Noname manuscript No.
(will be inserted by the editor)

Modeling cumulative biological phenomena with Suppes-Bayes Causal Networks

Daniele Ramazzotti · Alex Graudenzi · Giulio Caravagna · Marco Antoniotti

Received: date / Accepted: date

Abstract Several diseases related to cell proliferation are characterized by the accumulation of somatic DNA changes, with respect to wildtype conditions. Cancer and HIV are two common examples of such diseases, where the mutational load in the cancerous/viral population increases over time. In these cases, selective pressures are often observed along with competition, cooperation and parasitism among distinct cellular clones. Recently, we presented a mathematical framework to model these phenomena, based on a combination of Bayesian inference and Suppes' theory of probabilistic causation, depicted in graphical structures dubbed Suppes-Bayes Causal Networks (SBCNs). SBCNs are generative probabilistic graphical models that recapitulate the potential ordering of accumulation of such DNA changes during the progression of the disease. Such models can be inferred from data by exploiting likelihood-based model-selection strategies with regularization. In this paper we discuss the theoretical foundations of our approach and we investigate in depth the influence on the model-selection task of: (*i*) the poset based on Suppes' theory and (*ii*) different regularization strategies. Furthermore, we provide an example of application of our framework to HIV genetic data highlighting the valuable insights provided by the inferred SBCN.

Keywords Cumulative Phenomena · Bayesian Graphical Models · Probabilistic Causality

Daniele Ramazzotti
Department of Pathology, Stanford University, Stanford, CA 94305, USA E-mail: daniele.ramazzotti@stanford.edu

Alex Graudenzi
Department of Informatics, Systems and Communication, University of Milan-Bicocca, Milan, Italy

Giulio Caravagna
School of Informatics, University of Edinburgh, Edinburgh, UK

Marco Antoniotti
Department of Informatics, Systems and Communication, University of Milan-Bicocca, Milan, Italy

1 Introduction

A number of diseases are characterized by the accumulation of genomic lesions in the DNA of a population of cells. Such lesions are often classified as *mutations*, if they involve one or few nucleotides, or *chromosomal alterations*, if they involve wider regions of a chromosome. The effect of these lesions, occurring randomly and inherited through cell divisions (i.e., they are *somatic*), is that of inducing a phenotypic change in the cells. If the change is advantageous, then the clonal population might enjoy a *fitness* advantage over competing clones. In some cases, a natural selection process will tend to select the clones with more advantageous and inheritable traits. This particular picture can be framed in terms of Darwinian evolution as a scenario of survival of the fittest where, however, the prevalence of multiple heterogenous populations is often observed [6].

Cancer and HIV are two diseases where the mutational[1] load in the cancerours/viral population of cells increases over time, and drives phenotypic switches and disease progression. In this paper we specifically focus on these desease, but many biological systems present similar characteristics, see [38, 26, 20].

The emergence and development of cancer can be characterized as an *evolutionary process* involving a large population of cells, heterogeneous both in their genomes

[1] From now on, we will use the term mutation to refer to the types of genomic lesions mentioned above.

and in their epigenomes. In fact, genetic and epigenetic random alterations commonly occurring in any cell, can occasionally be beneficial to the neoplastic cells and confer to these clones a *functional selective advantage*. During clonal evolution, clones are generally selected for increased proliferation and survival, which may eventually allow the cancer clones to outgrow the competing cells and, in turn, may lead to invasion and metastasis [23,21]. By means of such a multistep stochastic process, cancer cells acquire over time a set of biological capabilities, i.e., *hallmarks* [14,15]. However, not all the alterations are involved in this acquisition; as a matter of fact, in solid tumors, we observe an average of 33 to 66 genes displaying a somatic mutations [36]. But only some of them are involved in the hallmark acquisition, i.e., *drivers*, while the remaining ones are present in the cancer clones without increasing their *fitness*, i.e., *passengers* [36].

The onset of AIDS is characterized by the collapse of the immune system after a prolonged asymptomatic period, but its progression's mechanistic basis is still unknown. It was recently hypothesized that the elevated turnover of lymphocytes throughout the asymptomatic period results in the accumulation of deleterious mutations, which impairs immunological function, replicative ability and viability of lymphocytes [10,31]. The failure of the modern combination therapies (i.e., highly active antiretroviral therapy) of the disease is mostly due to the capability of the virus to escape from drug pressure by developing drug resistance. This mechanism is determined by HIV's high rates of replication and mutation. In fact, under fixed drug pressure, these mutations are virtually non-reversible because they confer a strong selective advantage to viral populations [25,35].

In the last decades huge technological advancements led to the development of *Next Generation Sequencing* (NGS) techniques. These allow, in different forms and with different technological characteristics, to read out genomes from single cells or bulk [22,37,12,11]. Thus, we can use these technologies to quantify the presence of mutations in a sample. With this data at hand, we can therefore investigate the problem of inferring a *Progression Model* (PM) that recapitulates the ordering of accumulation of mutations during disease origination and development [7]. This problem allows different formulation according to the type of diseases that we are considering, the type of NGS data that we are processing and other factors. We point the reader to [3,7] for a review on progression models.

This work is focused on a particular class of mathematical models that are becoming successful to represent such mutational ordering. These are called *Suppes-Bayes Causal Networks*[2] (SBCN, [5]), and derive from a more general class of models, Bayesian Networks (BN, [17]), that has been successfully exploited to model cancer and HIV progressions [8,2,13]. SBCNs are probabilistic graphical models that are derived within a statistical framework based on Patrick Suppes' theory of *probabilistic causation* [34]. Thus, the main difference between standard Bayesian Networks and SBCNs is the encoding in the model of a set of causal axioms that describe the accumulation process. Both SBCNs and BNs are generative probabilistic models that induces a distribution of observing a particular mutational signature in a sample. But, the distribution induced by a SBCN is also consistent with the causal axioms and, in general, is different from the distribution induced by a standard BN [28].

The motivation for adopting a causal framework on top of standard BNs is that, in the particular case of cumulative biological phenomena, SBCNs allow better inferential algorithms and data-analysis pipelines to be developed [19,28,7]. Extensive studies in the inference of cancer progression have indeed shown that model-selection strategies to extract SBCNs from NGS data achieve better performance than algorithms that infer BNs. In fact, SBCN's inferential algorithms have higher rate of detection of true positive ordering relations, and higher rate of filtering out false positive ones. In general, these algorithms also show better *scalability*, *resistance to noise in the data*, and *ability to work with datasets with few samples* (see, e.g., [19,28]).

In this paper we give a formal definition of SBCNs, we assess their relevance in modeling cumulative phenomena and investigate the influence of (a) Suppes' poset, and (b) distinct maximum likelihood regularization strategies for model-selection. We do this by carrying out extensive synthetic tests in operational settings that are representative of different possible types of progressions, and data harbouring signals from heterogenous populations.

2 Suppes-Bayes Causal Networks

In [34], Suppes introduced the notion of *prima facie causation*. A prima facie relation between a cause u and its effect v is verified when the following two conditions are true: (i) *temporal priority* (TP), i.e., any cause happens before its effect and (ii) *probability raising* (PR), i.e., the presence of the cause raises the probability of observing its effect.

[2] The first use of these networks appears in [28], and it's earliest formal definition in [5].

Definition 1 (Probabilistic causation, [34]) For any two events u and v, occurring respectively at times t_u and t_v, under the mild assumptions that $0 < P(u), P(v) < 1$, the event u is called a *prima facie cause* of v if it occurs *before* and *raises the probability* of u, i.e.,

$$\textsf{TP:} \quad t_u < t_v \tag{1}$$

$$\textsf{PR:} \quad P(v \mid u) > P(v \mid \overline{u}) \tag{2}$$

While the notion of prima facie causation has known limitations in the context of the general theories of causality [16], this formulation seems to intuitively characterizes the dynamics of phenomena driven by the *monotonic accumulation of events*. In these cases, in fact, a temporal order among the events is implied and, furthermore, the occurrence of an early event *positively correlates* to the subsequent occurrence of a later one. Thus, this approach seems appropriate to capture the notion of selective advantage emerging from somatic mutations that accumulate during, e.g., cancer or HIV progression.

Let us now consider a graphical representation of the aforementioned dynamics in terms of a Bayesian graphical model.

Definition 2 (Bayesian network, [17]) The pair $\mathcal{B} = \langle G, P \rangle$ is a *Bayesian Network (BN)*, where G is a *directed acyclic graph (DAG)* $G = (V, E)$ of V nodes and E arcs, and P is a distribution induced over the nodes by the graph. Let $V = \{v_1, \ldots, v_n\}$ be *random variables* and the *edges/arcs* $E \subseteq V \times V$ encode the conditional dependencies among the variables. Define, for any $v_i \in V$, the *parent set* $\pi(v_i) = \{x \mid x \to v_i \in E\}$, then P defines the *joint probability distribution* induced by the BN as follow:

$$P(v_1, \ldots, v_n) = \prod_{v_i \in V} P(v_i \mid \pi(v_i)) \ . \tag{3}$$

All in all, a BN is a statistical model which succinctly represents the conditional dependencies among a set of *random variables* V through a DAG. More precisely, a BN represents a factorization of the *joint distribution* $P(v_1, \ldots, v_n)$ in terms of marginal (when $\pi(v) = \emptyset$) and conditional probabilities $P(\cdot \mid \cdot)$.

We now consider a common situation when we deal with data (i.e., observations) obtained at one (or a few) points in time, rather than through a timeline. In this case, we are resticted to work with cross-sectional data, where no explicit information of time is provided. Therefore, we can model the dynamics of cumulative phenomena by means of a specific set of the general BNs where the nodes V represent the accumulating events as *Bernoulli random variables* taking values in $\{0, 1\}$ based on their occurrence: the value of the variable is 1 if the event is observed and 0 otherwise. We then define a dataset D of s cross-sectional samples over n Bernoulli random variables as follow.

$$\begin{array}{c} \\ s_1 \\ s_2 \\ \vdots \\ s_m \end{array} \begin{array}{c} \begin{array}{cccc} v_1 & v_2 & \cdots & v_n \end{array} \\ \begin{pmatrix} d_{1,1} & d_{1,2} & \ldots & d_{1,n} \\ d_{2,1} & d_{2,2} & \ldots & d_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{m,1} & d_{m,2} & \ldots & d_{m,n} \end{pmatrix} \end{array} = D \tag{4}$$

In order to extend BNs to account for Suppes' theory of probabilistic causation we need to estimate for any variable $v \in V$ its timing t_v. Because we are dealing with cumulative phenomena and, in the most general case, data that do not harbour any evidente temporal information[3], we can use the *marginal probability* $P(v)$ as a proxy for t_v (see also the commentary at the end of this section). In cancer and HIV, for instance, this makes sense since mutations are inherited through cells divisions, and thus will fixate in the clonal populations during disease progression, i.e., they are persistent.

Definition 3 (Suppes-Bayes Causal Network, [5]) A BN $\mathcal{B}$ is a SBCN if and only if, for any edge $v_i \to v_j \in E$, Suppes' conditions (Definition 1) hold, that is

$$P(v_i) > P(v_j) \quad \text{and} \quad P(v_j \mid v_i) > P(v_j \mid \neg v_i) \ . \tag{5}$$

It should be noted that SBCNs and BNs *have the same likelihood function*. Thus, SBCNs *do not embed any constraint of the cumulative process in the likelihood computation*, while approaches based on cumulative BNs do [13]. Instead, the structure of the model, E, is consistent with the causal model à-la-Suppes and, of course, this in turn reflects in the induced distribution. Even though this difference seem subtle, this is arguably the most interesting advantage of SBCNs over ad-hoc BNs for cumulative phenomena.

Model-selection to infer a network from data. The structure G of a BN (or of a SBCN) can be inferred from a data matrix D, as well as the parameters of the conditional distributions that define P. The model-selection task is that of inferring such information from data; in general, we expect different models (i.e., edges) if we infer a SBCN or a BN, as SBCNs encode Suppes' additional constraints.

[3] In many cases, the data that we can access are *cross-sectional*, meaning that the samples are collected at independent and unknown time-points. For this reason, we have to resort on the simplest possible approach to estimate timings. However, in the case we were provided with explicit observations of time, the temporal priority would be directly and, yet, more efficiently assessable.

The general structural learning, i.e., *the model-selection problem*, for BNs is NP-HARD [17], hence one needs to resort on approximate strategies. For each BN $\mathcal{B}$ a *log-likelihood function* $\mathcal{L}(D \mid E)$ can be used to search in the space of structures (i.e., the set of edges E), together with a regularization function $\mathcal{R}(\cdot)$ that penalizes overly complicated models. The network's structure is then inferred by solving the following optimization problem

$$E_* = \arg\max_E \left[\mathcal{L}(D|E) - \mathcal{R}(E)\right]. \tag{6}$$

Moreover, the parameters of the conditional distributions can be computed by *maximum-likelihood estimation* for the set of edges E_*; the overall solution is *locally optimal* [17].

Model-selection for SBCNs works in this very same way, but constrains the search for valid solutions (see, e.g., [28]). In particular, it scans only the subset of edges that are consistent with Definition 1 – while a BN search will look for the full $V \times V$ space. To filter pairs of edges, Suppes' conditions can be estimated from the data with solutions based, for instance, on bootstrap estimates [28]. The resulting model will satisfy, by construction, the conditions of probabilistic causation. It has been shown that if the underlying phenomenon that produced our data is characterised by an accumulation, then the inference of a SBCN, rather than a BN, leads to *much better precision and recall* [19,28].

We conclude this Section by discussing in details the characteristics of the SBCNs and, in particular, to which extent they are capable of modeling cumulative phenomena.

Temporal priority. Suppes' first constraint ("event u is temporally preceding event v", Definition 1) assumes an underlying *temporal (partial) order* $\sqsubseteq_{\mathsf{TP}}$ among the events/ variables of the SBCN, that we need to compute.

Cross-sectional data, unfortunately, are not provided with an explicit measure of time and hence $\sqsubseteq_{\mathsf{TP}}$ needs to be estimated from data D[4]. The cumulative nature of the phenomenon that we want to model leads to a simple estimation of $\sqsubseteq_{\mathsf{TP}}$: *the temporal priority TP is assessed in terms of marginal frequencies [28],*

$$v_j \sqsubseteq_{\mathsf{TP}} v_i \iff P(v_i) > P(v_j)\ . \tag{7}$$

Thus, more frequent events, i.e., v_i, are assumed to occur earlier, which is sound when we assume the accumulating events to be irreversible.

[4] We notice that in the case we were provided with explicit observations of time, $\sqsubseteq_{\mathsf{TP}}$ would be directly and, yet, more efficiently assessable.

TP is combined with probability raising to complete Suppes' conditions for prima facie (see below). Its contribution is fundamental for model-selection, as we now elaborate.

First of all, recall that the model-selection problem for BNs is in general NP-HARD [17], and that, as a result of the assessment of Suppes' conditions (TP *and* PR), we constrain our search space to the networks with a given order. Because of time irreversibility, marginal distributions induce a total ordering $\sqsubseteq_{\mathsf{TP}}$ on the v_i, i.e., reflexing $\leq$. Learning BNs given a fixed order $\sqsubseteq$ – even partial [17] – of the variables bounds the cardinality of the parent set as follows

$$|\pi(v_x)| \leq |\{v_j \mid v_x \sqsubseteq v_j\}|\,, \tag{8}$$

and, in general, it make inference easier than the general case [17]. Thus, by constraining Suppes' conditions via $\sqsubseteq_{\mathsf{TP}}$'s total ordering, we drop down the model-selection complexity. It should be noted that, after model-selection, the ordering among the variables that we practically have in the selected arcs set E is in general partial; in the BN literature this is sometimes called *poset*.

Probability raising. Besides TP, as a second constraint we further require that the arcs are consistent to the condition of PR: this leads to another relation $\sqsubseteq_{\mathsf{PR}}$. Probability raising is equivalent to constraining for *positive statistical dependence* [19]

$$\begin{aligned} v_j \sqsubseteq_{\mathsf{PR}} v_i & \\ &\iff P(v_j \mid v_i) > P(v_j \mid \overline{v}_i) \\ &\iff P(v_i, v_i) > P(v_i)\, P(v_j)\ , \end{aligned} \tag{9}$$

thus we model all and only the positive dependant relations. Definition 1 is thus obtained by selecting those probability raising relations that are consistent with TP,

$$\sqsubseteq_{\mathsf{TP}} \cap \sqsubseteq_{\mathsf{PR}}\ , \tag{10}$$

as the core of Suppes' characterization of causation is relevance [34].

If $\sqsubseteq_{\mathsf{TP}}$ reduces the search space of the possible valid structures for the network by setting a specific total order to the nodes, $\sqsubseteq_{\mathsf{PR}}$ instead reduces the search space of the possible valid parameters of the network by requiring that the related conditional probability tables, i.e., $P(\cdot)$, account only for positive statistical dependencies. It should be noted that these constraints affect the structure and the parameters of the model, but the likelihood-function is the same for BNs and SBCNs.

Network simplification, regularization and spurious causality. Suppes' criteria are known to be necessary but not sufficient to evaluate general causal claims [34]. Even if we restrict to causal cumulative phenomena, the expressivity of probabilistic causality needs to be taken into account.

When dealing with small sample sized datasets (i.e., small m), many pairs of variables that satisfy Suppes' condition may be *spurious causes*, i.e., *false positive*[5]. *False negatives* should be few, and mostly due to noise in the data. Thus, it follows

- we expect all the "statistically relevant" relations to be also prima facie [28];
- we need to filter out spurious causality instances[6], as we know that prima facie overfits .

A model-selection strategy which exploits a regularization schema seems thus the best approach to the task. Practically, this strategy will select simpler (i.e., *sparse*) models according to a *penalized likelihood fit criterion* – for this reason, it will filter out edges proportionally to how much the regularization is stringent. Also, it will rank spurious association according to a criterion that is consistent with Suppes' intuition of causality, as likelihood relates to statistical (in)dependence. Alternatives based on likelihood itself, i.e., without regularization, do not seem viable to minimize the effect of likelihood's overfit, that happens unless $m \to \infty$ [17]. In fact, one must recall that due to statistical noise and sample size, exact statistical (in)dependence between pair of variables is never (or very unlikely) observed.

2.1 Modeling heterogeneous populations

Complex biological processes, e.g., *proliferation, nutrition, apoptosis*, are orchestrated by multiple co-operative networks of proteins and molecules. Therefore, different "mutants" can evade such control mechanisms in different ways. Mutations happen as a random process that is unrelated to the relative fitness advantage that they confer to a cell. As such, different cells will deviate from wildtype by exhibiting different mutational signatures during disease progression. This has an implication for many cumulative disease that arise from populations that are *heterogeneous*, both at the level of the single patient (intra-patient heterogeneity) or in the population of patients (inter-patient heterogeneity). Heterogeneity introduces significant challenges in designing effective treatment strategies, and major efforts are ongoing at deciphering its extent for many diseases such as cancer and HIV[2,28,7].

We now introduce a class of mathematical models that are suitable at modeling heterogenous progressions. These models are derived by augmenting BNs with logical formulas, and are called *Monotonic Progression Networks* (MPNs) [9,18]. MPNs represent the progression of events that accumulate *monotonically*[7] over time, where the conditions for any event to happen is described by a probabilistic version of the canonical boolean operators, i.e., conjunction ($\wedge$), inclusive disjunction ($\vee$), and exclusive disjunction ($\oplus$).

Following [9,18], we define one type of MPNs for each boolean operator: the *conjunctive* (CMPN), the *disjunctive semi-monotonic* (DMPN), and the *exclusive disjunction* (XMPN). The operator associated with each network type refers to the logical relation among the parents that eventually lead to the common effect to occur.

Definition 4 (Monotonic Progression Networks, [9,18]) MPNs are BNs that, for $\theta, \epsilon \in [0,1]$ and $\theta \gg \epsilon$, satisfy the conditions shown in Table 1 for each $v \in V$.

Here, θ represents the conditional probability of any "effect" to follow its preceding "cause" and ϵ models the probability of any noisy observation – that is the observation of a sample where the effects are observed without their causes. Note that the above inequalities define, for each type of MPN, specific constraints to the induced distributions. These are sometimes termed, according to the probabilistic logical relations, *noisy-AND*, *noisy-OR* and *noisy-XOR* networks [24,18].

Model-selection with heterogeneous populations. When dealing with heterogeneous populations, the task of model selection, and, more in general, any statistical analysis, is non-trivial. One of the main reason for this state of affairs is the emergence of statistical paradoxes such as Simpson's paradox [39,32]. This phonomenon refers to the fact that sometimes, associations among dichotomous variables, which are similar within subgroups of a population, e.g., females and males, change their statistical trend if the individuals of the subgroups are pooled together. Let us know recall a famous example to this regard. The admissions of the University of

[5] An edge is spurious when it satisfies Definition 1, but it is not actually the true model edge. For instance, for a linear model $u \to v \to w$, transitive edge $u \to w$ is spurious. Small m induces further spurious associations in the data, not necessarily related to particular topological structures.

[6] A detailed account of these topics, the particular types of spurious structures and their interpretation for different types of models are available in [19,28,27].

[7] The events accumulate over time and when later events occur earlier events are observed as well.

CMPN :	$P\left(v \mid \sum \pi(v) = \lvert\pi(v)\rvert\right) = \theta$	$P\left(v \mid \sum \pi(v) < \lvert\pi(v)\rvert\right) \leq \epsilon\,,$	(11)
DMPN :	$P\left(v \mid \sum \pi(v) > 0\right) = \theta$	$P\left(v \mid \sum \pi(v) = 0\right) \leq \epsilon\,,$	(12)
XMPN :	$P\left(v \mid \sum \pi(v) = 1\right) = \theta$	$P\left(v \mid \sum \pi(v) \neq 1\right) \leq \epsilon\,.$	(13)

Table 1: Definitions for CMPN, DMPN, and XMPN.

Berkeley for the fall of 1973 showed that men applying were much more likely than women to be admitted with a difference that was unlikely to be due to chance. But, when looking at the individual departments separately, it emerged that 6 out of 85 were indead biased in favor of women, while only 4 presented a slighly bias against them. The reason for this inconsistency was due to the fact that women tended to apply to competitive departments which had low rates of admissions while men tended to apply to less-competitive departments with high rates of admissions, leading to an apparent bias toward them in the overall population [4].

Similar situations may arise in cancer when different populations of cancer samples are mixed. As an example, let us consider an hypothetical progression leading to the alteration of gene e. Let us now assume that the alterations of this gene may be due to the previous alterations of either gene c_1 or gene c_2 exclusively. If this was the case, then we would expect a significant pattern of selective advantage from any of its causes to e if we were able to stratify the patients accordingly to either alteration c_1 or c_2, but we may lose these associations when looking at all the patients together.

In [28], the notion of *progression pattern* is introduced to describe this situation, defined as a boolean relation among all the genes, members of the parent set of any node as the ones defined by MPNs. To this extent, the authors extend Suppes' definition of prima facie causality in order to account for such patterns rather than for relations among atomic events as for Definition 1. Also, they claim that general MPNs can be learned in polynomial time provided that the dataset given as input is *lifted* [28] with a Bernoulli variable per causal relation representing the logical formula involving any parent set.

Following [28,27], we now consider any formula in *conjunctive normal form* (CNF)

$$\varphi = \boldsymbol{c}_1 \wedge \ldots \wedge \boldsymbol{c}_n,$$

where each $\boldsymbol{c}_i$ is a *disjunctive clause* $\boldsymbol{c}_i = c_{i,1} \vee \ldots \vee c_{i,k}$ over a set of literals and each literal represents an event (a Boolean variable) or its negation. By following analogous arguments as the ones used before, we can extend Definition 1 as follows.

Definition 5 (CNF probabilistic causation, [28, 27]) For any CNF formula φ and e, occurring respectively at times t_φ and t_e, under the mild assumptions that $0 < P(\varphi), P(e) < 1$, φ is a *prima facie cause* of e if

$$t_\varphi < t_e \quad \text{and} \quad P(e \mid \varphi) > P(e \mid \overline{\varphi})\ . \tag{14}$$

Given these premises, we can now define the Extended Suppes-Bayes Causal Networks, an extension of SBCNs which allows to model heterogeneity as defined probabilistically by MPNs.

Definition 6 (Extended Suppes-Bayes Causal Network) A BN $\mathcal{B}$ is an Extended SBCN if and only if, for any edge $\varphi_i \to v_j \in E$, Suppes' generalized conditions (Definition 5) hold, that is

$$P(\varphi_i) > P(v_j) \quad \text{and} \quad P(v_j \mid v_\varphi) > P(v_j \mid \neg v_\varphi)\ . \tag{15}$$

3 Evaluation on simulated data

We now evaluate the performance of the inference of Suppes-Bayes Causal Network on simulated data, with specific attention on the impact of the constraints based on Suppes' probabilistic causation on the overall performance. All the simulations are performed with the following settings.

We consider 6 different topological structures: the first two where any node has at the most one predecessor, i.e., (*i*) trees, (*ii*) forests, and the others where we set a limit of 3 predecessors and, hence, we consider (*iii*) directed acyclic graphs with a single source and conjunctive parents, (*iv*) directed acyclic graphs with multiple sources and conjunctive parents, (*v*) directed acyclic graphs with a single source and disjunctive parents, (*vi*) directed acyclic graphs with multiple sources and disjunctive parents. For each of these configurations, we generate 100 random structures.

Moreover, we consider 4 different sample sizes (50, 100, 150 and 200 samples) and 9 noise levels (i.e., probability of a random entry for the observation of any node in a sample) from 0% to 20% with step 2.5%. Furthermore, we repeat the above settings for networks of 10

and 15 nodes. Any configuration is then sampled 10 times independently, for a total of more than 4 million distinct simulated datasets.

Finally, the inference of the structure of the SBCN is performed using the algorithm proposed in [28] and the performance is assessed in terms of: $accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)}$, $sensitivity = \frac{TP}{(TP+FN)}$ and $specificity = \frac{TN}{(FP+TN)}$ with TP and FP being the true and false positive (we define as positive any arc that is present in the network) and TN and FN being the true and false negative (we define negative any arc that is not present in the network). All these measures are values in $[0, 1]$ with results close to 1 indicators of good performance.

In Figures 1, 2 and 3 we show the performance of the inference on simulated datasets of 100 samples and networks of 15 nodes in terms of accurancy, sensitivity and specificity for different settings which we discuss in details in the next Paragraphs.

Suppes' prima facie conditions are necessary but not sufficient We first discuss the performance by applying *only* the prima facie criteria and we evaluate the obtained prima facie network in terms of accurancy, sensitivity and specificity on simulated datasets of 100 samples and networks of 15 nodes (see Figures 1, 2 and 3). As expected, the sensitivity is much higher than the specificity implying the significant impact of false positives rather than false negatives for the networks of the prima facie arcs. This result is indeed expected being Suppes' criteria mostly capable of removing some of the arcs which do not represent valid causal relations rather than asses the exact set of valid arcs. Interestingly, the false negatives are still limited even when we consider DMPN, i.e., when we do not have guarantees for the algoritm of [28] to converge. The same simulations with different sample sizes (50, 150 and 200 samples) and on networks of 10 nodes present a similar trend (results not shown here).

The likelihood score overfits the data In Figures 1, 2 and 3 we also show the performance of the inference by likelihood fit (without any regularizator) on the prima facie network in terms of accurancy, sensitivity and specificity on simulated datasets of 100 samples and networks of 15 nodes. Once again, in general, sensitivity is much higher than specificity implying also in this case a significant impact of false positives rather than false negatives for the inferred networks. These results make explicit the need for a regularization heuristic when dealing with real (not infinite) sample sized datasets as discussed in the next paragraph. Another interesting consideration comes from the observation that the prima facie networks and the networks inferred via likelihood fit without regularization seem to converge to the same performance as the noise level increases. This is due to the fact that, in general, the prima facie constraints are very conservative in the sense that false positives are admitted as long as false negatives are limited. When the noise level increases, the positive dependencies among nodes are generally reduced and, hence, less arcs pass the prima facie cut for positive dependency. Also in this case, the same simulations with different sample sizes (50, 150 and 200 samples) and on networks of 10 nodes present a similar trend (results not shown here).

Model-selection with different regularization strategies We now investigate the role of different regularizations on the performance. In particular, we consider two commonly used regularizations: (*i*) the *Bayesian information criterion* (BIC) [30] and (*ii*) the *Akaike information criterion* (AIC) [1].

Although BIC and AIC are both scores based on maximum likelihood estimation and a penalization term to reduce overfitting, yet with distinct approaches, they produce significantly different behaviors. More specifically, BIC assumes the existence of one *true* statistical model which is generating the data while AIC aims at finding the best approximating model to the unknown data generating process. As such, BIC may likely underfit, whereas, conversely, AIC might overfit[8].

The performance on simulated datasets are shown in Figure 1, 2, 3. In general, the performance is improved in all the settings with both regularizators, as they succeed in shrinking toward sparse networks.

Furthermore, we observe that the performance obtained by SBCNs is still good even when we consider simulated data generated by DMPN. Although in this case we do not have any guarantee of convergence, in practice the algorithm seems efficient in approximating the generative model. In conclusion, without any further input, SBCNs can model CMPNs and, yet, depict the more significant arcs of DMPNs. To infer XMPN, the dataset needs to be lifted [28].

The same simulations with different sample sizes (50, 150 and 200 samples) and on networks of 10 nodes present a similar trend (results not shown here).

[8] Thus, BIC tends to make a trade-off between the likelihood and model complexity with the aim of inferring the statistical model which generates the data. This makes it useful when the purpose is to detect the best model describing the data. Instead, asymptotically, minimizing AIC is equivalent to minimizing the cross validation value [33]. It is this property that makes the AIC score useful in model selection when the purpose is prediction. Overall, the choise of the regularizator tunes the level of sparsity of the retrieved SBCN and, yet, the confidence of the inferred arcs.

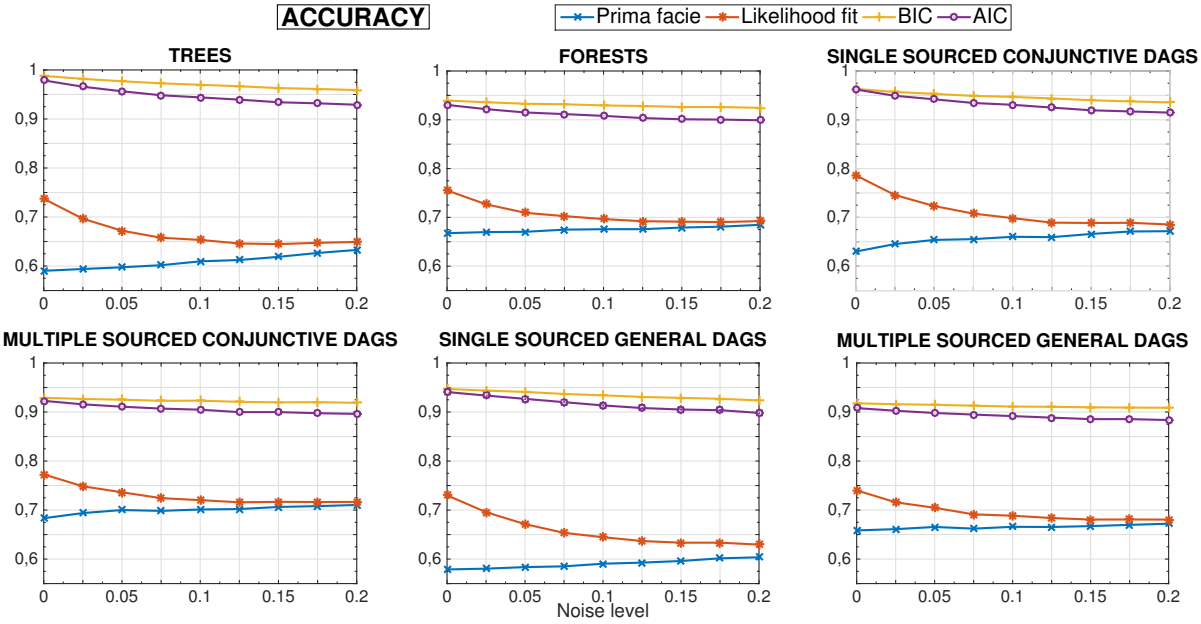

Fig. 1: Performance of the inference on simulated datasets of 100 samples and networks of 15 nodes in terms of accurancy for the 6 considered topological structures. The y axis refers to the performance while the x axis represent the different noise levels.

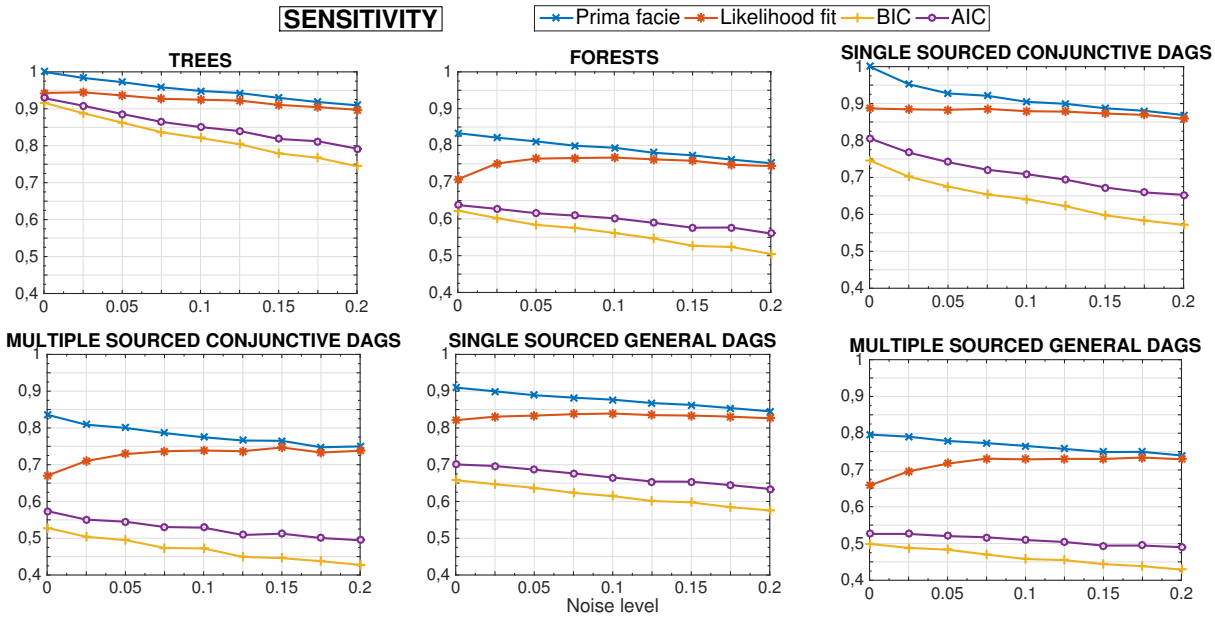

Fig. 2: Performance of the inference on simulated datasets of 100 samples and networks of 15 nodes in terms of sensitivity for the 6 considered topological structures. The y axis refers to the performance while the x axis represent the different noise levels.

4 Application to HIV genetic data

We now present an example of application of our framework on HIV genomic data. In particular we study drug resistance in patients under antiretroviral therapy and we select a set of 7 amino acid alterations in the HIV genome to be depicted in the resulting graphical model, namely $K20R$, $M36I$, $M46I$, $I54V$, $A71V$, $V82A$, $I84V$, where, as an example, the genomic event $K20R$ describes a mutation from lysine (K) to arginine (R) at position 20 of the HIV protease.

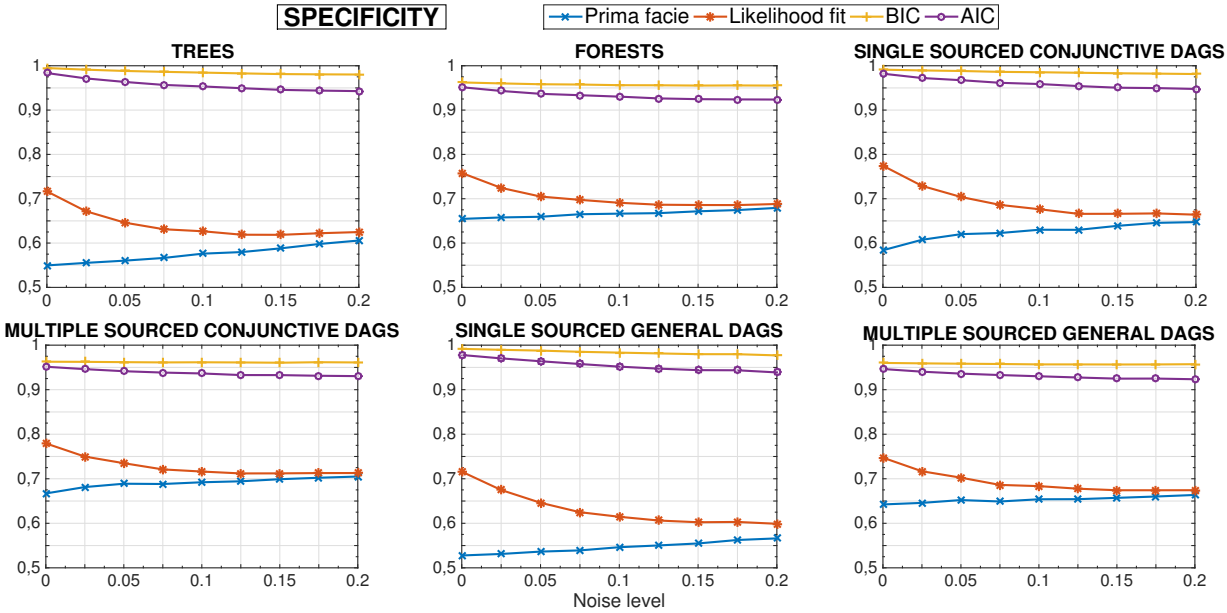

Fig. 3: Performance of the inference on simulated datasets of 100 samples and networks of 15 nodes in terms of specificity for the 6 considered topological structures. The y axis refers to the performance while the x axis represent the different noise levels.

In this study, we consider datasets from the Stanford HIV Drug Resistance Database (see, [29]) for 2 protease inhibitors, ritonavir (RTV) and indinavir (IDV). The first dataset consists of 179 samples (see Figure 4) and the second of 1035 samples (see Figure 4).

We then infer a model on these datasets by both Bayesian Network and Suppes-Bayes Causal Network. We show the results in Figures 5 where each node represent a mutation and the scores on the arcs measure the confidence in the found relation by non-parametric bootstrap.

In this case, it is interesting to observe that the set of dependency relations (i.e., any pair of nodes connected by an arc, without considering its direction) depicted both by SBCNs and BNs are very similar, with the main different being in the direction of some connection. This difference is expected and can be attributed to the constrain of temporal priority adopted in the SBCNs. Furthermore, we also observe that most of the found relations in the SBCN are more confidence (i.e., higher bootstrap score) than the one depicted in the related BN, leading us to observe a higher statistical confidence in the models inferred by SBCNs.

5 Conclusions

In this work we investigated the properties of a constrained version of Bayesian network, named SBCN, which is particularly sound in modeling the dynamics of system driven by the monotonic accumulation of events, thanks to encoded poset based on Suppes' theory of probabilistic causation. In particular, we showed how SBCNs can, in general, describe different types of MPN, which makes them capable of characterizing a broad range of cumulative phenomena not limited to cancer evolution and HIV drug resistance.

Besides, we investigated the influence of Suppes' poset on the inference performance with cross-sectional synthetic datasets. In particular, we showed that Suppes' constraints are effective in defining a partially order set accounting for accumulating events, with very few false negatives, yet many false positives. To overcome this limitation, we explored the role of two maximum likelihood regularization parameters, i.e., BIC and AIC, the former being more suitable to test previously conjectured hypotheses and the latter to predict novel hypotheses.

Finally, we showed on a dataset of HIV genomic data how SBCN can be effectively adopted to model cumulative phenomena, with results presenting a higher statistical significance compared to standard BNs.

Acknowledgements This work has been partially supported by grants from the SysBioNet project, a MIUR initiative for the Italian Roadmap of European Strategy Forum on Research Infrastructures (ESFRI).

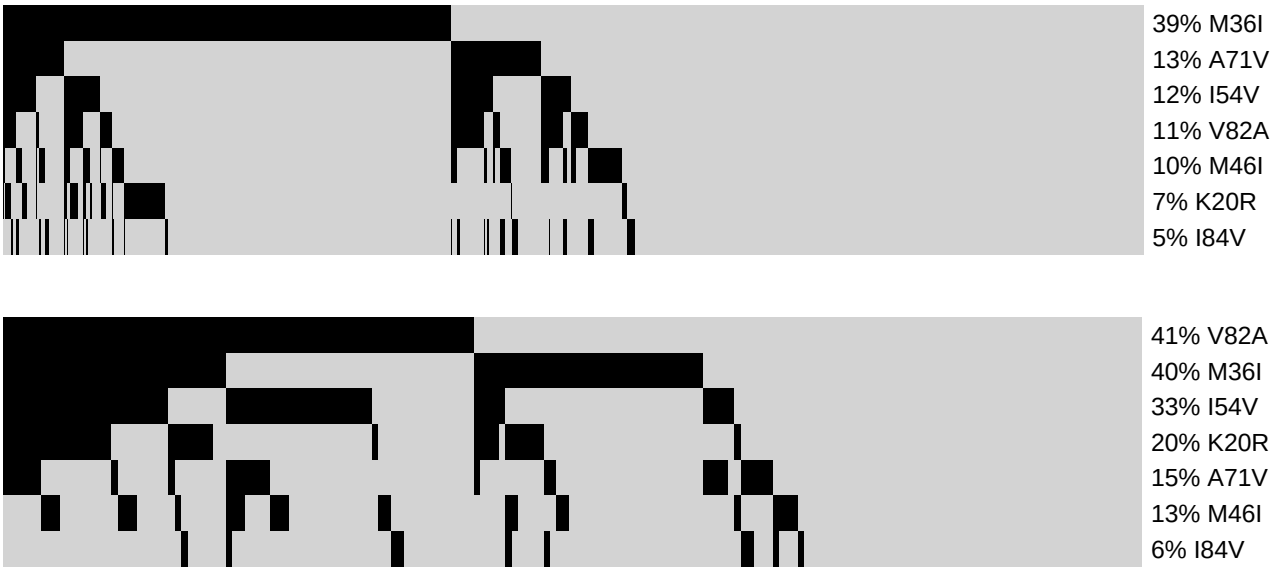

Fig. 4: Mutations detected in the genome for 179 HIV patients under ritonavir (top), and 1035 under indinavir (bottom). Each black rectangle denotes the presence of a mutation in the gene annotated to the right of the plot; percentages correspond to marginal probabilities.

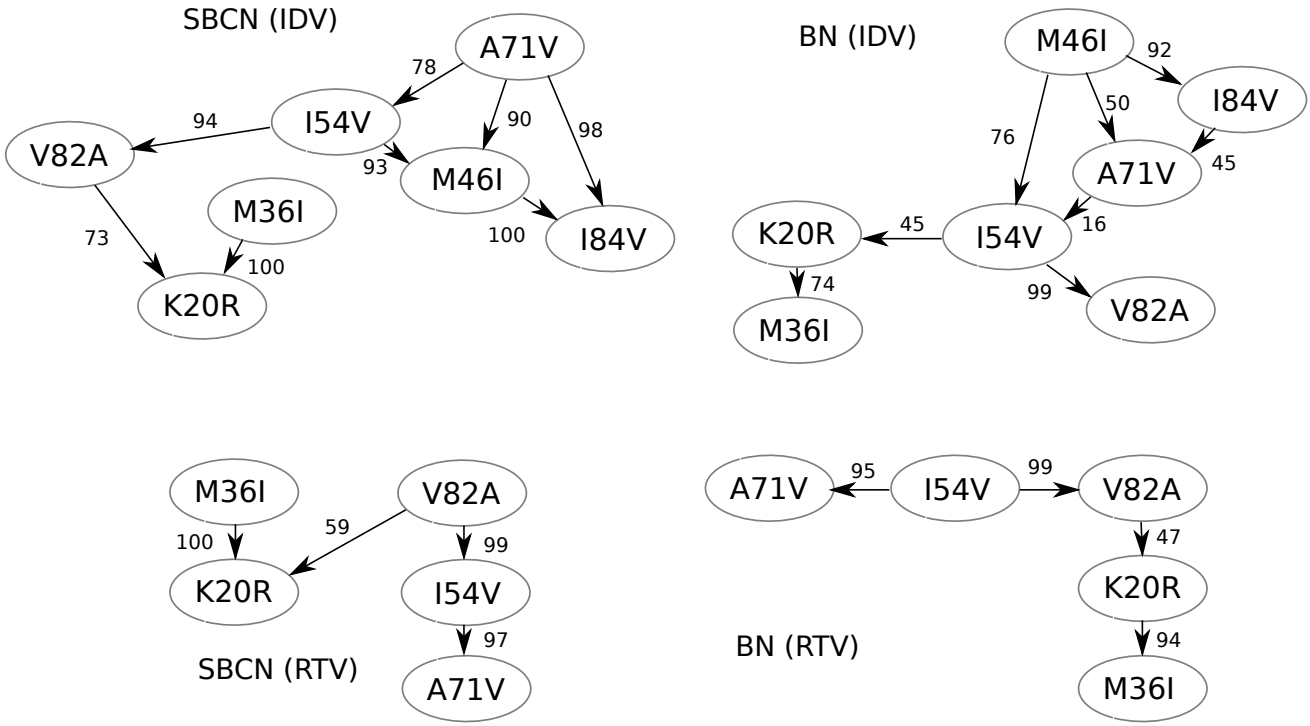

Fig. 5: HIV progression of patients under ritonavir or indinavir, (Figure 4) described as a Bayesian Network or as a Suppes-Bayes Causal Network. Edges are annotated with non parametric bootstrap scores.

References

1. Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: Selected Papers of Hirotugu Akaike, pp. 199–213. Springer (1998)
2. Beerenwinkel, N., Eriksson, N., Sturmfels, B.: Conjunctive bayesian networks. Bernoulli pp. 893–909 (2007)
3. Beerenwinkel, N., Schwarz, R.F., Gerstung, M., Markowetz, F.: Cancer evolution: mathematical models

and computational inference. Systematic biology **64**(1), e1–e25 (2015)
4. Bickel, P.J., Hammel, E.A., OConnell, J.W., et al.: Sex bias in graduate admissions: Data from berkeley. Science **187**(4175), 398–404 (1975)
5. Bonchi, F., Hajian, S., Mishra, B., Ramazzotti, D.: Exposing the probabilistic causal structure of discrimination. Submitted. Available on arXiv.org. (2015)
6. Burrell, R.A., McGranahan, N., Bartek, J., Swanton, C.: The causes and consequences of genetic heterogeneity in cancer evolution. Nature **501**(7467), 338–345 (2013)
7. Caravagna, G., Graudenzi, A., Ramazzotti, D., Sanz-Pamplona, R., De Sano, L., Mauri, G., Moreno, V., Antoniotti, M., Mishra, B.: Algorithmic methods to infer the evolutionary trajectories in cancer progression. PNAS **113**(28), E4025–E4034 (2016)
8. Desper, R., Jiang, F., Kallioniemi, O.P., Moch, H., Papadimitriou, C.H., Schäffer, A.A.: Inferring tree models for oncogenesis from comparative genome hybridization data. Journal of computational biology **6**(1), 37–51 (1999)
9. Farahani, H.S., Lagergren, J.: Learning oncogenetic networks by reducing to mixed integer linear programming. PloS one **8**(6), e65,773 (2013)
10. Galvani, A.P.: The role of mutation accumulation in hiv progression. Proceedings of the Royal Society of London B: Biological Sciences **272**(1574), 1851–1858 (2005)
11. Gerlinger, M., Horswell, S., Larkin, J., Rowan, A.J., Salm, M.P., Varela, I., Fisher, R., McGranahan, N., Matthews, N., Santos, C.R., et al.: Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. Nature genetics **46**(3), 225–233 (2014)
12. Gerlinger, M., Rowan, A.J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., et al.: Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. New England journal of medicine **366**(10), 883–892 (2012)
13. Gerstung, M., Baudis, M., Moch, H., Beerenwinkel, N.: Quantifying cancer progression with conjunctive bayesian networks. Bioinformatics **25**(21), 2809–2815 (2009)
14. Hanahan, D., Weinberg, R.A.: The hallmarks of cancer. cell **100**(1), 57–70 (2000)
15. Hanahan, D., Weinberg, R.A.: Hallmarks of cancer: the next generation. cell **144**(5), 646–674 (2011)
16. Hitchcock, C.: Probabilistic causation. In: E.N. Zalta (ed.) The Stanford Encyclopedia of Philosophy, winter 2012 edn. (2012)
17. Koller, D., Friedman, N.: Probabilistic graphical models: principles and techniques. MIT press (2009)
18. Korsunsky, I., Ramazzotti, D., Caravagna, G., Mishra, B.: Inference of cancer progression models with biological noise. arXiv preprint arXiv:1408.6032 (2014)
19. Loohuis, L.O., Caravagna, G., Graudenzi, A., Ramazzotti, D., Mauri, G., Antoniotti, M., Mishra, B.: Inferring tree causal models of cancer progression with probability raising. PLoS One **9**(10) (2014)
20. Lozovsky, E.R., Chookajorn, T., Brown, K.M., Imwong, M., Shaw, P.J., Kamchonwongpaisan, S., Neafsey, D.E., Weinreich, D.M., Hartl, D.L.: Stepwise acquisition of pyrimethamine resistance in the malaria parasite. Proceedings of the National Academy of Sciences **106**(29), 12,025–12,030 (2009)
21. Merlo, L.M., Pepper, J.W., Reid, B.J., Maley, C.C.: Cancer as an evolutionary and ecological process. Nature Reviews Cancer **6**(12), 924–935 (2006)
22. Navin, N.E.: Cancer genomics: one cell at a time. Genome biology **15**(8), 1 (2014)
23. Nowell, P.C.: The clonal evolution of tumor cell populations. Science **194**(4260), 23–28 (1976)
24. Pearl, J.: Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann (2014)
25. Perrin, L., Telenti, A.: Hiv treatment failure: testing for hiv resistance in clinical practice. Science **280**(5371), 1871–1873 (1998)
26. Poelwijk, F.J., Kiviet, D.J., Weinreich, D.M., Tans, S.J.: Empirical fitness landscapes reveal accessible evolutionary paths. Nature **445**(7126), 383–386 (2007)
27. Ramazzotti, D.: A model of selective advantage for the efficient inference of cancer clonal evolution. arXiv preprint arXiv:1602.07614 (2016)
28. Ramazzotti, D., Caravagna, G., Olde Loohuis, L., Graudenzi, A., Korsunsky, I., Mauri, G., Antoniotti, M., Mishra, B.: CAPRI: efficient inference of cancer progression models from cross-sectional data. Bioinformatics **31**(18), 3016–3026 (2015). DOI 10.1093/bioinformatics/btv296
29. Rhee, S.Y., Gonzales, M.J., Kantor, R., Betts, B.J., Ravela, J., Shafer, R.W.: Human immunodeficiency virus reverse transcriptase and protease sequence database. Nucleic acids research **31**(1), 298–303 (2003)
30. Schwarz, G., et al.: Estimating the dimension of a model. The annals of statistics **6**(2), 461–464 (1978)
31. Seifert, D., Di Giallonardo, F., Metzner, K.J., Günthard, H.F., Beerenwinkel, N.: A framework for inferring fitness landscapes of patient-derived viruses using quasispecies theory. Genetics **199**(1), 191–203 (2015)
32. Simpson, E.H.: The interpretation of interaction in contingency tables. Journal of the Royal Statistical Society. Series B (Methodological) pp. 238–241 (1951)
33. Stone, M.: An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. Journal of the Royal Statistical Society. Series B (Methodological) pp. 44–47 (1977)
34. Suppes, P.: A probabilistic theory of causality. North-Holland Publishing Company Amsterdam (1970)
35. Vandamme, A.M., Van Laethem, K., De Clercq, E.: Managing resistance to anti-hiv drugs. Drugs **57**(3), 337–361 (1999)
36. Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., Kinzler, K.W.: Cancer genome landscapes. science **339**(6127), 1546–1558 (2013)
37. Wang, Y., Waters, J., Leung, M.L., Unruh, A., Roh, W., Shi, X., Chen, K., Scheet, P., Vattathil, S., Liang, H., et al.: Clonal evolution in breast cancer revealed by single nucleus genome sequencing. Nature **512**(7513), 155–160 (2014)
38. Weinreich, D.M., Delaney, N.F., DePristo, M.A., Hartl, D.L.: Darwinian evolution can follow only very few mutational paths to fitter proteins. science **312**(5770), 111–114 (2006)
39. Yule, G.U.: Notes on the theory of association of attributes in statistics. Biometrika **2**(2), 121–134 (1903)