increased false positive rates (Table 3). In

Table 3: NI-CBN performance for the poset of Example 1 (4 events) and the poset of Figure 3A (7 events). Symbols are defined in the main text and in the legend of Table 1.

$\lvert\mathscr{E}\rvert$	ε	σ	N	fpr $\pm$ se	fnr $\pm$ se	Δ_{MSE}	$\log_{10} p$	runs
4	0.001	0.1	500	0.003±0.002	0	0.078	–17.6	100
4	0.001	0.1	1000	0.002±0.002	0	0.074	–17.7	100
4	0.001	1	500	0.002±0.002	0.003±0.003	0.079	–16.5	100
4	0.001	1	1000	0.010±0.004	0	0.072	–16.2	100
4	0.01	0.1	500	0.007±0.003	0	0.084	–17.7	100
4	0.01	0.1	1000	0.005±0.003	0	0.090	–17.6	100
4	0.01	1	500	0.007±0.003	0	0.082	–16.0	100
4	0.01	1	1000	0.003±0.002	0	0.085	–17.6	100
4	0.1	0.1	500	0.042±0.008	0.063±0.018	0.073	–17.5	100
4	0.1	0.1	1000	0.040±0.008	0.028±0.012	0.077	–17.7	100
4	0.1	1	500	0.085±0.013	0.143±0.022	0.065	–17.6	100
4	0.1	1	1000	0.058±0.011	0.053±0.015	0.059	–17.7	100
7	0.001	0.1	500	0.001±0.001	0.014±0.004	0.084	–17.7	100
7	0.001	0.1	1000	0	0.003±0.002	0.076	–17.7	100
7	0.001	1	500	0.007±0.002	0.017±0.005	0.052	–12.5	100
7	0.001	1	1000	0.003±0.002	0.007±0.003	0.055	–16.3	100
7	0.01	0.1	500	0.011±0.003	0.029±0.006	0.077	–17.7	100
7	0.01	0.1	1000	0.003±0.002	0.004±0.002	0.073	–17.7	100
7	0.01	1	500	0.014±0.003	0.027±0.006	0.052	–15.3	100
7	0.01	1	1000	0.002±0.001	0.004±0.002	0.083	–17.3	100
7	0.1	0.1	500	0.082±0.008	0.543±0.023	0.041	–15.1	100
7	0.1	0.1	1000	0.090±0.009	0.404±0.023	0.052	–16.3	100
7	0.1	1	500	0.130±0.011	0.566±0.023	0.040	–10.7	100
7	0.1	1	1000	0.118±0.011	0.499±0.025	0.047	–14.8	100

general, poset reconstruction is increasingly difficult for larger posets, higher error rates ε, and smaller sample size N, while the impact of the phenotype variance σ^2 appears to be small (Tables 1–3).

For most posets and parameter constellations, the NI-CBN model significantly outperformed the linear model in terms of the MSE of predicted phenotypes. This was not the case only for some of the models defined by the empty poset on seven events, which was also the most difficult model to fit (Table 1). This expected superiority of the NI-CBN model confirms that linear models are not appropriate for many types of fitness landscapes.

7 Application to HIV drug resistance

We consider genetic changes in the HIV genome in response to drug therapy and analyze two dataset obtained from the Stanford HIV Drug Resistance Databse (Rhee, Gonzales, Kantor, Betts, Ravela, and Shafer, 2003). The first dataset consists of 617 observations of the HIV reverse transcriptase (RT) genotype and paired measurements of phenotypic resistance to the RT inhibitor zidovudine. Resistance levels are reported as the logarithm of the fold-change in susceptibility of the virus to the drug as compared to the wild type. The genetic events are the amino acid changes $\mathscr{E} = \{41\text{L}, 67\text{N}, 69\text{D}, 70\text{R}, 210\text{W}, 215\text{Y}, \text{and } 219\text{Q}\}$, where, for example, 41L stands for the occurrence of leucine (L) at position 41 of the HIV RT. These mutations are known to be involved in the development of zidovudine resistance (Shafer and Schapiro, 2008).

The poset of the ML NI-CBN found by simulated annealing is shown in Figure 3A. It exhibits two independent mutational pathways, one involving mutations 41L and 215Y, the other 67N and 70R, that have been described before (Boucher, O'Sullivan, Mulder, Ramautarsing, Kellam, Darby, Lange, Goudsmit, and Larder, 1992, Larder, 1994). In previous work, a more restrictive model class of tree posets was not able to find the independence of both pathways, but a much more complex mixture model of tree posets was (Beerenwinkel, Rahnenführer, Däumer, Hoffmann, Kaiser, Selbig, and Lengauer, 2005). The model applied here offers more structural flexibility with the same number of free model parameters and it integrates both genotypic and phenotypic data into a single model.

The induced genotype lattice $G(\mathscr{E}, \prec)$ and the predicted drug resistance levels are visualized in Figure 3B and listed in the Appendix (Table 4). The lattice consists of 28 genotypes and the estimated isotonic regression function groups these into twelve genotype blocks of identical resistance to zidovudine. This description of the evolutionary process is much simpler than considering all $|\mathscr{G}| = 2^7 = 128$ combinatorially possible genotypes. The model suggests that under the selective pressure of zidovudine, neutral networks of neighboring genotypes of (near) identical fitness exist.

Linear regression of zidovudine resistance on the genetic events $\mathscr{E}$ was slightly less accurate than the NI-CBN predictions with a MSE of 0.45 ± 0.024 versus 0.44 ± 0.025 as estimated by 10-fold cross-validation ($p = 0.053$, one-sided, paired Wilcoxon rank sum test). Despite the comparable predictive performance, the two models have a very different structure. The NI-CBN model allows for non-linear effects of mutations and for context dependancy, whereas in the linear model, the effect per mutation is averaged over all genetic contexts. For the zidovudine data, the linear model tends to underestimate resistance in genotypes with few mutations and to overestimate resistance when many mutations have occurred

(Appendix, Table 4).

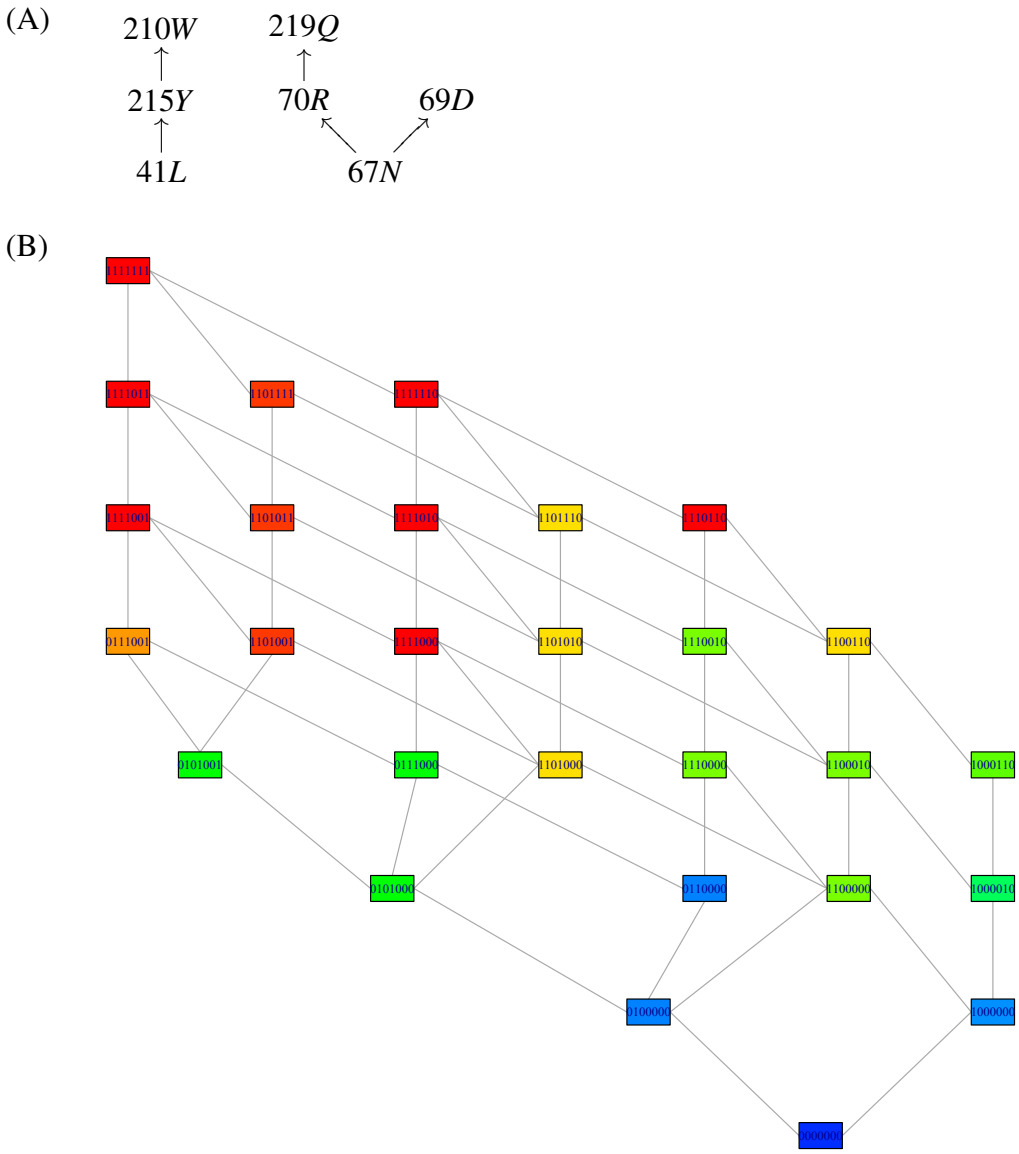

Figure 3: Cover relations of the optimal poset (A) and induced genotype lattice (B) for the development of HIV resistance to the nucleotide RT inhibitor zidovudine. Genotypes are encoded as binary strings that refer to the seven amino acid substitutions 41L, 67N, 69D, 70R, 210W, 215Y, and 219Q in the RT gene. The predicted levels of phenotypic resistance are color-coded (blue = fully susceptible, red = highly resistant). Further details, including the remaining model parameters and confidence intervals are given in the Appendix, Table 4 and Table 5.

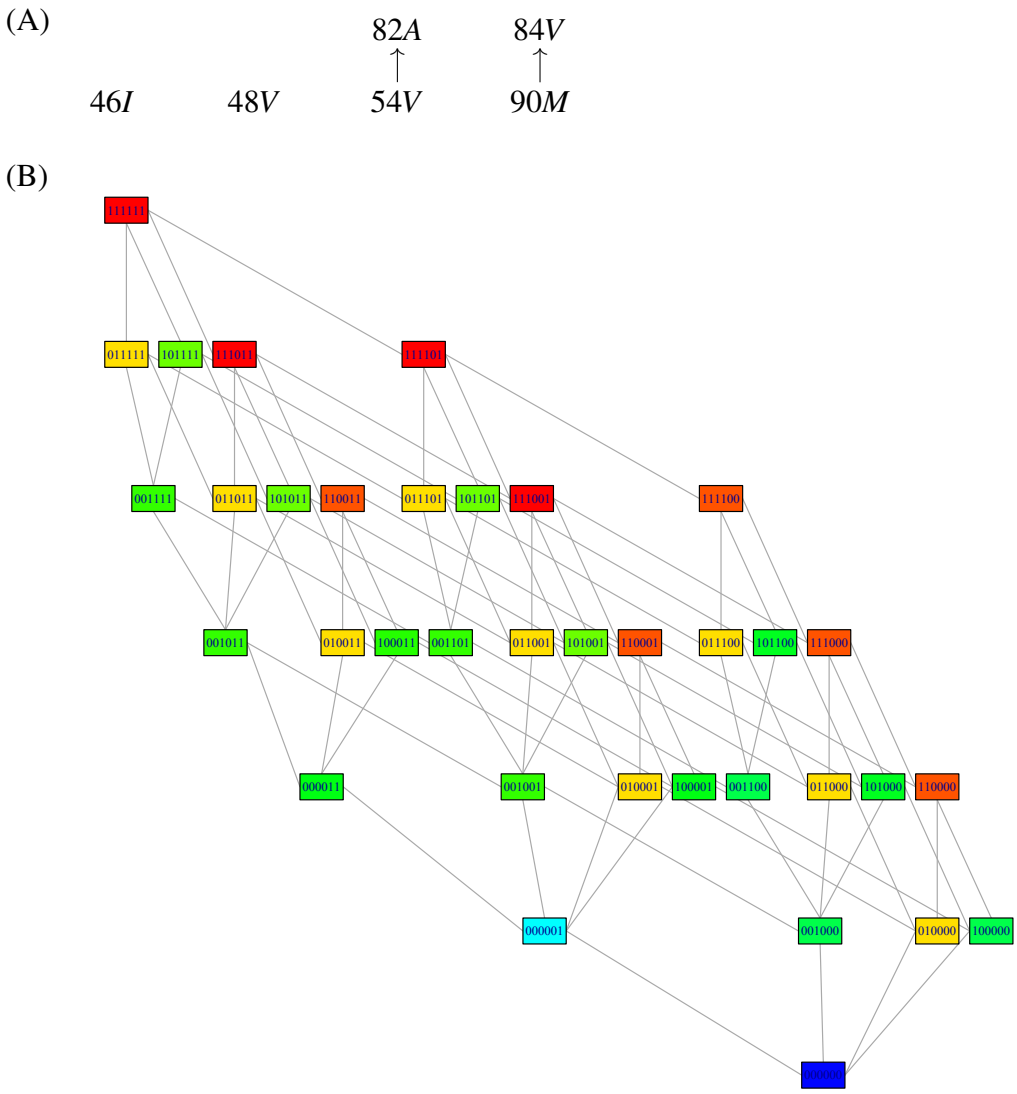

Figure 4: Cover relations of the optimal poset (A) and induced genotype lattice (B) for the development of HIV resistance to the PR inhibitor indinavir. Genotypes are encoded as binary strings that refer to the six amino acid substitutions 46I, 48V, 54V, 82A, 84V, and 90M in the PR gene. The predicted levels of phenotypic resistance are color-coded (blue = fully susceptible, red = highly resistant). Further details, including the remaining model parameters and confidence intervals are given in the Appendix, Table 7 and Table 8.

We assessed the uncertainty associated with model estimation using the bootstrap. For the fixed optimal model structure shown in Figure 3, the model parameters θ, μ, σ, and ε were re-estimated from 100 bootstrap samples. The resulting 95% confidence intervals are given in the Appendix, Tables 4 and 5. Another 100 bootstrap samples were used to quantify the uncertainty in estimating the model structure. In Table 6, the abundance of each cover relation (or equivalently, of each edge in the Bayesian network) among the 100 optimal posets is shown. This analysis strongly supports the optimal poset of Figure 3. The only appreciable uncertainty of the model structure that we detected is the order in which mutations 41L and 215Y occur. The data appears to favor the relation 41L $\prec$ 215Y, but it also provides some support for 215Y $\prec$ 41L, which indicates that both single mutants are almost equally likely to occur.

The second dataset consists of 1473 genotypes defined on the resistance-associated amino acid substitutions $\mathscr{E} = \{46\text{I}, 48\text{V}, 54\text{V}, 82\text{A}, 84\text{V}, 90\text{M}\}$ in the HIV protease (PR) and paired measurements of resistance to the PR inhibitor indinavir (Shafer and Schapiro, 2008). The optimal poset contains only two relations, inducing a genotype lattice of size 36 (Figure 4). The NI-CBN model groups these genotypes into 13 blocks of identical resistance levels (Appendix, Table 7). Again, the effect of several mutations appears to depend on the genetic background in which they occur. Because the linear regression model can not capture these dependencies, it is outperformed by the NI-CBN model in terms of MSE (0.27 ± 0.013 versus 0.25 ± 0.013, $p = 0.003$). All model parameters and their bootstrap confidence intervals are given in the Appendix, Tables 7 and 8. The structural uncertainty about the optimal poset is summarized in Table 9 of the Appendix, emphasizing the general stability of the poset while suggesting the cover relation 82A $\prec$ 54V as an alternative to 54V $\prec$ 82A, although with less than half the bootstrap support.

8 Conclusions

We have introduced a statistical model for jointly estimating the dynamics of accumulating mutations in a population and the associated phenotypic changes. The I-CBN model is a CBN model coupled with isotonic regression. It estimates constraints on the order in which mutations occur by a poset and the genotype-phenotype map (or fitness landscape) by a monotonic function. Parameter estimation and model selection are straightforward and efficient for this model. The NI-CBN model accounts for noisy observations and we have presented an EM algorithm for parameter estimation in this setting. For model selection, we propose a stochastic search procedure and we have implemented a simulated annealing algorithm.

The model has been tested on simulated data and applied to paired genotype-phenotype HIV drug resistance data. The NI-CBN model generalizes earlier efforts to estimate dependencies among HIV mutations from genotype data alone based on posets (Beerenwinkel et al., 2007, Beerenwinkel and Sullivant, 2009, Gerstung et al., 2009), tree posets or mixtures of trees (Beerenwinkel et al., 2005), and general Bayesian networks (Deforche et al., 2006). It can also be regarded as a model for regressing viral resistance phenotype on genotype. The isotonic regression model on the genotype lattice applied here combines the ability of non-linear models to account for context specificity with model interpretability.

Estimating drug resistance and the probability of evolutionary escape have been shown to improve predictions of clinical outcomes of antiretroviral therapy (Beerenwinkel, Lengauer, Däumer, Kaiser, Walter, Korn, Hoffmann, and Selbig, 2003b, Altmann, Beerenwinkel, Sing, Savenkov, Däumer, Kaiser, Rhee, Fessel, Shafer, and Lengauer, 2007). The NI-CBN model estimates both quantities jointly, and thus, will be a natural choice for enhancing clinical response predictions.

The monotonic block structure of the regression function highlights two features of evolutionary escape from drug pressure: the process is directed towards increasing levels of resistance and genotype blocks of identical resistance phenotype indicate connected neutral networks. Evolutionary escape may thus include neutral mutations within blocks and selectively advantages mutations that cause the transition to a new block. A similar drift-and-shift pattern of evolutionary escape from immune pressure has been described for Influenza A virus (Koelle, Cobey, Grenfell, and Pascual, 2006, van Nimwegen, 2006).

The NI-CBN model presented here can offer new insights into the structure of mutational pathways and the dynamics of evolutionary escape. In the future, the model might be improved in several ways. For example, large genetic event sets can not be handled with the current algorithms and often a pre-selection is necessary. The number of model parameters grows linearly with the lattice size, which in turn can be at worst exponential in the number of events. This raises the issue of overfitting of the regression function, and additional regularization may be beneficial. On the other hand, additional parameters could make the model more flexible and allow for better fitting of the obsevred data. For example, we have chosen to model phenotypic variance by a single parameter σ for all genotypes in order to keep the total number of model parameters small and because there was no obvious reason to believe that this term differs between genotypes. In principle, however, one can assume different variance parameters σ_g for each genotype g. Similarly, more detailed error models are conceivable that account separately for false positive and false negative observations (Beerenwinkel and Drton, 2007), or explicitly model the error process of the measuring device.

Although we have restricted our applications here to the development of HIV drug resistance, we expect the NI-CBN model to be useful also for other pathogens and for modeling the genetic progression of cancer, where the events may range from single nucleotide variants to large-scale genomic rearrangements.

Appendix

Table 4: HIV RT genotype lattice for zidovudine; see Figure 3.

	Data		Linear model	NI-CBN model			
g	n_g	$\bar{y}_g$	$\hat{\phi}_g$	$\hat{u}_g$	$\hat{\mu}_g$	95% CI	Block
0000000	196	-0.09	-0.07	241.6	-0.06	[-0.13, 0.04]	1
0100000	5	0.13	0.39	14.2	0.13	[-0.06, 0.54]	2
0110000	0	NA	0.62	3.3	0.13	[-0.06, 0.54]	2
1000000	17	0.21	0.34	13.9	0.15	[-0.08, 0.81]	3
1000010	34	0.71	0.78	43.1	0.80	[0.63, 1.02]	4
0101000	24	0.99	0.84	26.1	1.00	[0.61, 1.21]	5
0101001	35	0.97	1.09	57.1	1.00	[0.77, 1.28]	5
0111000	0	NA	1.07	0.9	1.00	[0.65, 2.23]	5
1000110	50	1.23	1.16	74.2	1.23	[1.01, 1.41]	6
1100000	1	1.60	0.80	1.9	1.30	[0.04, 1.69]	7
1100010	8	1.31	1.25	11.7	1.30	[0.97, 1.69]	7
1110000	0	NA	1.03	0.2	1.30	[0.04, 1.69]	7
1110010	6	1.03	1.48	4.6	1.30	[0.94, 1.69]	7
1100110	52	1.67	1.63	49.5	1.67	[1.49, 2.09]	8
1101000	7	2.13	1.25	5.3	1.67	[1.19, 2.05]	8
1101010	7	1.08	1.70	6.9	1.67	[1.19, 2.05]	8
1101110	7	1.29	2.08	9.0	1.67	[1.49, 2.09]	8
0111001	18	1.55	1.31	15.1	1.83	[1.20, 2.28]	9
1101001	11	1.77	1.50	8.2	2.06	[1.47, 2.48]	10
1101011	2	1.62	1.94	4.0	2.06	[1.53, 2.56]	10
1101111	7	2.11	2.32	8.4	2.06	[1.66, 2.61]	11
1110110	14	2.00	1.86	13.8	2.24	[1.88, 2.47]	12
1111000	1	2.86	1.48	0.9	2.24	[1.38, 2.69]	12
1111001	1	2.16	1.72	1.0	2.24	[1.91, 2.73]	12
1111010	0	NA	1.93	0.2	2.24	[1.38, 2.69]	12
1111011	0	NA	2.17	0.1	2.24	[1.97, 2.78]	12
1111110	1	0.62	2.30	0.7	2.24	[1.89, 2.69]	12
1111111	0	NA	2.55	1.3	2.24	[1.97, 2.87]	12

Table 5: Parameter estimates and their 95% bootstrap confidence intervals for the zidovudine NI-CBN model displayed in Figure 3. The estimates for the parameters μ_g are shown in Table 4.

Parameter	MLE	95% CI
θ_{41L}	0.42	[0.37, 0.46]
θ_{67N}	0.40	[0.35, 0.43]
θ_{69D}	0.17	[0.13, 0.24]
θ_{70R}	0.59	[0.53, 0.67]
θ_{210W}	0.69	[0.60, 0.75]
θ_{215Y}	0.88	[0.82, 0.93]
θ_{219Q}	0.65	[0.57, 0.74]
σ^2	0.33	[0.18, 0.41]
ε	0.047	[0.039, 0.059]

Table 6: Bootstrap analysis of the structural stability of the zidovudine NI-CBN model displayed in Figure 3. The entry with row index mutation e and colum index mutation f denotes the number of times the relation $e \prec f$ appeared as a cover relation (or equivalently, the edge $e \rightarrow f$ appeared in the graph of the Bayesian network model) among 100 bootstrap samples. Numbers in bold face indicate the presence of the corresponding edge in the optimal ML poset of Figure 3.

	41L	67N	69D	70R	210W	215Y	219Q
41L	0	3	7	0	37	**60**	1
67N	0	0	**77**	**72**	0	1	12
69D	0	0	0	0	0	0	1
70R	0	7	4	0	0	1	**94**
210W	1	0	6	0	0	0	1
215Y	34	0	5	1	**67**	0	0
219Q	0	1	14	2	0	0	0

Table 7: HIV PR genotype lattice for indinavir; see Figure 4.

	Data		Linear model	NI-CBN model			
g	n_g	$\bar{y}_g$	$\hat{\phi}_g$	$\hat{u}_g$	$\hat{\mu}_g$	95% CI	Block
000000	469	-0.06	0.11	546.2	-0.05	[-0.08, -0.03]	1
000001	109	0.65	0.56	137.7	0.59	[0.46, 0.75]	2
001000	23	1.08	0.51	32.3	1.07	[0.97, 1.13]	3
001100	79	1.02	0.81	132.4	1.07	[0.97, 1.13]	3
100000	60	0.88	0.53	110.4	1.08	[0.76, 1.19]	4
101000	18	1.51	0.93	18.6	1.16	[1.06, 1.50]	5
101100	43	0.95	1.23	50.0	1.16	[1.06, 1.50]	5
100001	70	1.04	0.97	74.1	1.20	[1.10, 1.38]	6
000011	63	1.10	0.93	95.0	1.23	[1.14, 1.30]	7
100011	61	1.29	1.35	58.6	1.35	[1.24, 1.47]	8
001001	30	1.09	0.96	21.3	1.40	[1.24, 1.48]	9
001011	14	1.49	1.33	12.5	1.40	[1.32, 1.50]	9
001101	64	1.26	1.26	71.1	1.40	[1.29, 1.48]	9
001111	8	1.10	1.63	17.6	1.40	[1.32, 1.50]	9
101001	26	1.35	1.37	11.7	1.53	[1.33, 1.68]	10
101011	20	1.31	1.75	9.8	1.53	[1.40, 1.68]	10
101101	30	1.43	1.67	27.0	1.53	[1.42, 1.68]	10
101111	3	1.43	2.05	6.5	1.53	[1.44, 1.68]	10
010000	9	1.18	0.63	15.1	1.99	[1.72, 2.13]	11
010001	9	1.26	1.07	7.4	1.99	[1.82, 2.24]	11
010011	0	NA	1.45	1.6	1.99	[1.83, 2.37]	11
011000	2	0.74	1.03	0.8	1.99	[1.72, 2.15]	11
011001	6	0.80	1.47	0.5	1.99	[1.82, 2.24]	11
011011	0	NA	1.85	0.1	1.99	[1.83, 2.37]	11
011100	8	1.16	1.33	4.0	1.99	[1.72, 2.15]	11
011101	9	1.65	1.77	2.9	1.99	[1.81, 2.24]	11
011111	4	1.46	2.15	1.0	1.99	[1.83, 2.37]	11
110000	1	-0.22	1.05	4.4	2.38	[2.15, 2.56]	12
110001	1	0.00	1.49	0.6	2.38	[2.15, 2.60]	12
110011	1	0.71	1.86	0.1	2.38	[2.15, 2.60]	12
111000	0	NA	1.45	0.6	2.38	[2.17, 2.56]	12
111100	0	NA	1.75	0.6	2.38	[2.15, 2.56]	12
111001	0	NA	1.89	0.3	2.61	[2.18, 2.68]	13
111011	0	NA	2.27	0.0	2.61	[2.18, 2.68]	13
111101	0	NA	2.19	0.1	2.61	[2.17, 2.68]	13
111111	0	NA	2.57	0.0	2.61	[2.18, 2.68]	13

Table 8: Parameter estimates and their 95% bootstrap confidence intervals for the zidovudine NI-CBN model displayed in Figure 4. The estimates for the parameters μ_g are shown in Table 7.

Parameter	MLE	95% CI
$\theta_{46\mathrm{I}}$	0.25	[0.23, 0.28]
$\theta_{48\mathrm{V}}$	0.03	[0.02, 0.04]
$\theta_{54\mathrm{V}}$	0.29	[0.26, 0.31]
$\theta_{82\mathrm{A}}$	0.74	[0.67, 0.79]
$\theta_{84\mathrm{V}}$	0.36	[0.32, 0.41]
$\theta_{90\mathrm{M}}$	0.38	[0.34, 0.41]
σ^2	0.12	[0.09, 0.13]
ε	0.077	[0.068, 0.085]

Table 9: Bootstrap analysis of the structural stability of the indinavir NI-CBN model displayed in Figure 4. The entry with row index mutation e and colum index mutation f denotes the number of times the relation $e \prec f$ appeared as a cover relation (or equivalently, the edge $e \to f$ appeared in the graph of the Bayesian network model) among 100 bootstrap samples. Numbers in bold face indicate the presence of the corresponding edge in the optimal ML poset of Figure 4.

	46I	48V	54V	82A	84V	90M
46I	0	2	0	0	4	0
48V	0	0	0	0	0	0
54V	0	1	0	**69**	0	0
82A	0	4	31	0	0	0
84V	0	0	0	0	0	0
90M	2	3	0	0	**81**	0

References

Altmann, A., N. Beerenwinkel, T. Sing, I. Savenkov, M. Däumer, R. Kaiser, S.-Y. Rhee, W. J. Fessel, R. W. Shafer, and T. Lengauer (2007): "Improved prediction of response to antiretroviral combination therapy using the genetic barrier to drug resistance." *Antivir Ther*, 12, 169–178.

Barlow, R. E., D. J. Bartholomew, J. M. Bremner, and H. D. Brunk (1972): *Statistical Inference under Order Restrictions*, John Wiley & Sons.

Beerenwinkel, N. and M. Drton (2007): "A mutagenetic tree hidden Markov model for longitudinal clonal HIV sequence data." *Biostatistics*, 8, 53–71, URL http://dx.doi.org/10.1093/biostatistics/kxj033.

Beerenwinkel, N., M. Däumer, M. Oette, K. Korn, D. Hoffmann, R. Kaiser, T. Lengauer, J. Selbig, and H. Walter (2003a): "Geno2pheno: Estimating phenotypic drug resistance from HIV-1 genotypes." *Nucleic Acids Res*, 31, 3850–3855.

Beerenwinkel, N., N. Eriksson, and B. Sturmfels (2006): "Evolution on distributive lattices." *J Theor Biol*, 242, 409–420, URL http://dx.doi.org/10.1016/j.jtbi.2006.03.013

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

Beerenwinkel, N., T. Lengauer, M. Däumer, R. Kaiser, H. Walter, K. Korn, D. Hoffmann, and J. Selbig (2003b): "Methods for optimizing antiviral combination therapies." *Bioinformatics*, 19 Suppl 1, i16–i25, iSMB 2003.

Beerenwinkel, N., J. Rahnenführer, M. Däumer, D. Hoffmann, R. Kaiser, J. Selbig, and T. Lengauer (2005): "Learning multiple evolutionary pathways from cross-sectional data." *J Comput Biol*, 12, 584–598, URL http://dx.doi.org/10.1089/cmb.2005.12.584

Beerenwinkel, N., B. Schmidt, H. Walter, R. Kaiser, T. Lengauer, D. Hoffmann, K. Korn, and J. Selbig (2002): "Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype." *Proc Natl Acad Sci U S A*, 99, 8271–8276, URL http://dx.doi.org/10.1073/pnas.112177799

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

Boucher, C. A., E. O'Sullivan, J. W. Mulder, C. Ramautarsing, P. Kellam, G. Darby, J. M. Lange, J. Goudsmit, and B. A. Larder (1992): "Ordered appearance of zidovudine resistance mutations during treatment of 18 human immunodeficiency virus-positive subjects." *J Infect Dis*, 165, 105–110.

Carlson, J. M., Z. L. Brumme, C. M. Rousseau, C. J. Brumme, P. Matthews, C. Kadie, J. I. Mullins, B. D. Walker, P. R. Harrigan, P. J. R. Goulder, and D. Heckerman (2008): "Phylogenetic dependency networks: inferring patterns of CTL escape and codon covariation in HIV-1 Gag." *PLoS Comput Biol*, 4, e1000225, URL http://dx.doi.org/10.1371/journal.pcbi.1000225.

de Leeuw, J., K. Hornik, and P. Mair (2009): "Isotone optimization in R: Pool-adjacent-violators algorithm (PAVA) and active set methods," *Journal of Statistical Software*, 32, 1–24, URL http://www.jstatsoft.org/v32/i05

Deforche, K., T. Silander, R. Camacho, Z. Grossman, M. A. Soares, K. V. Laethem, R. Kantor, Y. Moreau, and A.-M. Vandamme (2006): "Analysis of HIV-1 pol sequences using Bayesian networks: implications for drug resistance." *Bioinformatics*, 22, 2975–2979.

Dempster, A., N. Laird, and D. Rubin (1977): "Maximum likelihood from incomplete data via the EM algorithm (with discussions)," *J R Statist Soc B*, 39, 1–38.

Draghici, S. and R. B. Potter (2003): "Predicting HIV drug resistance with neural networks." *Bioinformatics*, 19, 98–107.

Gerstung, M., M. Baudis, H. Moch, and N. Beerenwinkel (2009): "Quantifying cancer progression with conjunctive Bayesian networks." *Bioinformatics*, 25, 2809–2815, URL http://dx.doi.org/10.1093/bioinformatics/btp505.

Iwasa, Y., F. Michor, and M. A. Nowak (2003): "Evolutionary dynamics of escape from biomedical intervention." *Proc Biol Sci*, 270, 2573–2578, URL http://dx.doi.org/10.1098/rspb.2003.2539

Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi (1983): "Optimization by simulated annealing." *Science*, 220, 671–680, URL http://dx.doi.org/10.1126/science.220.4598.671

Klingler, T. M. and D. L. Brutlag (1994): "Discovering structural correlations in alpha-helices." *Protein Sci*, 3, 1847–1857, URL http://dx.doi.org/10.1002/pro.5560031024.

Koelle, K., S. Cobey, B. Grenfell, and M. Pascual (2006): "Epochal evolution shapes the phylodynamics of interpandemic influenza A (H3N2) in humans." *Science*, 314, 1898–1903, URL http://dx.doi.org/10.1126/science.1132745.

Larder, B. A. (1994): "Interactions between drug resistance mutations in human immunodeficiency virus type 1 reverse transcriptase." *J Gen Virol*, 75 (Pt 5), 951–957.

Lozovsky, E. R., T. Chookajorn, K. M. Brown, M. Imwong, P. J. Shaw, S. Kamchonwongpaisan, D. E. Neafsey, D. M. Weinreich, and D. L. Hartl (2009): "Stepwise acquisition of pyrimethamine resistance in the malaria parasite." *Proc Natl Acad Sci U S A*, 106, 12025–12030, URL http://dx.doi.org/10.1073/pnas.0905922106

Poelwijk, F. J., D. J. Kiviet, D. M. Weinreich, and S. J. Tans (2007): "Empirical fitness landscapes reveal accessible evolutionary paths." *Nature*, 445, 383–386, URL http://dx.doi.org/10.1038/nature05451.

Poon, A. F. Y., F. I. Lewis, S. L. K. Pond, and S. D. W. Frost (2007): "An evolutionary-network model reveals stratified interactions in the V3 loop of the HIV-1 envelope." *PLoS Comput Biol*, 3, e231, URL http://dx.doi.org/10.1371/journal.pcbi.0030231.

Rabinowitz, M., L. Myers, M. Banjevic, A. Chan, J. Sweetkind-Singer, J. Haberer, K. McCann, and R. Wolkowicz (2006): "Accurate prediction of HIV-1 drug response from the reverse transcriptase and protease amino acid sequences using sparse models created by convex optimization." *Bioinformatics*, 22, 541–549, URL http://dx.doi.org/10.1093/bioinformatics/btk011

Reidys, C. M. and P. F. Stadler (2002): "Combinatorial landscapes," *SIAM Review*, 44, 3–54.

Rhee, S.-Y., M. J. Gonzales, R. Kantor, B. J. Betts, J. Ravela, and R. W. Shafer (2003): "Human immunodeficiency virus reverse transcriptase and protease sequence database." *Nucleic Acids Res*, 31, 298–303.

Rhee, S.-Y., J. Taylor, G. Wadhera, A. Ben-Hur, D. L. Brutlag, and R. W. Shafer (2006): "Genotypic predictors of human immunodeficiency virus type 1 drug resistance." *Proc Natl Acad Sci U S A*, 103, 17355–17360, URL http://dx.doi.org/10.1073/pnas.0607274103

Segal, M. R., J. D. Barbour, and R. M. Grant (2004): "Relating HIV-1 sequence variation to replication capacity via trees and forests," *Stat Appl Genet Mol Biol*, 3, 2.

Sevin, A. D., V. DeGruttola, M. Nijhuis, J. M. Schapiro, A. S. Foulkes, M. F. Para, and C. A. Boucher (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. W. and J. M. Schapiro (2008): "HIV-1 drug resistance mutations: an updated framework for the second decade of HAART," *AIDS Rev*, 10, 67–84.

van Nimwegen, E. (2006): "Influenza escapes immunity along neutral networks." *Science*, 314, 1884–1886, URL http://dx.doi.org/10.1126/science.1137300.

Wang, D. and B. Larder (2003): "Enhanced prediction of lopinavir resistance from genotype by use of artificial neural networks," *J Infect Dis*, 188, 653–660.

Weinreich, D. M., N. F. Delaney, M. A. Depristo, and D. L. Hartl (2006): "Darwinian evolution can follow only very few mutational paths to fitter proteins." *Science*, 312, 111–114, URL http://dx.doi.org/10.1126/science.1123539.