Inferring tree causal models of cancer progression with probability raising

Loes Olde Loohuis*†
Department of Computer Science
City University New York, The Graduate Center
New York, USA.

Giulio Caravagna* Alex Graudenzi*
Daniele Ramazzotti* Giancarlo Mauri
Marco Antoniotti
Dipartimento di Informatica Sistemistica e Comunicazione
Università degli Studi Milano-Bicocca
Milano, Italy.

Bud Mishra
Courant Institute of Mathematical Sciences
New York University
New York, USA.

Abstract

Existing techniques to reconstruct tree models of progression for accumulative processes, such as cancer, seek to estimate causation by combining correlation and a frequentist notion of temporal priority. In this paper, we define a novel theoretical framework called CAPRESE (CAncer PRogression Extraction with Single Edges) to reconstruct such models based on the notion of probabilistic causation defined by Suppes. We consider a general reconstruction setting complicated by the presence of noise in the data due to biological variation, as well as experimental or measurement errors. To improve tolerance to noise we define and use a shrinkage-like estimator. We prove the correctness of our algorithm by showing asymptotic convergence to the correct tree under mild constraints on the level of noise. Moreover, on synthetic data, we show that our approach outperforms the state-of-the-art, that it is efficient even with a relatively small number of samples and that its performance quickly converges to its asymptote as the number of samples increases. For real cancer datasets obtained with different technologies, we highlight biologically significant differences in the progressions inferred with respect to other competing techniques and we also show how to validate conjectured biological relations with progression models.

Introduction

Cancer is a disease of evolution. Its initiation and progression are caused by dynamic somatic alterations to the genome manifested as point mutations, structural alterations, DNA methylation and histone modification changes [1].

These genomic alterations are generated by random processes, and since individual tumor cells compete for space and resources, the fittest variants are naturally selected for. For example, if through some mutations

*Equal contributors.

†L. Olde Loohuis is currently at the Center for Neurobehavioral Genetics, University of California, Los Angeles, USA.

a cell acquires the ability to ignore anti-growth signals from the body, this cell may thrive and divide, and its progeny may eventually dominate some part(s) of the tumor. This *clonal expansion* can be seen as a *discrete state* of the cancer's progression, marked by the acquisition of a set of genetic events. Cancer progression can then be thought of as a sequence of these discrete steps, where the tumor acquires certain distinct properties at each state. Different progression sequences are possible, but some are more common than others, and not every order is viable [2].

In the last two decades, many specific genes and genetic mechanisms that are involved in different types of cancer have been identified (see e.g. [3, 4] for an overview of common cancer genes and [5, 6] for specific genetic analyses of ovarian carcinoma and lung adenocarcinoma, respectively), and *therapies* targeting the activity of these genes are now being developed at a fast pace [2]. However, unfortunately, the *causal and temporal relations* among the genetic events driving cancer progression remain largely elusive.

The main reason for this state of affairs is that information revealed in the data is usually obtained only at one (or a few) points in time, rather than over the course of the disease. Extracting this dynamic information from the available *cross-sectional* data is challenging, and a combination of mathematical, statistical and computational techniques is needed. In recent years, several methods to extract progression models from cross-sectional data have been developed, starting from the seminal work on single-path-models by Fearon and Vogelstein [7]. In particular, different models of oncogenetic trees were developed over the years. At the core of some of these methods, e.g. [8, 9], is the use of *correlation* to identify relations among genetic events. These techniques reconstruct *tree* models of progression as independent acyclic paths with branches and no confluences. Distinct models of oncogenetic trees are instead based on *maximum likelihood estimation*, e.g., [10, 11, 12]. More general *Markov chain* models, e.g., [13], describe more flexible probabilistic networks, despite the computationally expensive parameter estimation. Other recent models are Conjunctive Bayesian Networks, CBNs [14, 15], that extract *directed acyclic graphs*, yet imposing specific constraints on the joint occurrence of events. Finally, in a slightly different context, temporal models were reconstructed from time-course gene expression data [16, 17].

In this paper we present a novel theoretical framework called CAPRESE (CAncer PRogression Extraction with Single Edges) to reconstruct cumulative progressive phenomena, such as cancer progression. We assume the original problem setting of [8], and propose a new a technique to infer *probabilistic progression trees* from cross-sectional data. Unlike maximum likelihood estimation-based techniques, our goal is the extraction of the *minimal* progression model explaining the order in which mutations occur and accumulate. The method is technology agnostic, i.e., it can be applied to dataset derived from all types of (epi-)genetic data such as deep exome sequencing, bisulfite sequencing, SNP arrays, etc., (see Results), and takes as input a set of pre-selected genetic events of which the presence or the absence of each event is recorded for each sample.

CAPRESE is based on two main ingredients: 1) rather than using *correlation* to infer progression structures, we base our technique on a notion of *probabilistic causation*, and 2), to increase robustness against noise, we adopt a *shrinkage-like estimator* to measure causation among any pair of events.

More specifically, with respect to our first ingredient, we adopt the notion of (prima facie) causation proposed by Suppes in [18]. Its basic intuition is simple: event a causes event b if (i) a occurs *before* b and (ii) the occurrence of a *raises the probability* of observing b. This is a very basic notion of probabilistic causation that in itself does not address many of the problems associated with it (such as asymmetry, common causes, and screening off [19]), and includes *spurious* as well as *genuine* causes. However, as it turns out, this basic notion combined with a filter for independent progressions starting from the same root, is an excellent tool to guide progression extraction from cross-sectional data – one that outperforms the commonly used correlation-based methods.

Probabilistic causation was used in biomedical applications before (e.g., to find driver genes from CNV data in [20], and to extract causes from biological time series data in [21]), but, to the best of our knowledge, never to infer *progression models* in the *absence* of direct temporal information.

The extraction problem is complicated by the presence of both false positive and false negative observations (see [22] for a discussion on this issue based on the reconstruction by [8]), such as the one provided by the intrinsic variability of biological processes (e.g., *genetic heterogeneity*) and *experimental errors*. This poses a problem, because while probability raising is a very precise tool, it, by itself, is not robust enough

against noise. Conditional on the amount of noise, we will rely both on probabilistic causation and on a more robust (but less precise) correlation-based metric in an optimal way. To do this we introduce our second ingredient, a *shrinkage-like estimator* to measure causation among any pair of events. The intuition behind this estimator, which is closely related to a shrinkage estimator from [23], is to find the optimal balance between probability raising on the one hand and correlation on the other, depending on the amount of noise.

We prove correctness of our algorithm by showing that with increasing sample sizes, the reconstructed tree asymptotically converges to the correct one (Theorem 3). Under mild constraints on the noise rates, this result holds for the reconstruction problem in the presence of uniform noise as well.

We also study the performance of CAPRESE in more realistic settings with limited sample sizes. Using synthetic data, we show that under these conditions, our algorithm outperforms the state-of-the-art tree reconstruction algorithm of [8] (see Results). In particular, our shrinkage-like estimator provides, on average, an increased robustness to noise which ensures it to outperform oncotrees [8]. Performance is defined in terms of *structural similarity* between the reconstructed tree and the actual tree, rather than on their induced distribution as is done, e.g., in [11]. This metric is especially appropriate for the goal of reconstructing a progression model where data-likelihood fit is secondary to "calling" the possibly minimal set of causal relations.

Also, we show that CAPRESE works well already with a relatively low number of samples and that its performance quickly converges to its asymptote as the number of samples increases. This outcome hints at the applicability of the algorithm with relatively small datasets without compromising its efficiency.

We remark that further analyses on synthetic data suggests that CAPRESE outperforms a well known bayesian probabilistic graphical model as well (i.e., *Conjunctive Bayesian Networks* [14, 15]), which was originally conceived for the reconstruction of more complex topologies, e.g. DAGs, but was proven effective in reconstructing tree topologies as well [24] (see Results).

Finally, we apply our technique to alterations assessed with both Comparative Genomic Hybridization and Next Generation Sequencing techniques (see Results). In the former case, we show that the algorithm of [8] and CAPRESE highlight biologically important differences in ovarian, gastrointestinal and oral cancer, but our inferences are statistically more significant. In the latter, we validate a recently discovered relation among two key genes involved in leukemia.

Methods

Problem setting

The set-up of the reconstruction problem is as follows. Assuming that we have a set G of n mutations (*events*, in probabilistic terminology) and s samples, we represent a cross-sectional dataset as an $s \times n$ binary matrix in which an entry $(k, l) = 1$ if the mutation l was observed in sample k, and 0 otherwise. The problem we solve in this paper is to extract a set of edges E yielding a progression *tree* $\mathcal{T} = (G \cup \{\diamond\}, E, \diamond)$ from this matrix which, we remark, only implicitly provides information of progression timing. The root of $\mathcal{T}$ is modeled using a (special) event $\diamond \notin G$ such that *heterogenous progression paths* or *forests* can be reconstructed. More precisely, we aim at reconstructing a *rooted tree* that satisfies: (i) each node has at most one incoming edge, (ii) the root has no incoming edges (iii) there are no *cycles*.

Each progression tree subsumes a distribution of observing a subset of the mutations in a cancer sample that can be formalized as follows:

Definition 1 (Tree-induced distribution)**.** *Let $\mathcal{T}$ be a tree and $\alpha : E \to [0, 1]$ a labeling function denoting the independent probability of each edge, $\mathcal{T}$ generates a distribution where the probability of observing a sample with the set of alterations $G^* \subseteq G$ is*

$$\mathcal{P}(G^*) = \prod_{e \in E'} \alpha(e) \cdot \prod_{\substack{(u,v) \in E \\ u \in G^*, v \notin G}} \Big[1 - \alpha(u, v)\Big] \tag{1}$$

where all events in G^* are assumed to be reachable from the root $\diamond$, and $E' \subseteq E$ is the set of edges connecting the root to the events in G^*.

We would like to emphasize two properties related to the tree-induced distribution. First, the distribution subsumes that, given any oriented edge $(a \to b)$, an observed sample contains alteration b with probability $\mathcal{P}(a)\mathcal{P}(b)$, that is the probability of observing b after a. For this reason, if a causes b, the probability of observing a will be greater than the probability of observing b accordingly to the temporal priority principle which states that all causes must precede, in time, their effects [25].

Second, the input dataset is a set of samples generated, ideally, from an unknown distribution induced by an unknown tree or forest that we aim at reconstructing. However, in some cases, it could be that no tree exists whose induced distribution generates *exactly* those input data. When this happens, the set of observed samples slightly diverges from any tree-induced distribution. To model these situations a notion of *noise* can be introduced, which depends on the context in which data are gathered. Adding noise to the model complicates the reconstruction problem (see Results).

The *oncotree* approach. In [8] Desper *et al.* developed a method to extract progression trees, named *"oncotrees"*, from static CNV data. In [22] Szabo *et al.* extended the setting of Desper's reconstruction problem to account for both *false positives* and *negatives* in the input data. In these oncotrees, nodes represent CNV events and edges correspond to possible progressions from one event to the next.

The reconstruction problem is exactly as described above, and each tree is rooted in the special event $\diamond$. The choice of which edge to include in a tree is based on the estimator

$$w_{a\to b} = \log\left[\frac{\mathcal{P}(a)}{\mathcal{P}(a)+\mathcal{P}(b)} \cdot \frac{\mathcal{P}(a,b)}{\mathcal{P}(a)\mathcal{P}(b)}\right], \tag{2}$$

which assigns to each edge $a \to b$ a weight accounting for both the relative and joint frequencies of the events – thus measuring *correlation*. The estimator is evaluated after including $\diamond$ to each sample of the dataset. In this definition the rightmost term is the (symmetric) *likelihood ratio* for a and b occurring together, while the leftmost is the asymmetric *temporal priority* measured by rate of occurrence. This implicit form of timing assumes that, if a occurs *more often* than b, then it likely occurs *earlier*, thus satisfying

$$\frac{\mathcal{P}(a)}{\mathcal{P}(a)+\mathcal{P}(b)} > \frac{\mathcal{P}(b)}{\mathcal{P}(a)+\mathcal{P}(b)}\,.$$

An oncotree is the rooted tree whose total weight (i.e., sum of all the weights of the edges) is maximized, and can be reconstructed in $O(|G|^2)$ steps using Edmond's algorithm [26]. By construction, the resulting graph is a proper tree rooted in $\diamond$: each event occurs only once, *confluences* are absent, i.e., any event is caused by at most one other event. This method has been used to derive progressions for various cancer datasets e.g., [27, 28, 29]), and even though several methods that extend this framework exists (e.g. [9, 11, 15]), to the best of our knowledge, it is currently the only method that aims to solve exactly the same problem as the one investigated in this paper and thus provide a benchmark to compare against.

A probabilistic approach to causation

We briefly review the approach to probabilistic causation, on which our method is based. For an extensive discussion on this topic we refer to [19].

In his seminal work [18], Suppes proposed the following notion.

Definition 2 (Probabilistic causation, [18])**.** *For any two events c and e, occurring respectively at times t_c and t_e, under the mild assumptions that $0 < \mathcal{P}(c), \mathcal{P}(e) < 1$, the event c is a* prima facie cause *of the event e if it occurs* before *the effect and the cause* raises the probability *of the effect, i.e.,*

$$t_c < t_e \quad \textit{and} \quad \mathcal{P}(e \mid c) > \mathcal{P}(e \mid \overline{c})\,. \tag{3}$$

As discussed in [19] the above conditions are not, in general, sufficient to claim that event c is a cause of event e. In fact a prima facie cause is either *genuine* or *spurious*. In the latter case, the fact that the conditions hold in the observations is due either to coincidence or to the presence of a certain third *confounding factor*, related both to c and to e [18]. Genuine causes, instead, satisfy Definition 2 and are not screened-off by any confounding factor. However, they need not be direct causes. See Figure 1.

Note that we consider cross-sectional data where no information about t_c and t_e is available, so in our reconstruction setting we are restricted to consider solely the *probability raising* (PR) property, i.e., $\mathcal{P}(e \mid c) > \mathcal{P}(e \mid \overline{c})$, which makes it harder to discriminate among genuine and spurious causes. Now we review some of its properties.

Proposition 1 (Dependency)**.** *Whenever the* PR *holds between two events a and b, then the events are* statistically dependent *in a positive sense, i.e.,*

$$\mathcal{P}(b \mid a) > \mathcal{P}(b \mid \overline{a}) \iff \mathcal{P}(a, b) > \mathcal{P}(a)\mathcal{P}(b)\,. \tag{4}$$

This and the next proposition are well-known facts of the PR; their derivation as well as the proofs of all the results we present is in the Supplementary Materials. Notice that the opposite implication holds as well: when the events a and b are still dependent but in a negative sense, i.e., $\mathcal{P}(a, b) < \mathcal{P}(a)\mathcal{P}(b)$, the PR does not hold, i.e., $\mathcal{P}(b \mid a) < \mathcal{P}(b \mid \overline{a})$.

We would like to use the asymmetry of the PR to determine whether a pair of events a and b satisfy a causation relation so to place a before b in the progression tree but, unfortunately, the PR satisfies the following property.

Proposition 2 (Mutual PR)**.** $\mathcal{P}(b \mid a) > \mathcal{P}(b \mid \overline{a}) \iff \mathcal{P}(a \mid b) > \mathcal{P}(a \mid \overline{b})$.

That is, if a raises the probability of observing b, then b raises the probability of observing a too.

Nevertheless, in order to determine causes and effects among the genetic events, we can use our *degree of confidence* in our estimate of probability raising to decide the direction of the causation relationship between pairs of events. In other words, if a raises the probability of b *more* than the other way around, then a is a more likely cause of b than b of a. Notice that this is sound as long as each event has *at most* one cause; otherwise, *frequent late events* with more than one cause, which are rather common in biological progressive phenomena, should be treated differently. As mentioned, the PR is not symmetric, and the *direction* of probability raising depends on the relative frequencies of the events. We make this asymmetry precise in the following proposition.

Proposition 3 (Probability raising and temporal priority)**.** *For any two events a and b such that the probability raising $\mathcal{P}(a \mid b) > \mathcal{P}(a \mid \overline{b})$ holds, we have*

$$\mathcal{P}(a) > \mathcal{P}(b) \iff \frac{\mathcal{P}(b \mid a)}{\mathcal{P}(b \mid \overline{a})} > \frac{\mathcal{P}(a \mid b)}{\mathcal{P}(a \mid \overline{b})}\,. \tag{5}$$

That is, given that the PR holds between two events, a raises the probability of b *more* than b raises the probability of a, if and only if a is observed more frequently than b. Notice that we use the ratio to assess the PR inequality. The proof of this proposition is technical and can be found in the Supplementary Materials. From this result it follows that if we measure the timing of an event by the rate of its occurrence (that is, $\mathcal{P}(a) > \mathcal{P}(b)$ implies that a happens before b), this notion of PR subsumes the same notion of temporal priority induced by a tree. We also remark that this is also the temporal priority made explicit in the coefficients of Desper's method. Given these results, we define the following notion of causation.

Definition 3. *We state that a is a* prima facie cause *of b if a is a probability raiser of b, and it occurs more frequently: $\mathcal{P}(b \mid a) > \mathcal{P}(b \mid \overline{a})$ and $\mathcal{P}(a) > \mathcal{P}(b)$.*

We term *prima facie topology* a directed acyclic graph (over some events) where each edge represents a prima facie cause. When at most a single incoming edge is assigned to each event (i.e., an event has at most a *unique cause*, in the real world), we term this structure *single-cause prima facie topology*. Intuitively,

this last class of topologies correspond to the trees or, more generally forests when they have disconnected components, that we aim at reconstructing.

Before moving on to introducing our algorithm let us discuss our definition of *causation*, its role in the definition of the reconstruction problem and some of its limitations. As already mentioned, it may be that for some prima facie cause c of an event e, there is a third event a prior to both, such that a causes c and ultimately c causes e. Alternatively, a may cause both c and e independently, and the causation relationship observed from c to e is merely *spurious*. In the context of the tree-reconstruction problem, namely when it is assumed that each event has at most a unique cause, the aim is to filter out the spurious edges from a general prima facie topology, so to extract a single-cause prima facie structure (see Figure 1).

Definition 3 summarizes Suppes basic notion of prima facie cause, while it is ignoring deeper discussions of causation that aim at distinguishing between actual genuine and spurious causes, e.g. screening-off, background context, d-separation [30, 31, 19]. For our purposes however, the above definition is sufficient when (i) all the significant events are considered, i.e., all the genuine causes are observed as in a closed-world assumption, and (ii) we aim at extracting the *order* of progression among them (or determine that there is no apparent relation), rather than extracting causalities *per se*. Note that these assumptions are strong and might be weakened in the future (see Discussions), but are shared by us and [8].

Finally, we recall a few algebraic requirements necessary for our framework to be well-defined. First of all, the PR must be computable: every mutation a should be observed with probability strictly $0 < \mathcal{P}(a) < 1$. Moreover, we need each pair of mutations (a, b) to be *distinguishable* in terms of PR, that is, for each pair of mutations a and b, $\mathcal{P}(a \mid b) < 1$ or $\mathcal{P}(b \mid a) < 1$ similarly to the above condition. Any non-distinguishable pair of events can be merged as a single composite event. From now on, we will assume these conditions to be verified.

Performance measure and synthetic datasets

We made use of *synthetic data* to evaluate the performance of CAPRESE as a function of dataset size and the false positive and negative rates. Many distinct synthetic datasets were created for this purpose, as explained below. The algorithm's performance was measured in terms of *Tree Edit Distance* (TED, [32]), i.e., the minimum-cost sequence of node edit operations (relabeling, deletion and insertion) that transforms the reconstructed trees into the ones generating the data. The choice of this measure of evaluation is motivated by the fact that we are interested in the *structure* behind the progressive phenomenon of cancer evolution and, in particular, we are interested in a measure of the genuine causes that we miss and of the spurious causes that we fail to recognize (and eliminate). Also, since topologies with similar distributions can be structurally different we choose to measure performance using structural distance rather than a distance in terms of distributions. Within the realm of 'structural metrics' however, we have also evaluated the performance with the *Hamming Distance* [33], another commonly used structural metric, and we obtained analogous results (not shown here).

Synthetic data generation and experimental setting. Synthetic datasets were generated by sampling from various random trees constrained to have depth $\log(|G|)$, since wide branches are harder to reconstruct than straight paths, and by sampling event probabilities in $[0.05, 0.95]$ (see Supplementary Materials).

Unless explicitly specified, in all the experiments we used 100 distinct random trees (or forests, accordingly to the test to perform) of 20 events each. This seems a fairly reasonable number of events and is in line with the usual size of reconstructed trees, e.g. [34, 35, 36, 37]. The *scalability* of the techniques was tested against the number of samples by ranging $|G|$ from 50 to 250, with a step of 50, and by replicating 10 independent datasets for each parameters setting (see the caption of the figures for details).

We included a form of *noise* in generating the datasets, in order to account for (i) the realistic presence of *biological noise* (such as the one provided by bystander mutations, genetic heterogeneity, etc.) and (ii) *experimental errors*. A noise parameter $0 \leq \nu < 1$ denotes the probability that any event assumes a random value (with uniform probability), after sampling from the tree-induced distribution. Algorithmically this process generates, on average, $|G|\nu/2$ random entries in each sample (e.g. with $\nu = 0.1$ we have, on average,

one error per sample). We wish to assess whether these noisy samples can mislead the reconstruction process, even for low values of ν. Notice that assuming a uniformly distributed noise may appear simplistic since some events may be more robust, or easy to measure, than others. However, introducing in the data both *false positives* (at rate $\epsilon_+ = \nu/2$) and *negatives* (at rate $\epsilon_- = \nu/2$) makes the inference problem substantially harder, and was first investigated in [22].

In the Results section, we refer to datasets generated with rate $\nu > 0$ as noisy synthetic dataset. In the numerical experiments, ν is usually discretized by 0.025, (i.e., 2.5% noise).

Results

Extracting progression trees with probability raising and a shrinkage-like estimator

The CAPRESE reconstruction method is described in Algorithm 1. The algorithm is similar to Desper's and Szabo's algorithm, the main difference being an alternative weight function based on a shrinkage-like estimator.

Definition 4 (Shrinkage-like estimator)**.** *We define the* shrinkage-like estimator $m_{a\to b}$ *of the confidence in the causation relationship from a to b as*

$$m_{a\to b} = (1-\lambda)\alpha_{a\to b} + \lambda\beta_{a\to b}\,, \tag{6}$$

where $0 \le \lambda \le 1$ *and*

$$\alpha_{a\to b} = \frac{\mathcal{P}(b \mid a) - \mathcal{P}(b \mid \overline{a})}{\mathcal{P}(b \mid a) + \mathcal{P}(b \mid \overline{a})} \qquad \beta_{a\to b} = \frac{\mathcal{P}(a,b) - \mathcal{P}(a)\mathcal{P}(b)}{\mathcal{P}(a,b) + \mathcal{P}(a)\mathcal{P}(b)}\,. \tag{7}$$

This estimator is similar in spirit to a shrinkage estimator (see [23]) and combines a normalized version of PR, the *raw estimate* α, with a *correction factor* β (in our case a correlation-based measure of temporal distance among events), to define a proper order in the confidence of each causation relationship. Our λ is the analogous of the *shrinkage coefficient* and can have a Bayesian interpretation based on the strength of our belief that a and b are causally relevant to one another and the evidence that a raises the probability of b. In the absence of a closed form solution for the optimal value of λ, one may rely on cross-validation of simulated data. The power of shrinkage (and our shrinkage-like estimator) lies in the possibility of determining an optimal value for λ to balance the effect of the correction factor on the raw model estimate to ensure optimal performances on ill posed instances of the inference problem. A crucial difference, however, between our estimator and classical shrinkage, is that our estimator aims at improving the performance of the *overall* reconstruction process, not limited to the performance of the estimator itself as is the case in shrinkage. That is, the metric m induces an ordering to the events reflecting our confidence for their causation. Furthermore, since we make no assumption about the underlying distribution, we learn it empirically by cross-validation. In the next sections we show that the shrinkage-like estimator is an effective way to get such an ordering especially when data are noisy. In CAPRESE we use a pairwise matrix version of the estimator.

The raw estimator and the correction factor. By considering only the raw estimator α, we would include an edge $(a \to b)$ in the tree consistently in terms of (i) Definition 3 (Methods) and (ii) if a is the best probability raiser for b. When $\mathcal{P}(a) = \mathcal{P}(b)$ the events a and b are indistinguishable in terms of temporal priority, thus α is not sufficient to decide their causal relation, if any. This intrinsic ambiguity is unlikely in practice even if, in principle, it is possible. Notice that this formulation of α is a monotonic normalized version of the PR ratio.

Proposition 4 (Monotonic normalization)**.** *For any two events a and b we have*

$$\mathcal{P}(a) > \mathcal{P}(b) \iff \frac{\mathcal{P}(b \mid a)}{\mathcal{P}(b \mid \overline{a})} > \frac{\mathcal{P}(a \mid b)}{\mathcal{P}(a \mid \overline{b})} \iff \alpha_{a\to b} > \alpha_{b\to a}\,. \tag{8}$$

This raw model estimator satisfies $-1 \leq \alpha_{a\to b}, \alpha_{b\to a} \leq 1$: when it tends to -1 the pair of events appear disjointly (i.e., they show an anti-causation pattern), when it tends to 0 no causation or anti-causation can be inferred and the two events are statistically independent and, when it tends to 1, the causation relationship between the two events is genuine. Therefore, α provides a quantification of the degree of confidence for a PR causation relationship. In fact, for any given possible causation edge (a, b), the term $\mathcal{P}(b \mid \overline{a})$ gives an estimate of the *error rate* of b, therefore the numerator of the raw model α provides an estimate of how often b is actually caused by a. The α estimator is then normalized to range between -1 and $+1$.

However, α does not provide a general criterion to disambiguate among genuine causes of a given event. We show a specific case in which α is not a sufficient estimator. Let us consider, for instance, a causal linear path: $a \to b \to c$. In this case, when evaluating the candidate parents a and b for c we have: $\alpha_{a\to c} = \alpha_{b\to c} = 1$, so a and b are genuine causes of c, though we would like to select b, instead of a. Accordingly, we can only infer that $t_a < t_c$ and $t_b < t_c$, i.e., a partial ordering, which does not help to disentangle the relation among a and b with respect to c.

In this case, the β coefficient can be used to determine which of the two genuine causes occurs closer, in time, to c (b, in the example above). In general, such a correction factor provides information on the *temporal distance* between events, in terms of statistical dependency. In other words, the higher the β coefficient, the closer two events are. Therefore, when dealing with noisy data and limited sample sizes, there are situations where, by using the α estimator alone, we could infer a wrong transitive edge to be the most likely cause even in the presence of the real cause. For this reason, we reduce the α estimator to the correction factor β, which, for each given edge (a, b), is normalized within -1 and $(1 - \max[\mathcal{P}(a), \mathcal{P}(b)])/(1 + \max[\mathcal{P}(a), \mathcal{P}(b)]) < +1$.

The shrinkage-like estimator m then results in the combination of the raw PR estimator α and of the β correction factor, which respects the temporal priority induced by α.

Proposition 5 (Coherence in dependency and temporal priority)**.** *The β correction factor is* symmetrical *and subsumes the same notion of dependency of the raw estimator α, that is*

$$\mathcal{P}(a, b) > \mathcal{P}(a)\mathcal{P}(b) \;\Leftrightarrow \alpha_{a\to b} > 0 \Leftrightarrow \beta_{a\to b} > 0 \quad \textit{and} \quad \beta_{a\to b} = \beta_{b\to a}\,. \tag{9}$$

The independent progressions filter. As in Desper's approach, we also add a *root* $\diamond$ with $\mathcal{P}(\diamond) = 1$ in order to separate different progression paths, i.e., the different sub-trees rooted in $\diamond$. CAPRESE initially builds a unique tree by using the estimator; typically, the most likely event will be at the top of the progression even if there may be rare cases where more than one event has no valid parent, in these cases we would already be reconstructing a forest. In the reconstructed tree, all the edges represent the most confident prima facie cause, although some of those could still be spurious causes. Then the correlation-like weight between any node j and $\diamond$ is computed as

$$\frac{\mathcal{P}(\diamond)}{\mathcal{P}(\diamond) + \mathcal{P}(j)} \frac{\mathcal{P}(\diamond, j)}{\mathcal{P}(\diamond)\mathcal{P}(j)} = \frac{1}{1 + \mathcal{P}(j)}\,.$$

If this quantity is greater than the weight of j with each upstream connected element i, we consider the best prima facie cause of j to be a spurious cause and we substitute the edge $(i \to j)$ with the edge $(\diamond \to j)$. Note that in this work we are ignoring deeper discussions of probabilistic causation that aim at distinguishing between actual genuine causes and spurious causes. Instead, we remove spurious causes by using a filter based on correlation because the probability raising of the omnipresent event $\diamond$ is not well defined (see Methods). In addition, we remark that the evaluation for an edge to be a genuine or a spurious cause takes into account all the given events. Because of this, if events are added or removed from the dataset, the same edge can be defined to be genuine or spurious as the set of events included in the model is varied arbitrarily. However, since we do not consider the problem of selecting the set of progression events, we assume that all and only the relevant events for the problem at hand are alreadyknown a priori and included in the model.

Finally, note that this filter is indeed implying a non-negative threshold for the shrinkage-like estimator, when an edge is valid.

Theorem 1 (Independent progressions)**.** *Let $G^* = \{a_1, \ldots, a_k\} \subset G$ a set of k* prima facie causes *for some $b \notin G^*$, and let $a^* = \max_{a_i \in G^*}\{m_{a_i \to b}\}$. The reconstructed tree by CAPRESE contains edge $\diamond \to b$ instead of $a^* \to b$ if, for all $a_i \in G^*$*

$$\mathcal{P}(a_i, b) < \mathcal{P}(a_i)\mathcal{P}(b)\frac{1}{1+\mathcal{P}(b)} + \frac{\mathcal{P}(b)^2}{1+\mathcal{P}(b)}\,. \tag{10}$$

The proof of this theorem can be found in the Supplementary Materials. What this theorem suggests is that, in principle, by examining the level of statistical dependency of each pair of events, it would be possible to determine how many trees compose the reconstructed forest. Furthermore, it suggests that CAPRESE could be defined by first processing the independent progressions filter, and then using m to build the independent progression trees in the forest.

To conclude, the algorithm reconstructs a well defined tree (or, more in general, forest).

Theorem 2 (Algorithm correctness)**.** *CAPRESE reconstructs a well defined tree $\mathcal{T}$ without disconnected components, transitive connections and cycles.*

Additionally, asymptotically with the number of samples, the reconstructed tree is the correct one.

Theorem 3 (Asymptotic convergence)**.** *Let $T = (G \cup \{\diamond\}, E, \diamond)$ be the forest to reconstruct from a set of s input samples, given as the input matrix D. If D is strictly sampled from the distribution induced by T and infinite samples are available, i.e., $s \to \infty$, CAPRESE with $\lambda \to 0$ correctly reconstructs T.*

The proof of these Theorems are also in the Supplementary Materials. These theorems considered datasets where the observed and theoretical probabilities match, because of $s \to \infty$. However, data often contains false positives and negatives (i.e., data are noisy) and resistance to their effects is desirable in any inferential technique. With this in mind, we prove a corollary of the theorem analoguos to a result appearing in [22].

Corollary 1 (Uniform noise)**.** *Let the input matrix D be strictly sampled from the distribution induced by T with sample size $s \to \infty$, but let it be corrupted by noise levels of false positives ϵ_+ and false negatives ϵ_-. Let $p_{\min} = \min_{x \in G}\{\mathcal{P}(x)\}$, CAPRESE correctly reconstructs T for $\lambda \to 0$ whenever*

$$\epsilon_+ < \sqrt{p_{\min}}(1 - \epsilon_+ - \epsilon_-)$$

and $\epsilon_+ + \epsilon_- < 1$.

Essentially, this corollary states that CAPRESE (and so the estimator m) is robust against a noise affecting all samples equally. Also, the fact that it holds for $\lambda \to 0$ is sound with the theory of shrinkage estimators for which, asymptotically, the corrector factor is not needed to regularize the ill posed problem.

Optimal shrinkage-like coefficient

Theorem 3 and Corollary 1 state that with infinite samples and mild constraints on the false positive/negative rates we get optimal results with $\lambda \to 0$. Precisely, for the uniform noise model that we applied to synthetic data (see Methods) we have $\epsilon_+ = \epsilon_- = \nu/2$, thus the hypothesis required by Corollary 1 is

$$\nu < \frac{\sqrt{p_{\min}}}{1/2 + \sqrt{p_{\min}}}\,.$$

For $p_{\min} = 0.05$, which we set in data generation (see Supplementary Materials), this inequality implies correct reconstruction for $\nu < 0.3$ (a 15% error rate), with infinite samples. However, we are interested in performance and the optimal value of λ in situations in which we have finite sample sizes as well. Here, we empirically estimate the optimal λ value, both in the case of trees and forests, as a function of noise and sample size. In the next section, we assess performance of our algorithm empirically.

In Figure 2, we show the variation of the performance of CAPRESE as a function of λ, for datasets with 150 samples generated from tree topologies. The optimal value, i.e., lowest Tree Edit Distance (TED, see

Methods), for noise-free datasets (i.e., $\nu = 0$) is obtained for $\lambda \to 0$, whereas for the noisy datasets a series of U-shaped curves suggests a unique optimum value for $\lambda \to 1/2$, immediately observable for $\nu < 0.15$. Identical results are obtained when dealing with forests (not shown here). In addition, further experiments with n varying around the typical sample size ($n = 150$) show that the optimal λ is largely insensitive to the dataset size (see Figure 3). Thus we have limited our analysis to datasets with the typical sample size that is characteristic of data currently available.

Summarizing, Figures 2 and 3 suggest that for sample size below 250 without false positives and negatives the PR raw estimate α suffices to drive reconstruction to good results (TED is 0 with 250 samples), i.e.,

$$m_{a\to b} \overset{\lambda\to 0}{\approx} \alpha_{a\to b} \tag{11}$$

which is obtained by setting λ to a very small value, e.g. 10^{-2}, in order to consider at least a small contribution of the correction factor too. Conversely, when $\nu > 0$, the best performance is obtained by averaging the shrinkage-like effect, i.e.,

$$m_{a\to b} \overset{\lambda=1/2}{=} \frac{\alpha_{a\to b}}{2} + \frac{\beta_{a\to b}}{2}\,. \tag{12}$$

These results suggest that, in general, a unique optimal value for the shrinkage-like coefficient can be determined, even in situations not captured by Theorem 3 and Corollary 1.

Performance of CAPRESE compared to *oncotrees*

An analogue of Theorem 3 holds for Despers's oncotrees (Theorem 3.3, [8]), and an analogue of Corollary 1 holds for Szabo's extension with uniform noise (Reconstruction Theorem 1, [22]). Thus, with infinite samples both approaches reconstruct the correct trees/forests. With finite samples and noise, however, their performance may show different patterns, as speed of convergence may vary. We investigate this issue in the current section.

In Figure 4 we compare the performance of CAPRESE with oncotrees, for the case of noise-free synthetic data with the optimal shrinkage-like coefficient: $\lambda \to 0$, equation (11). Since Szabo's algorithm is equivalent to Desper's without false negatives and positives, we rely solely on Szabo's implementation [22]. In Figure 5 we show an example of reconstructed tree where CAPRESE infers the correct tree while oncotrees mislead a causation relation.

In general, one can observe that the TED of CAPRESE is, on average, always bounded above by the TED of oncotrees, both in the case of trees and forests. For trees, with 50 samples the average TED of CAPRESE is around 6, whereas for Desper's technique it is around 13. The performance of both algorithms improves as long as the number of samples is increased: CAPRESE has the best performance (i.e., TED ≈ 0) with 250 samples, while oncotrees have TED around 6. When forests are considered, the difference between the performance of the algorithms reduces slightly, but also in this case CAPRESE clearly outperforms oncotrees.

Notice that the improvement due to the increase in the sample size seems to reach a *plateau*, and the initial TED for our estimator seems rather close to the plateau value. This empirical analysis suggests that CAPRESE has already good performances with few samples, a favorable adjoint to Theorem 3. This result has some important practical implications, particularly considering the scarcity of available biological data.

In Figure 6 we extend the comparison to *noisy* datasets. In this case, we used the optimal shrinkage-like coefficient: $\lambda \to 1/2$, equation (12). The results confirm what observed without false positives and negatives, as CAPRESE outperforms oncotrees up to $\nu = 0.15$, for all the sizes of the sample sets. In the Supplementary Materials we show similar plots for the noise-free case.

We can thus draw the conclusion that our algorithm performs better with finite samples and noise, since less samples are required to get good performances and a higher resistance to false positives and negatives is shown.

Performance of CAPRESE compared to *Conjunctive Bayesian Networks*

Inspired by Desper's seminal work, Beerenwinkel and others developed methods to estimate the constraints on the order in which mutations accumulate during cancer progression, using a probabilistic graphical model called *Conjuntive Bayesian Networks* (CBN)[14, 15]. While the goal of this research was to reconstruct *direct acyclic graphs* and not trees per se, evidence presented in [24] suggests that, in the absence of noise, these models perform better than oncotrees even at reconstructing *trees*. For this reason, we performed experiments similar to the ones suggested above, comparing CAPRESE to the extension of CBN called *hidden-CBN* (h-CBN) that accounts for noisy genotype observations [15]. This method combines CBNs with a simulated annealing algorithm for structure search and a denoising of the genotypes via the maximum a posteriori estimates to compute the most likely progression. One aspect that complicates a comparison between CAPRESE and (h-)CBN is that the methods assume different models. For example, at the heart of CBN is a monotonicity assumption (i.e., an event can only occur if all its predecessors have occurred) not assumed by CAPRESE. Despite the differences between the model assumptions, we present a preliminary comparison between the methods in Appendix B.3, indicating that we not only outperform oncotrees, but h-CBNs as well. In particular, this suggests that CAPRESE converges much faster than h-CBNs with respect to the sample size, also in the presence of noise.

We also analyze the rate of *false positives/negatives* reconstructed by CAPRESE when (synthetic) data are sampled from DAGs (Appendix B.3). The rate of *false positives* goes to 0 as the sample size increases, implying that CAPRESE is capable of reconstructing a tree subsumed by the underlying causal DAG topology. In addition, the number of *false negatives* approaches a value proportional to the connectivity of the model from which the data was generated. This is expected, since CAPRESE will assign at most one cause to each considered event. However, it should be noted that further experiments with samples selected from a wider array of topologies should be performed to confirm these results and compare both methods in full. While not within the scope of the current paper, these issues will be addressed in future work.

Case studies

In the next subsections we apply CAPRESE to real cancer data obtained by *Comparative Genomic Hybridization* (CGH) and *Next Generation Sequencing* (NGS). This shows the potential application of reconstruction techniques to various types of mutational profiles and various cancers.

Performance on cancer CGH datasets

Encouraged by the results in previous sections, we test our reconstruction approach on a real *ovarian cancer* dataset made available within the oncotree package [8]. The data was collected through the public platform SKY/M-FISH [38], used to allow investigators to share molecular cytogenetic data. The data was obtained by using the CGH technique on samples from *papillary serous cystadenocarcinoma* of the ovary. This technique uses fluorescent staining to detect CNV data at the resolution of chromosome arms. While the recent emergence of NGS approaches make the dataset itself rather outdated, the underlying principles remain the same and the dataset provides a valid test-case for our approach. The seven most commonly occurring events are selected from the 87 samples, and the set of events are the following gains and losses on chromosomes arms $G = \{8q+, 3q+, 1q+, 5q-, 4q-, 8p-, Xp-\}$ (e.g., $4q-$ denotes a deletion of the q arm of the 4th chromosome).

In Figure 7 we compare the trees reconstructed by the two approaches. Our technique differs from Desper's by predicting the causal sequence of alterations

$$8q+ \;\rightarrow\; 8p- \;\rightarrow\; Xp-\,,$$

when used either $\lambda \rightarrow 0$ or $\lambda = 1/2$. Notice that among the samples in the dataset some are not generated by the distribution induced by the recovered tree, thus comparing the reconstruction for both λs is necessary.

At this point, we do not have a biological interpretation for this result. However, we do know that common cancer genes reside in these regions, e.g. the tumor suppressor gene PDGFR on $5q$ and the oncogene MYC on $8q$), and loss of heterozygosity on the short arm of chromosome 8 is quite common (see, e.g.,

http://www.genome.jp/kegg/). Recently, evidence has been reported that $8p$ contains many cooperating cancer genes [39].

In order to assign a confidence level to these inferences we applied both parametric and non-parametric *bootstrapping methods* to our results. Essentially, these tests consist of using the reconstructed trees (in the parametric case), or the probability observed in the dataset (in the non-parametric case) to generate new synthetic datasets, and then reconstructs again the progressions (see, e.g., [40] for an overview of these methods and [41] for the use of bootstrap for evalutating the confidence of oncogenetic trees.). The confidence is given by the number of times the trees in Figure 7 are reconstructed from the generated data. A similar approach can be used to estimate the confidence of every edge separately. For oncotrees the *exact tree* is obtained 83 times out of 1000 non-parametric resamples, so its estimated confidence is 8.3%. For CAPRESE the confidence is 8.6%. In the parametric case with false positive and false negative error rates of 0.21 and 0.027, following [22], the confidence of oncotrees is 17% while the confidence of our method is much higher: 32%. When error rates are forced to 0 the confidence of oncotrees raises to 86.6% and 90.9% respectively.

For the non-parametric case, edges confidence is shown in Table 8. Most notably, our algorithm reconstructs the inference $8q+ \to 8p-$ with high confidence (confidence 62%, and 26% for $5q- \to 8p-$), while the confidence of the edge $8q+ \to 5q-$ is only 39%, almost the same as $8p- \to 8q+$ (confidence 40%). The confidences are similar with either $\lambda \to 0$ or $\lambda = 1/2$.

Analysis of other CGH datasets. We report the differences between the reconstructed trees also based on datasets of gastrointestinal and oral cancer ([35, 37] respectively). In the case of gastrointestinal stromal cancer, among the 13 CGH events considered in [35] (gains on $5p$, $5q$ and $8q$, losses on $14q$, $1p$, $15q$, $13q$, $21q$, $22q$, $9p$, $9q$, $10q$ and $6q$), oncotrees identify the path progression

$$1p- \to 15q- \to 13q- \to 21q-$$

while CAPRESE reconstructs the branch

$$1p- \to 15q- \qquad\qquad 1p- \to 13q- \to 21q- \ .$$

In the case of oral cancer, among the 12 CGH events considered in [37] (gains on $8q$, $9q$, $11q$, $20q$, $17p$, $7p$, $5p$, $20p$ and $18p$, losses on $3p$, $8p$ and $18q$), the reconstructed trees differ since oncotrees identifies the path

$$8q+ \to 20q+ \to 20p+$$

while our algorithm reconstructs the path

$$3p- \to 7p+ \to 20q+ \to 20p+ \ .$$

These examples show that CAPRESE provides important differences in the reconstruction compared to oncotrees.

Performance on cancer NGS datasets

In this section we show the application of reconstruction techniques to the validation of a specific relation among recurrent mutations involved in *atypical Chronic Myeloid Leukemia* (aCML).

In [42] Piazza *et al.* used high-throughput *exome sequencing technology* to identity somatically acquired mutations in 64 aCML patients, and found a previously unidentified recurring *missense point mutation* hitting Setbp1. By re-sequencing Setbp1 in samples with aCML and other common human cancers, they found that around 25% of the aCML patients tested positive for Setbp1, while most of the other types of tumors were negative. Assessing the possible relationship between Setbp1 variants and the mutations in many driver aCML oncogenes such as (e.g., Asxl1, Tet2, Kras, etc.) no significant association or mutual exclusion with Setbp1 was found but for Asxl1, which was frequently mutated together with Setbp1, hinting at a potential relation among the events. In particular, Asxl1 was presenting either a *non-sense point* or a *indel* type of somatic mutation.

Hence, we reconstructed aCML progression models from the datasets provided in [42], with the goal of assessing a *potential causal dependency* between mutated Setbp1 and Asxl1. A more extensive analysis is postponed, as we only seek to clearly illustrate the functionalities of the algorithm here.

As a first case (Figure 9, left), we treated the Asxl1 missense point and indel mutations as indistinguishable, and we merged the two events in the dataset. Afterwards, we separated the two types of mutations for Asxl1 (Figure 9, right).

In particular, it is interesting to notice that, when the Asxl1 mutations are considered equivalent, the inference suggests that the mutations belong to two independent progression paths (i.e., the independent progression filters "breaks" every potential causal relation among the events). Conversely, when the mutations are kept separate, the progression model suggests that: (*i*) the missense point mutation hitting Setbp1 can cause a non-sense point mutation in Asxl1 and (*ii*) the observed Asxl1 mutations seems to be independent. Concerning edges confidence, as before assessed via non-parametric bootstrap, it is worth noting that the confidence in the indel Asxl1 mutation being an early event raises consistently in the latter case.

All in all, it seems that a progression model allows to test the significance of the association firstly observed in [42] and also refines the knowledge by suggesting a specific causal and temporal relations among events. With this this in mind, ad-hoc sequencing experiments might be set up to assess these predictions, eventually providing a strong evidence that could be used to, e.g., synthesize a progression-specific aCML-effective drug.

Discussion

In this work we have introduced a novel framework for the reconstruction of the causal topologies underlying cumulative progressive phenomena, based on the *probability raising* notion of causation. Besides such a probabilistic notion, we also introduced the use of a *shrinkage-like estimator* to efficiently unravel ambiguous causal relations, often present when data are noisy. As a first step towards the definition of our new framework, we have here presented an effective novel technique called CAPRESE (CAncer PRogression Extraction with Single Edges) for the reconstruction of tree or, more generally, forest models of progression which combines probabilistic causation and a shrinkage-like estimation.

We prove correctness of CAPRESE by showing asymptotic convergence to the correct tree. Under mild constraints on the noise rates, this result holds for the reconstruction problem in the presence of uniform noise as well. Moreover, we also compare our technique to the standard tree reconstruction algorithm based on correlation (i.e., Oncontrees) and to a more general bayesian probabilistic graphical model (i.e., Conjunctive Bayesian Networks), and show that CAPRESE outperforms the state-of-the-art on synthetic data, also exhibiting a noteworthy efficiency with relatively small datasets. Furthermore, we tested our technique on ovarian, gastrointestinal and oral cancer CGH data and NGS leukemia data. The CGH analysis suggested that our approach can infer, with high confidence, novel causal relationships which would remain unpredictable in a correlation-based attack. The NGS analysis allowed validating a causal and temporal relation among key mutations in atypical chronic myeloid leukemia.

One of the strong points of CAPRESE is that it can be applied to genomic data of any kind, even heterogeneous, and at any resolution, as shown. In fact, it simply requires a set of samples in which the presence or the absence of some alterations supposed to be involved in a causal cumulative process have been assessed. Notice also that the results of our technique can be used not only to describe the *progression* of the process, but also to *classify* different progression types. In the case of cancer, for instance, this genome-level classifier could be used to group patients according to the position of the detected individual genomic alterations in the progression model (e.g., at a specific point of the tree) and, consequently, to set up a genome-specific *therapy design* aimed, for instance, at blocking or slowing certain progression paths instead of others, as was studied in [43].

Several future research directions are possible. First, more complex models of progression, e.g. directed acyclic graphs, could be inferred with probability raising and compared to the standard approaches of [14, 15, 44], as we explained in the introduction. These models, rather than trees, could explain the common phenomenon of *preferential progression paths* in the target process via, e.g., *confluence* among events. In

the case of cancer, for instance, these models would be certainly more suitable than trees to describe the accumulation of mutations.

Second, the shrinkage-like estimator itself could be improved by introducing, for instance, different correction factors. In addition, an analytical formulation of the optimal shrinkage-like coefficient could be investigated by starting from the hypotheses which apply to our problem setting, along the lines of [45].

Third, advanced statistical techniques such as *bootstrapping* [40] could be used to account for more sophisticated models of noise within data, so as to decipher complex causal dependencies.

Finally, a further development of the framework could involve the introduction of *timed data*, in order to extend our techniques to settings where a temporal information on the samples is available.

Software availability

The implementation of CAPRESE is part of the TRanslational ONCOlogy (TRONCO) R package and is available for download at standard R repositories.

Acknowledgements

We are grateful for the many excellent comments we received from anonymous reviewers.

References

[1] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: The next generation," *Cell*, vol. 144, pp. 646–674, 2011.

[2] J. Luo, N. L. Solimini, and S. J. Elledge, "Principles of cancer therapy: Oncogene and non-oncogene addiction," *Cell*, vol. 136, pp. 823–837, Mar. 2009.

[3] B. Vogelstein and K. Kinzler, "Cancer genes and the pathways they control," *Nature medicine*, vol. 10, no. 8, pp. 789–799, 2004.

[4] S. A. Frank, *Dynamics of Cancer*. Princeton University Press, 2007.

[5] D. Bell, A. Berchuck, M. Birrer, J. Chien, D. Cramer, F. Dao, R. Dhir, P. DiSaia, H. Gabra, and P. Glenn, "Integrated genomic analyses of ovarian carcinoma," 2011.

[6] M. Imielinski *et al.*, "Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing," *Cell*, vol. 150, no. 6, pp. 1107–1120, 2012.

[7] B. Vogelstein, E. R. Fearon, S. R. Hamilton, S. E. Kern, A. C. Preisinger, M. Leppert, A. M. Smits, and J. L. Bos, "Genetic alterations during colorectal-tumor development," *New England Journal of Medicine*, vol. 319, no. 9, pp. 525–532, 1988.

[8] R. Desper, F. Jiang, O. Kallioniemi, H. Moch, C. Papadimitriou, and A. Schäffer, "Inferring tree models for oncogenesis from comparative genome hybridization data," *Journal of Computational Biology*, vol. 6, no. 1, pp. 37–51, 1999.

[9] R. Desper, F. Jiang, O. Kallioniemi, H. Moch, C. Papadimitriou, and A. Schäffer, "Distance-based reconstruction of tree models for oncogenesis," *Journal of Computational Biology*, vol. 7, no. 6, pp. 789–803, 2000.

[10] A. von Heydebreck, B. Gunawan, and L. Füzesi, "Maximum likelihood estimation of oncogenetic tree models," *Biostatistics*, vol. 5, no. 4, pp. 545–56, 2004.

[11] N. Beerenwinkel, J. Rahnenführer, M. Däumer, D. Hoffmann, R. Kaiser, J. Selbig, and T. Lengauer, "Learning multiple evolutionary pathways from cross-sectional data," *Journal of Computational Biology*, vol. 12, no. 6, pp. 584–598, 2005.

[12] N. Beerenwinkel, J. Rahnenführer, R. Kaiser, D. Hoffmann, J. Selbig, and T. Lengauer, "Mtreemix: a software package for learning and using mixture models of mutagenetic trees," *Bioinformatics*, vol. 21, no. 9, pp. 2106–2107, 2005.

[13] M. Hjelm, M. Höglund, and J. Lagergren, "New probabilistic network models and algorithms for oncogenesis," *Journal of Computational Biology*, vol. 13, pp. 853–865, 2006.

[14] N. Beerenwinkel, N. Eriksson, and B. Sturmfels, "Conjunctive bayesian networks," *Bernoulli*, pp. 893–909, 2007.

[15] M. Gerstung, M. Baudis, H. Moch, and N. Beerenwinkel, "Quantifying cancer progression with conjunctive bayesian networks," *Bioinformatics*, vol. 25, no. 21, pp. 2809–2815, 2009.

[16] A. Gupta and Z. Bar-Joseph, "Extracting dynamics from static cancer expression data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 5, pp. 172–182, June 2008.

[17] N. Ramakrishnan, S. Tadepalli, L. T. Watson, R. F. Helm, M. Antoniotti, and B. Mishra, "Reverse engineering dynamic temporal models of biological processes and their relationships.," *PNAS*, vol. 107, pp. 12511–12516, July 2010.

[18] P. Suppes, *A probabilistic theory of causality.* North Holland Publishing Company, 1970.

[19] C. Hitchcock, "Probabilistic causation," in *The Stanford Encyclopedia of Philosophy* (E. Zalta, ed.), winter 2012 ed., 2012.

[20] I. Ionita, R. Daruwala, and B. Mishra, "Mapping Tumor-Suppressor genes with multipoint statistics from Copy-Number–Variation data," *American Journal of Human Genetics*, vol. 79, pp. 13–22, July 2006. PMID: 16773561 PMCID: 1474131.

[21] S. Kleinberg, *Causality, Probability, and Time.* Cambridge University Press, 2012.

[22] A. Szabo and K. Boucher, "Estimating an oncogenetic tree when false negatives and positives are present," *Mathematical biosciences*, vol. 176, no. 2, pp. 219–236, 2002.

[23] B. Efron, *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction.* Cambridge University Press, 2013.

[24] K. Hainke, J. Rahnenführer, and R. Fried, "Cumulative disease progression models for cross-sectional data: A review and comparison," *Biometrical Journal*, vol. 54, pp. 617–640, Aug. 2012.

[25] H. Reichenbach, *The Direction of Time.* University of California Press, 1956.

[26] J. Edmonds, "Optimum branchings," *Journal of Research of the National Bureau of Standards B*, vol. 71, pp. 233–240, 1967.

[27] T. Kainu *et al.*, "Somatic deletions in hereditary breast cancers implicate 13q21 as a putative novel breast cancer susceptibility locus," *Proceedings of the National Academy of Sciences*, vol. 97, no. 17, pp. 9603–9608, 2000.

[28] Q. Huang, G. Yu, S. McCormick, J. Mo, B. Datta, M. Mahimkar, P. Lazarus, A. A. Schäffer, R. Desper, and S. Schantz, "Genetic differences detected by comparative genomic hybridization in head and neck squamous cell carcinomas from different tumor sites: construction of oncogenetic trees for tumor progression," *Genes, Chromosomes and Cancer*, vol. 34, no. 2, pp. 224–233, 2002.

[29] M. Radmacher, R. Simon, R. Desper, R. Taetle, A. Schäffer, and M. Nelson, "Graph models of oncogenesis with an application to melanoma," *Journal of theoretical biology*, vol. 212, no. 4, pp. 535–548, 2001.

[30] N. Cartwright, "Causal laws and effective strategies," *Noûs*, vol. 13, pp. 419–437, 1979.

[31] J. Pearl, *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[32] K. Zhang and D. Shasha, "Simple fast algorithms for the editing distance between trees and related problems," *SIAM journal on computing*, vol. 18, no. 6, pp. 1245–1262, 1989.

[33] R. W. Hamming, "Error-detecting and error-correcting codes," *Bell System Technical Journal*, vol. 29, no. 2, pp. 147–160, 1950.

[34] E. Samuelson, S. Karlsson, K. Partheen, S. Nilsson, C. Szpirer, and A. Behboudi, "Bac cgh-array identified specific small-scale genomic imbalances in diploid dmba-induced rat mammary tumors," *BMC cancer*, vol. 12, no. 1, p. 352, 2012.

[35] B. Gunawan *et al.*, "An oncogenetic tree model in gastrointestinal stromal tumours (gists) identifies different pathways of cytogenetic evolution with prognostic implications," *The Journal of pathology*, vol. 211, no. 4, pp. 463–470, 2007.

[36] T. Longerich, M. Mueller, K. Breuhahn, P. Schirmacher, A. Benner, and C. Heiss, "Oncogenetic tree modeling of human hepatocarcinogenesis," *International Journal of Cancer*, vol. 130, no. 3, pp. 575–583, 2012.

[37] S. Pathare, A. Schäffer, N. Beerenwinkel, and M. Mahimkar, "Construction of oncogenetic tree models reveals multiple pathways of oral cancer progression," *International journal of cancer*, vol. 124, no. 12, pp. 2864–2871, 2009.

[38] T. Knutsen, V. Gobu, R. Knaus, H. Padilla-Nash, M. Augustus, R. Strausberg, I. Kirsch, K. Sirotkin, and T. Ried, "The interactive online sky/m-fish & cgh database and the entrez cancer chromosomes search database: Linkage of chromosomal aberrations with the genome sequence," *Genes, Chromosomes and Cancer*, vol. 44, no. 1, pp. 52–64, 2005.

[39] W. Xue *et al.*, "A cluster of cooperating tumor-suppressor gene candidates in chromosomal deletions," *Proceedings of the National Academy of Sciences*, vol. 109, no. 21, pp. 8212–8217, 2012.

[40] B. Efron, *The jackknife, the bootstrap and other resampling plans*, vol. 38. SIAM, 1982.

[41] A. Szabo and K. M. Boucher, "Oncogenetic trees," *Handbook of Cancer Models with Applications to Cancer Screening, Cancer Treatment and Risk Assessment. World Scientific*, 2008.

[42] R. Piazza, S. Valletta, N. Winkelmann, S. Redaelli, R. Spinelli, A. Pirola, L. Antolini, L. Mologni, C. Donadoni, E. Papaemmanuil, S. Schnittger, D.-W. Kim, J. Boultwood, F. Rossi, G. Gaipa, G. P. D. Martini, P. F. di Celle, H. G. Jang, V. Fantin, G. R. Bignell, V. Magistroni, T. Haferlach, E. M. Pogliani, P. J. Campbell, A. J. Chase, W. J. Tapper, N. C. P. Cross, and C. Gambacorti-Passerini, "Recurrent setbp1 mutations in atypical chronic myeloid leukemia," *Nature Genetics*, vol. 45, pp. 18–24, 2013.

[43] L. Olde Loohuis, A. Witzel, and B. Mishra, "Cancer hybrid automata: Model, beliefs & therapy," *Information and Computation*, 2014.

[44] M. Gerstung, N. Eriksson, J. Lin, B. Vogelstein, and N. Beerenwinkel, "The temporal order of genetic and pathway alterations in tumorigenesis," *PloS one*, vol. 6, no. 11, p. e27136, 2011.

[45] B. Efron and C. Morris, "Stein's estimation rule and its competitors–an empirical bayes approach," *Journal of the American Statistical Association*, vol. 68, no. 341, pp. 117–130, 1973.

A Figures and tables

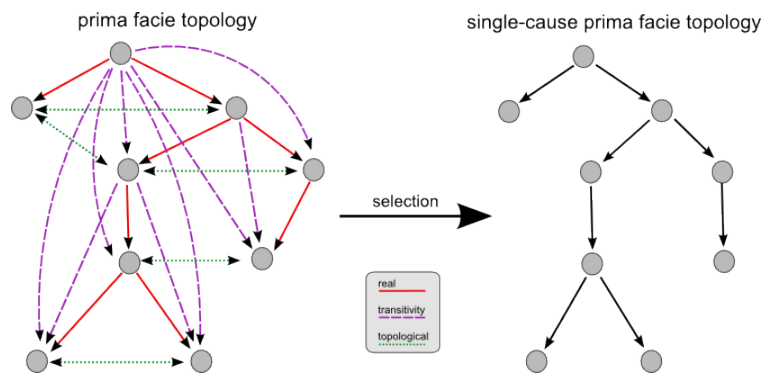


Figure 1: **Prima facie topology.** Example prima facie topology where all edges (a, b) represent prima facie causes, according to Definition 3: a is a probability raiser of b and it occurs more frequently. In left, we filter out spurious causes and select only the real ones among the genuine, yielding a single-cause prima facie topology.

Algorithm 1 CAPRESE: a tree-like reconstruction with a shrinkage-like estimator

1: consider a set of n genetic events G plus a special event $\diamond$, added to each sample of the dataset;

2: define a $m \times n$ matrix M where each entry contains the shrinkage-like estimator

$$m_{i \to j} = (1 - \lambda) \cdot \frac{\mathcal{P}(j \mid i) - \mathcal{P}(j \mid \bar{i})}{\mathcal{P}(j \mid i) + \mathcal{P}(j \mid \bar{i})} + \lambda \cdot \frac{\mathcal{P}(i, j) - \mathcal{P}(i)\mathcal{P}(j)}{\mathcal{P}(i, j) + \mathcal{P}(i)\mathcal{P}(j)}$$

according to the observed probability of the events i and j;

3: [PR causation] define a tree $\mathcal{T} = (G \cup \{\diamond\}, E, \diamond)$ where $(i \to j) \in E$ for $i, j \in G$ if and only if:

$$m_{i \to j} > 0 \quad \text{and} \quad m_{i \to j} > m_{j \to i} \quad \text{and} \quad \forall i' \in G,\ m_{i,j} > m_{i',j}\,.$$

4: [Independent progressions filter] define $G_j = \{x \in G \mid \mathcal{P}(x) > \mathcal{P}(j)\}$, replace edge $(i \to j) \in E$ with edge $(\diamond \to j)$ if, for all $x \in G_j$, it holds

$$\frac{1}{1 + \mathcal{P}(j)} > \frac{\mathcal{P}(x)}{\mathcal{P}(x) + \mathcal{P}(j)} \frac{\mathcal{P}(x, j)}{\mathcal{P}(x)\mathcal{P}(j)}\,.$$

B Supplementary Materials

B.1 Proofs

Here the proofs of all the propositions and theorems follow.

Proof of Proposition 1 (Dependency).

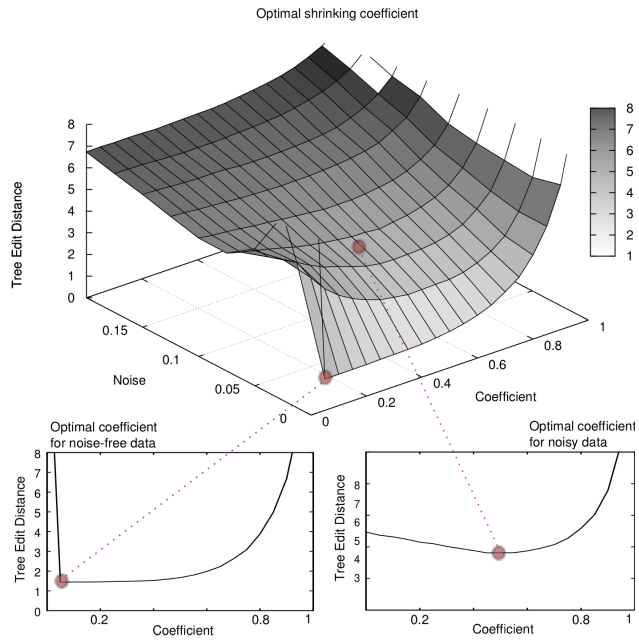


Figure 2: **Optimal shrinkage-like coefficient for reconstruction performance.** We show here the performance in the reconstruction of trees (TED surface) with $n = 150$ samples as a function of the shrinkage-like coefficient λ. Notice the global optimal performance for $\lambda \to 0$ when $\nu \to 0$ and for $\lambda \approx 1/2$ when $\nu > 0$.

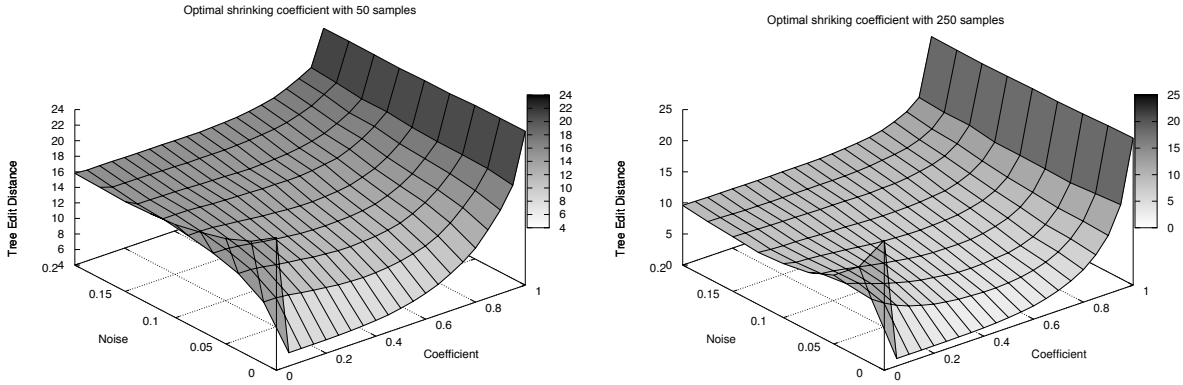


Figure 3: **Optimal λ with datasets of different size.** We show the analogous of Figure 2 with 50 and 250 samples. The estimation of the optimal shrinkage-like coefficient λ appears to be irrespective of the sample size.

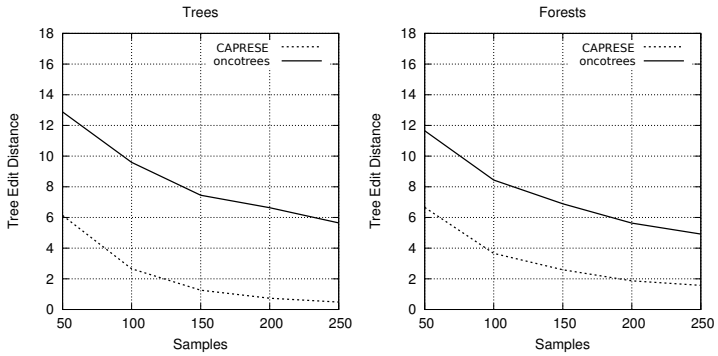


Figure 4: **Comparison on noise-free synthetic data.** Performance of CAPRESE (dashed line) and oncotrees (full line) in average TED when data are generated by random trees (left) and forests (right). In this case $\nu = 0$ (no false positives/negatives) and $\lambda \to 0$ in the estimator m.

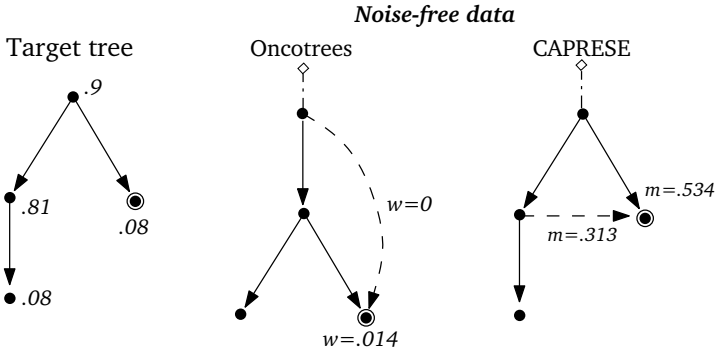


Figure 5: **Example of reconstructed trees.** Example of reconstruction from a dataset with 100 samples generated by the left tree (the theoretical probabilities are shown, i.e., the doubly-circled event appears in a sample with probability .08), with $\nu = 0$. In the sampled dataset oncotrees mislead the cause of the doubly-circled mutation ($w = 0$ for the true edge and $w = 0.014$ for the wrong one). CAPRESE infers the correct cause (the values of the estimator m with $\lambda = 1/2$ are shown, similar results are obtained for $\lambda \to 1$).

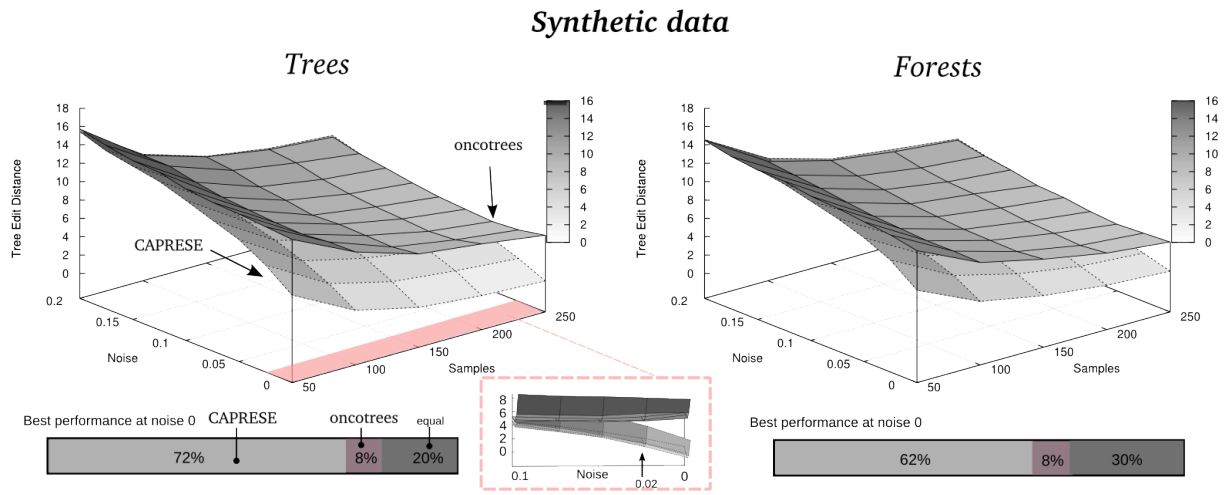


Figure 6: **Reconstruction with noisy synthetic data and $\lambda = 1/2$.** Performance of CAPRESE and oncotrees as a function of the number of samples and noise ν. According to Figure 2 the shrinkage-like coefficient is set to $\lambda = 1/2$. The magnified image shows the convergence to Desper's performance for $\nu \approx 0.1$. The barplot represents the percentage of times the best performance is achieved at $\nu = 0$.

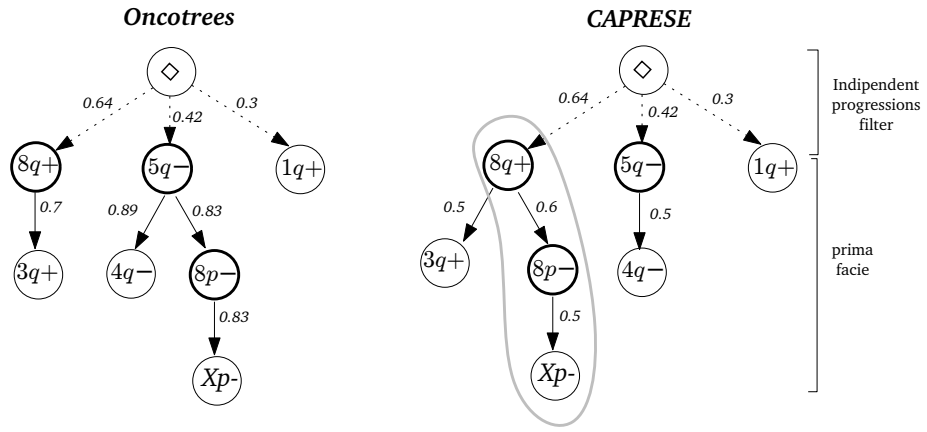


Figure 7: **Reconstruction of ovarian cancer progression.** Trees reconstructed by oncotrees and CAPRESE (with $\lambda \to 0$, with $\lambda = 1/2$ the same tree is reconstructed). The set of CGH events considered are gains on $8q$, $3q$ and $1q$ and losses on $5q$, $4q$, $8p$ and Xp. Events on chromosomes arms containing the key genes for ovarian cancer are in bolded circles. In the left tree all edge weights are the observed probabilities of events. In the right the full edges are the causation inferred with the PR and the weights represent the scores used by CAPRESE. Weights on dashed lines are as in the left tree.

Oncotrees (overall confidence 8.3%)

$\rightarrow$	$8q+$	$3q+$	$5q-$	$4q-$	$8p-$	$1q+$	$Xp-$
$\diamond$	**.99**	.06	**.51**	.22	.004	**.8**	.06
$8q+$	0	**.92**	.08	0.16	0.4	.02	.007
$3q+$	.002	0	.04	0	0	.09	.04
$5q-$	.001	.002	0	**.52**	**.39**	.009	.16
$4q-$	0	0	.27	0	.14	.05	.11
$8p-$	0	0	.07	.08	0	.004	**.59**
$1q+$	0	0	0	.004	0	0	0
$Xp-$	0	0	.003	.003	.04	.01	0

CAPRESE (overall confidence 8.6%)

$\rightarrow$	$8q+$	$3q+$	$5q-$	$4q-$	$8p-$	$1q+$	$Xp-$
$\diamond$	**.99**	.06	**.51**	.22	.004	**.8**	.06
$8q+$	0	**.92**	.06	.16	**.62**	.01	.008
$3q+$	.002	0	.03	.002	0	.09	.04
$5q-$	.001	.002	0	**.5**	.26	.009	.17
$4q-$	0	0	.29	0	.09	.05	.12
$8p-$	0	0	.07	.08	0	.004	**.59**
$1q+$	0	0	0	.004	0	0	0
$Xp-$	0	.001	.003	.004	.01	.01	0

Figure 8: **Estimated confidence for ovarian progression.** Frequency of edge occurrences in the non-parametric bootstrap test, for the trees shown in Figure 7. Colors represent confidence: light gray is < 0.4, mid gray is in the range $[0.4, 0.8]$ and dark gray is > 0.8. Bold entries are the edges recovered by the algorithms.

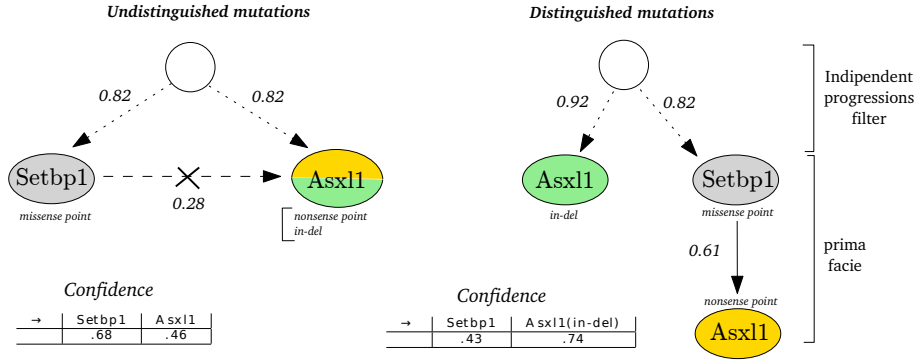


Figure 9: **Validating the Setbp1-Asxl1 relation in atypical Chronic Myeloid Leukemia.** Progression models where Asxl1 indel and non-sense point are merged (left) and separate (right) suggest that a missense point mutation hitting Setbp1 can cause a non-sense point mutation in Asxl1, that the observed Asxl1 mutations might be independent and that indel Asxl1 is an early event with high confidence.

Proof. For $\Rightarrow$ write $\mathcal{P}(\overline{a}, b) = \mathcal{P}(b) - \mathcal{P}(a, b)$, then write the PR as

$$\frac{\mathcal{P}(a,b)}{\mathcal{P}(a)} > \frac{\mathcal{P}(b) - \mathcal{P}(a,b)}{1 - \mathcal{P}(a)}$$

and, since $0 < \mathcal{P}(a) < 1$, the proposition follows by simple algebraic arrangements of $\mathcal{P}(a,b) \cdot [1 - \mathcal{P}(a)] > \mathcal{P}(a)\mathcal{P}(b) - \mathcal{P}(a,b) \cdot \mathcal{P}(a)$. The derivations are analogous but in reverse order for the implication $\Leftarrow$. □

Proof of Proposition 2 (Mutual probability raising).

Proof. The proof follows by Property 1 and the subsequent implication:

$$\mathcal{P}(b \mid a) > \mathcal{P}(b \mid \overline{a}) \Leftrightarrow \mathcal{P}(a,b) > \mathcal{P}(a)\mathcal{P}(b) \Leftrightarrow \mathcal{P}(a \mid b) > \mathcal{P}(a \mid \overline{b})\,.$$

□

Proof of Proposition 3 (Probability raising and temporal priority).

Proof. We first prove the forward direction $\Rightarrow$. Let $x = \mathcal{P}(\overline{a}, b)$, $y = \mathcal{P}(a, b)$ and $z = \mathcal{P}(a, \overline{b})$. We have two assumptions we will use later on:

1. $\mathcal{P}(a) > \mathcal{P}(b)$ which implies $\mathcal{P}(\overline{a}, b) < \mathcal{P}(a, \overline{b})$, i.e., $x < z$.
2. $\mathcal{P}(a \mid b) > \mathcal{P}(a \mid \overline{b})$ which, when $0 < x + y < 1$, implies by simple algebraic rearrangements the inequality
$$y[1 - x - y - z] > xz\,. \tag{13}$$

We proceed by rewriting $\mathcal{P}(b \mid a)/\mathcal{P}(b \mid \overline{a}) > \mathcal{P}(a \mid b)/\mathcal{P}(a \mid \overline{b})$ as

$$\frac{\mathcal{P}(a,b)\mathcal{P}(\overline{a})}{\mathcal{P}(\overline{a},b)\mathcal{P}(a)} > \frac{\mathcal{P}(a,b)\mathcal{P}(\overline{b})}{\mathcal{P}(a,\overline{b})\mathcal{P}(b)}$$

which means that

$$\frac{\mathcal{P}(b \mid a)}{\mathcal{P}(b \mid \overline{a})} > \frac{\mathcal{P}(a \mid b)}{\mathcal{P}(a \mid \overline{b})} \iff \frac{\mathcal{P}(\overline{a})}{\mathcal{P}(\overline{a},b)\mathcal{P}(a)} > \frac{\mathcal{P}(\overline{b})}{\mathcal{P}(a,\overline{b})\mathcal{P}(b)} \tag{14}$$

We can rewrite the right-hand side of (14) by using x, y, z where $\mathcal{P}(a) = \mathcal{P}(a,b) + \mathcal{P}(a,\overline{b}) = y + z$ and $\mathcal{P}(b) = \mathcal{P}(a,b) + \mathcal{P}(\overline{a},b) = x + y$, and then do suitable algebraic manipulations. We have

$$\frac{1 - y - z}{x(y+z)} > \frac{1 - x - y}{z(x+y)} \iff yz - y^2z - xz^2 - yz^2 > xy - x^2y - x^2z - xy^2 \tag{15}$$

when $x(y+z) \neq 0$ and $z(x+y) \neq 0$. To check that the right side of (15) holds we show that

$$(xy - x^2y - x^2z - xy^2) - (yz - y^2z - xz^2 - yz^2) < 0\,.$$

First, we rearrange it as $(x - z)[y - y^2 - xz - y(x+z)] < 0$ to show that

$$(x - z)[y(1 - y - x - z) - zx] < 0 \tag{16}$$

is always negative. By observing that, by assumption 1 we have $z > x$ and thus $(x - z) < 0$, and, by equation (13) we have $y(1 - y - x - z) - zx > 0$, we derive

$$\frac{\mathcal{P}(b \mid a)}{\mathcal{P}(b \mid \overline{a})} > \frac{\mathcal{P}(a \mid b)}{\mathcal{P}(a \mid \overline{b})}$$

which concludes the $\Rightarrow$ direction.

The other direction $\Leftarrow$ follows immediately by contraposition: assume that $\mathcal{P}(a \mid b) > \mathcal{P}(a \mid \overline{b})$, $\mathcal{P}(b \mid a)/\mathcal{P}(b \mid \overline{a}) > \mathcal{P}(a \mid b)/\mathcal{P}(a \mid \overline{b})$ and $\mathcal{P}(b) \leq \mathcal{P}(a)$. We distinguish two cases:

1. $\mathcal{P}(b) = \mathcal{P}(a)$, then $\mathcal{P}(b \mid a)/\mathcal{P}(b \mid \overline{a}) = \mathcal{P}(a \mid b)/\mathcal{P}(a \mid \overline{b})$.
2. $\mathcal{P}(b) < \mathcal{P}(a)$, then by symmetry $\mathcal{P}(b \mid a) > \mathcal{P}(b \mid \overline{a})$, and by the $\Rightarrow$ direction of the proposition it follows that $\mathcal{P}(b \mid a)/\mathcal{P}(b \mid \overline{a}) < \mathcal{P}(a \mid b)/\mathcal{P}(a \mid \overline{b})$.

In both cases we have a contradiction. This completes the proof. □

Proof of Proposition 4 (Monotonic normalization).

Proof. We prove the forward direction $\Rightarrow$, the converse follows by a similar argument. Let us assume

$$\frac{\mathcal{P}(b \mid a)}{\mathcal{P}(b \mid \overline{a})} > \frac{\mathcal{P}(a \mid b)}{\mathcal{P}(a \mid \overline{b})} \tag{17}$$

then $\mathcal{P}(b \mid a)\mathcal{P}(a \mid \overline{b}) > \mathcal{P}(a \mid b)\mathcal{P}(b \mid \overline{a})$. Now, to show the righthand side of the implication, we will show that

$$\Big[\mathcal{P}(b \mid a) - \mathcal{P}(b \mid \overline{a})\Big]\Big[\mathcal{P}(a \mid b) + \mathcal{P}(a \mid \overline{b})\Big] > \Big[\mathcal{P}(b \mid a) + \mathcal{P}(b \mid \overline{a})\Big]\Big[\mathcal{P}(a \mid b) - \mathcal{P}(a \mid \overline{b})\Big]$$

which reduces to show

$$\mathcal{P}(b \mid a)\mathcal{P}(a \mid \overline{b}) - \mathcal{P}(b \mid \overline{a})\mathcal{P}(a \mid b) > \mathcal{P}(b \mid \overline{a})\mathcal{P}(a \mid b) - \mathcal{P}(b \mid a)\mathcal{P}(a \mid \overline{b})\,.$$

By (17), two equivalent inequalities hold

$$\mathcal{P}(b \mid a)\mathcal{P}(a \mid \overline{b}) - \mathcal{P}(b \mid \overline{a})\mathcal{P}(a \mid b) > 0$$
$$\mathcal{P}(b \mid \overline{a})\mathcal{P}(a \mid b) - \mathcal{P}(b \mid a)\mathcal{P}(a \mid \overline{b}) < 0$$

and hence the implication holds. □

Proof of Proposition 5 (Coherence in dependency and temporal priority).

Proof. We make two assumptions:

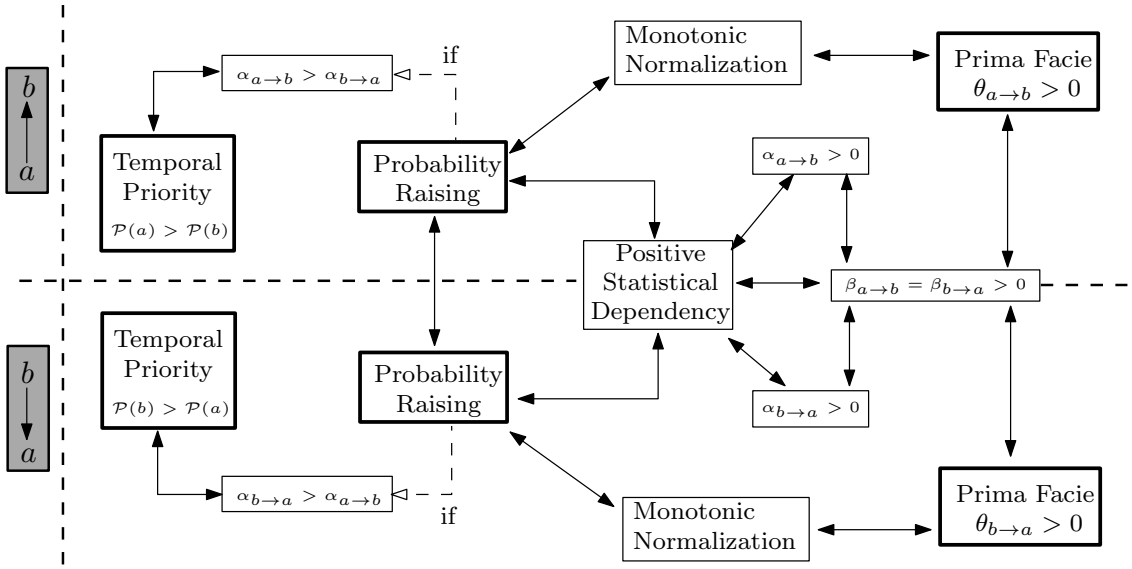
1. $\mathcal{P}(b \mid a) > \mathcal{P}(b \mid \overline{a})$ which implies $\alpha_{a\to b} > 0$.
2. $\mathcal{P}(a, b) > \mathcal{P}(a)\mathcal{P}(b)$ which implies $\beta_{a\to b} > 0$.

The proof for dependency follows by Proposition 1 and its implication:

$$\mathcal{P}(b \mid a) > \mathcal{P}(b \mid \overline{a}) \Leftrightarrow \mathcal{P}(a, b) > \mathcal{P}(a)\mathcal{P}(b) \Leftrightarrow \alpha_{a\to b} > 0 \Leftrightarrow \beta_{a\to b} > 0.$$

Moreover, being β symmetric by definition, the proof for temporal priority follows directly by Proposition 2 □

The properties outlined so far are sketched informally in the following diagram.



Proof of Theorem 1 (Independent progressions).

Proof. For any $a_i \in G^*$ it holds that

$$\mathcal{P}(a_i) > \mathcal{P}(b) \qquad m_{a_i \to b} > 0$$

being prima facie, also $a^* \to b$ is the edge selected by CAPRESE being the $\max\{\cdot\}$ over G^*. Thus, CAPRESE selects $\diamond \to b$ instead of $a^* \to b$ if, for any a_i, it holds

$$\frac{1}{1+\mathcal{P}(b)} > \frac{\mathcal{P}(a_i)}{\mathcal{P}(a_i)+\mathcal{P}(b)} \frac{\mathcal{P}(a_i, b)}{\mathcal{P}(a_i)\mathcal{P}(b)} \,.$$

With some algebraic manipulations we rewrite this as

$$\mathcal{P}(a_i)\mathcal{P}(b) + \mathcal{P}(b)^2 > \mathcal{P}(a_i, b)(1+\mathcal{P}(b)) \,,$$

which gives the inequality in the theorem statement. □

Proof of Theorem 2 (Algorithm correctness).

Proof. It is clear that CAPRESE does not create disconnected components since, to each node in G, a unique parent is attached (either from G or $\diamond$). For the same reason, no transitive connections can appear.

The absence of cycles results from Properties 3, 4 and 5. Indeed, suppose for contradiction that there is a cycle $(a_1, a_2), (a_2, a_3), \ldots, (a_n, a_1)$ in E, then by the three propositions we have

$$\mathcal{P}(a_1) > \mathcal{P}(a_2) > \ldots > \mathcal{P}(a_n) > \mathcal{P}(a_1)$$

which is a contradiction. □

Proof of Theorem 3 (Asymptotic convergence).

Proof. For any $u \in G$, when $s \to \infty$ the *observed probability* $\mathcal{P}(u)$ (evaluated from D) is equivalent to the product of the probabilities (in T) obtained by traversing the forest from the root $\diamond$ to u (Definition 1). Thus, $\mathcal{P}(u) \in (0, 1)$ since the traversal probabilities are in $(0, 1)$ too, hence all events are distinguishable and Algorithm 1 reconstructs a tree with the same events set G of T.

We now observe that the distribution induced by T (Definition 1) respects a *single-cause prima facie topology* where to each event is assigned *at most* a single cause. In other words, Definition 2 holds for any edge $(u, v) \in E$:

- by the event-persistence property usually assumed in cancer (fixating mutations are present in the progeny of a clone) the occurring times satisfy $t_u < t_v$ which, in a frequentist sense, implies $\mathcal{P}(u) > \mathcal{P}(v)$;
- it holds by construction (Definition 1) that $\mathcal{P}(v, u) = \mathcal{P}(v)$, thus $\mathcal{P}(v \mid u) = \mathcal{P}(v)/\mathcal{P}(u)$ which is strictly positive since $\mathcal{P}(v)$ and $\mathcal{P}(u)$ are, and that $\mathcal{P}(v, \overline{u}) = 0$, thus $\mathcal{P}(v \mid \overline{u}) = 0$.

To correctly reconstruct T we rely on the fact that our score $m_{u \to v}$ is consistent with the prima facie probabilistic causation because of:

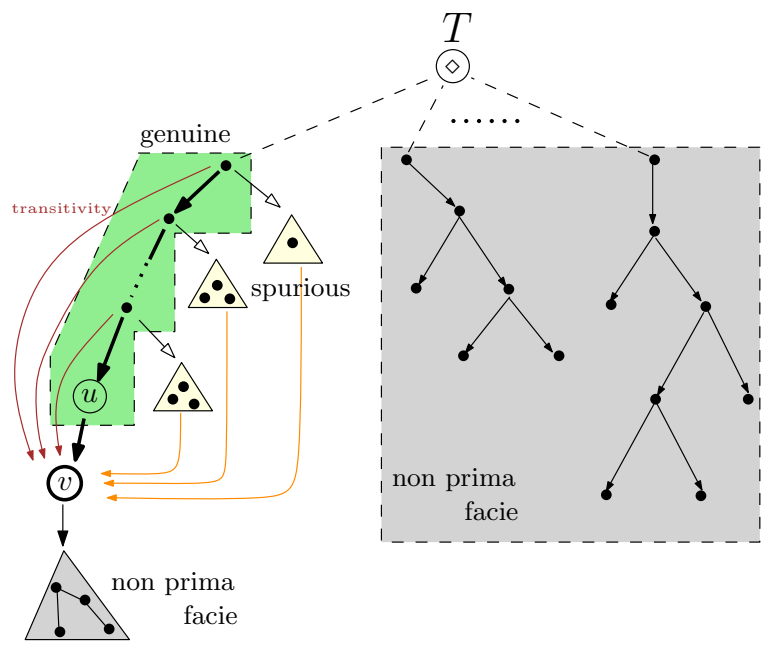
- Proposition 3, which states that PR (embedded as $\alpha_{u \to v}$ in m) subsumes a good temporal priority model of occurring times, as stated above;
- Proposition 4 and 5 which ensure the monotonicity and sign coherency among $\alpha_{u \to v}$ and $\beta_{u \to v}$ in m.

Thus, m is consistent with a single-cause prima facie topology. We now show that Algorithm 1 reconstructs correctly a generic edge in E, and hence also T.

Consider an event $v \in G$ and edge $(u, v) \in E$. The set of its "candidate" parent events is $G \setminus \{v\}$, we partition it in three disjoint sets $\mathcal{G}$, $\mathcal{S}$ and $\mathcal{N}$:

- $\mathcal{G}$*, genuine:* all the backward-reachable events, in $G \setminus \{v\}$, from v;
- $\mathcal{S}$*, spurious (or ambiguous):* all the events (but v) in the sub-forest which includes the path from $\diamond$ to v, which are not in $\mathcal{G}$;
- $\mathcal{N}$*, non prima facie:* all other events, i.e., $G \setminus (\{v, \} \cup \mathcal{S} \cup \mathcal{G})$;

Notice that $G = \{v\} \cup \mathcal{G} \cup \mathcal{S} \cup \mathcal{N}$, that $u \in \mathcal{G}$ and that all the effects of v are non prima facie to v because of the temporal priority, as shown below.



This way of partitioning events according to the structure of T subsumes a equivalent partitioning based on the score $\alpha \in [0, 1]$, which we use to prove correctness of our algorithm: for any $x \in G$ it holds that $\alpha_{x \to v} = 1$ if $x \in \mathcal{G}$, $0 < \alpha_{x \to v} < 1$ if $x \in \mathcal{S}$ and $\alpha_{x \to v} < 0$ if $x \in \mathcal{N}$.

We now show that CAPRESE correctly selects $u \in \mathcal{G}$:

- a non prima facie event $x \in \mathcal{N}$ either satisfies (Proposition 1)

  $$\mathcal{P}(x, v) \leq \mathcal{P}(x)\mathcal{P}(v)$$

  which means that $\alpha_{x \to v} < 0$, $\beta_{x \to v} < 0$ (Proposition 5) and thus $m_{x \to v} < 0$, or it is a descendant of v, which means that $\mathcal{P}(x) < \mathcal{P}(v)$. By construction, CAPRESE considers as candidate parents of v only not descendant events with positive score (see step 3);

- a spurious event $x \in \mathcal{S}$ is prima facie to v but $\alpha_{x \to v} < 1$ since:

  - $\mathcal{P}(x)\mathcal{P}(v) < \mathcal{P}(v, x) < \mathcal{P}(v)$, otherwise x would be backward reachable from x and thus in $\mathcal{G}$;
  - $0 < \mathcal{P}(v, \overline{x}) = \mathcal{P}(v) - \mathcal{P}(v, x)$ which means that $\mathcal{P}(v, \overline{x}) < \mathcal{P}(\overline{x})\mathcal{P}(v)$;
  - by all of the above $\mathcal{P}(v \mid \overline{x}) > 0$ which implies that $\alpha_{x \to v} < 1$.

  Recall now that $\lambda \to 0$, which means that $m_{x \to v} \approx \alpha_{x \to v} < 1$. CAPRESE will thus not select any of these events as cause of v if there exist an event with $m_{x \to v} = 1$, which is actually the case with genuine causes;

- genuine causes are the real cause of v, u, plus *all* the transitive backward-reachable events. Any x of these has maximum score $\alpha_{x \to v} = 1$ since:

  - $\mathcal{P}(x)\mathcal{P}(v) < \mathcal{P}(v, x) = \mathcal{P}(v)$ and $0 = \mathcal{P}(v, \overline{x})$;
  - by the above $\mathcal{P}(v \mid \overline{x}) = 0$ which implies $\alpha_{x \to v} = 1$.

Thus, CAPRESE will pick an event from $\mathcal{G}$, and not from $\mathcal{S}$. We need to show that u is the event with maximum score.

Enumerate the events in $\mathcal{G}$ as g_1 (which is u), ..., g_k in a way that

$$\mathcal{P}(g_1) < \ldots < \mathcal{P}(g_k)$$

and recall that this is a *total ordering* induced by the temporal priority, and that this is consistent with coefficient β, which means that

$$\beta_{g_1 \to v} > \ldots > \beta_{g_k \to v} \,.$$

Thus, in the limit $\lambda \to 0$

$$\max\{m_{g_i \to v} \mid g_i \in \mathcal{G}\} \overset{\lambda \to 0}{\approx} 1 + \max\{\beta_{g_i \to v} \mid g_i \in \mathcal{G}\} \approx 1 + \beta_{g_1 \to v} \approx m_{u \to v}$$

is the event closer in time to v, with respect to β. This event, namely u, is chosen by the algorithm as the real cause of v.

Finally, we show that the last step of the algorithm (the independent progression filter, step 4), does not invalidate the edge (u, v). In fact, the algorithm would replace such an edge with $(\diamond, v)$ if, for all nodes x backward-reachable from v (i.e., those in $\mathcal{G} \cup \mathcal{S}$) it was

$$\frac{1}{1 + \mathcal{P}(v)} > \frac{\mathcal{P}(x)}{\mathcal{P}(x) + \mathcal{P}(v)} \frac{\mathcal{P}(x, v)}{\mathcal{P}(x)\mathcal{P}(v)} \,.$$

It suffices thus to show that the above inequality is violated just by one of the backward-reachable nodes. We pick just $u \in \mathcal{G}$ and note that

$$\frac{\mathcal{P}(u)}{\mathcal{P}(u) + \mathcal{P}(v)} \frac{\mathcal{P}(u, v)}{\mathcal{P}(u)\mathcal{P}(v)} = \frac{\mathcal{P}(u \mid v)}{\mathcal{P}(u) + \mathcal{P}(v)} \,.$$

Also, we have that $\mathcal{P}(u) < 1$, $\mathcal{P}(v) < 1$ and, by construction, $\mathcal{P}(u \mid v) = 1$ because all the instances of v are co-occurring with those of u (but not the converse). Thus, inequality

$$\frac{1}{1 + \mathcal{P}(v)} < \frac{1}{\mathcal{P}(u) + \mathcal{P}(v)} \,,$$

is always true and ensures that edge (u, v) is maintained, which concludes the proof. □

Proof of Corollary 1 (Uniform noise).

Proof. As shown in [22], the uniform rates ϵ_+ and ϵ_- affect the observed probabilities as follows

$$\mathcal{P}(i)^* = \mathcal{P}(i)(1 - \epsilon_-) + (1 - \mathcal{P}(i))\epsilon_+ \tag{18}$$

$$\mathcal{P}(i, j)^* = \mathcal{P}(i, j)(1 - \epsilon_-)^2 + [\pi_{ij} - \mathcal{P}(i, j)](1 - \epsilon_-)\epsilon_+ + (1 - \pi_{ij})\epsilon_+^2 \,, \tag{19}$$

where $\pi_{ij} = \mathcal{P}(i) + \mathcal{P}(j) - \mathcal{P}(i, j)$. It is important to note (Lemma 1, [22]) that

$$\mathcal{P}(i) > \mathcal{P}(j) \implies \mathcal{P}(i)^* > \mathcal{P}(j)^* \,,$$

namely uniform noise is still implying temporal priority. Because of this, and since the raw estimate α is monotonic relative to temporal priority, all the derivations for Theorem 3 are still valid in this context, and the algorithm selects the correct genuine cause for each effect.

To guarantee that no valid connection is broken by the independent progressions filter, we again rely on Szabo's result (Reconstruction Theorem 1, [22]). In particular, for any correctly selected edge (u, v) in our algorithm, since we implement Desper's filter (or, analogously, Szabo's) for independent progressions we do not mistake by deleting (u, v) unless also their algorithms do. Since this is not the case when $\epsilon_+ < \sqrt{p_{\min}}(1 - \epsilon_+ - \epsilon_-)$ the proof is concluded. □

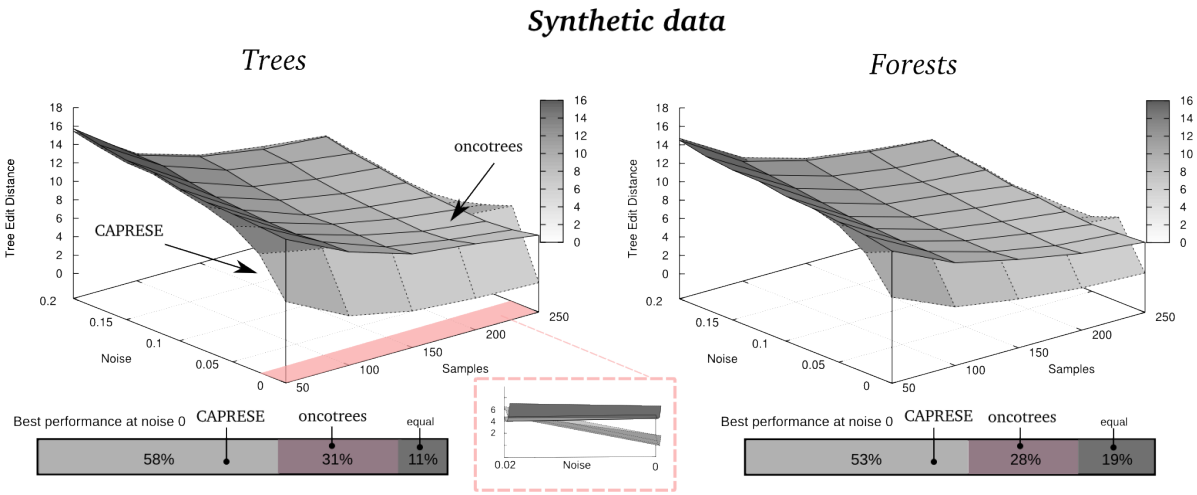


Figure 10: **Reconstruction with noisy synthetic data and $\lambda \to 0$.** The settings of the experiments are the same as those used in Figure 6, but in this case the estimator is shrank by $\lambda \to 0$, i.e., $\lambda = 0.01$. In the magnified image one can sees that the performance of CAPRESE converges to Desper's one already for $\nu \approx 0.01$, hence largely faster than in the case of $\lambda \approx 1/2$ (Fig. 6).

B.2 Synthetic data generation

A set of random trees is generated to prepare synthetic tests. Let n be the number of considered events and let $p_{\min} = 0.05 = 1 - p_{\max}$, a *single tree* with maximum depth $\log(n)$ is generated as follows:

1: pick an event $r \in G$ as the tree root;
2: assign to each event but r an integer value in $[2, \log(n)]$ representing its depth in the tree, ensure that for each level there is at least one event (0 is reserved for $\diamond$, 1 for r);
3: **for all** events $e \neq r$ **do**
4: let l be the level assigned to e;
5: assign a father to e selecting an event among those at which level $l - 1$ was assigned;
6: add the selected pair to the set of edges E;
7: **end for**
8: **for all** edges $(i, j) \in E$ **do**
9: assign $\alpha((i, j))$ a random value in $[p_{\min}, p_{\max}]$;
10: **end for**
11: **return** the generated tree;

When a forest is to be generated, we repeat the above algorithm to create its constituent trees. These trees (or forests), in turn, are used to sample the input matrix for the reconstruction algorithms, with the parameters described in the main text.

B.3 Further results

We show here the results of the experiments discussed but not presented in the main text.

Reconstruction of noisy synthetic data with $\lambda \to 0$. Although we know that $\lambda \to 0$ is not the optimal value of the shrinkage-like coefficient for noisy data, we show in Figure 10 the analogue of Figure 6 when the

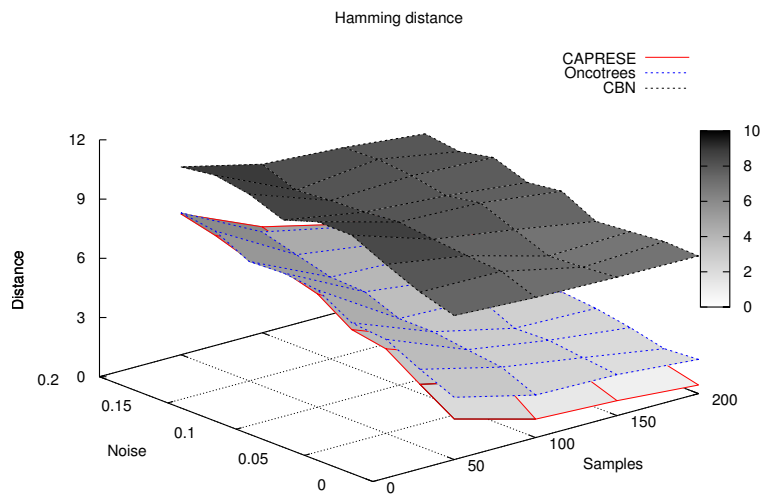


Figure 11: **Reconstruction with CAPRESE compared to oncotrees and h-CBN with noisy synthetic data.** Performance of CAPRESE compared to oncotrees and h-CBNs as a function of the number of samples and noise ν. The λ parameter used for CAPRESE is 1/2, and the reconstructed topologies contain 10 nodes each.

estimator is shrank by $\lambda \to 0$, i.e., $\lambda = 0.01$. When compared to Figure 6 it is clear that a best performance of CAPRESE is obtained with $\lambda \approx 1/2$, as suggested by Figure 2.

Comparison with hidden Conjunctive Bayesian Networks, h-CBNs. We here compare the performance of CAPRESE to hidden Conjunctive Bayesian Networks (h-CBN)[15], as well as to oncotrees. The settings of the experiment are slightly different from those of the previous analyses: we used 100 distinct random trees of 10 events each. We ranged the number of samples available for reconstruction from 50 to 200, with a step size of 50. The settings used for running h-CBNs are relatively standard settings: we allowed for 50 annealing iterations with initial temperature equal to 1. Since h-CBNs reconstruct DAGs, it is not possible to quantify its performance using Tree Edit Distance, as we did in the comparison with oncotrees. Instead, we here adopt Hamming Distance (computed on the connection adjacency matrix), as a closely related and computationally feasible alternative for measuring performance [33].

The results of the experiment can be found in Figure 11, and show that CAPRESE clearly outperforms h-CBNs. In particular, it is possible to notice that, for all the analyzed values of noise and sample sizes, both CAPRESE and oncotrees display a (average) Hamming Distance between the reconstructed model and the original tree topology that is significantly lower than h-CBNs, with the largest differences observed in the noise-free case. This result would point at a much faster convergence of CAPRESE with respect to the number of samples, also in presence of moderate levels of noise.

A few remarks are warranted about this experiment. First, in contrast to the comparison with oncotrees, we ran each experiment exactly once rather than averaging the results over 10 repetitions, and on relatively smaller trees. These limitations are a consequence of the extremely high time complexity of the simulated annealing step of h-CBNs. However, the comparison between CAPRESE and h-CBNs shows a so large difference in the performance that we do not expect this to be have significant impact. Second, the results obtained by h-CBNs are perhaps worse than expected based on results in the absence of noise presented in [24], which were however based on a unique tree topology. Yet, this outcome may have been potentially influenced by either the estimation procedure of the noise parameter in h-CBN, the adopted annealing procedure or by the used number of iterations. In future work we plan to extend our algorithm to extract more general topologies and to compare both methods in a greater detail.

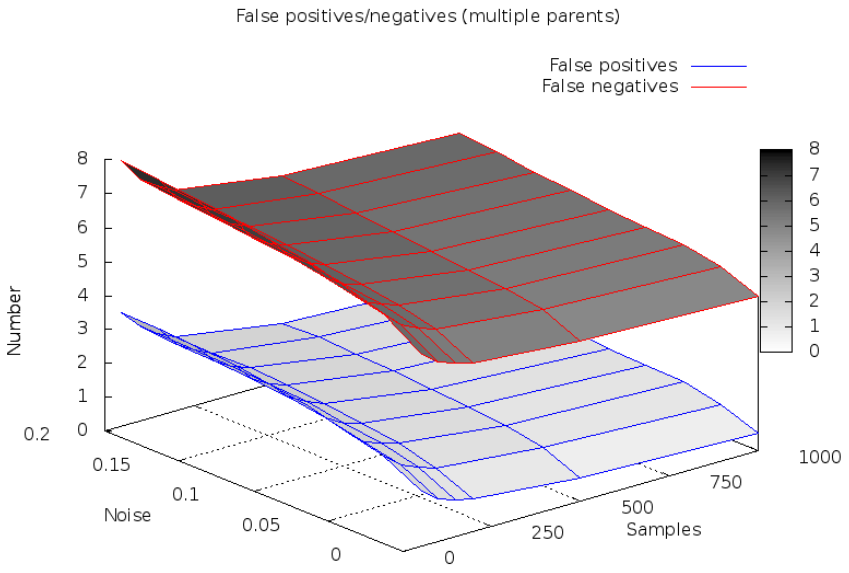


Figure 12: **Performance of CAPRESE to reconstruct models with conjunctive parents and noisy data.** Performance of CAPRESE measured in terms of the number of *false positives/negatives* in the reconstructed model, when data are generated from directed acyclic graphs with 10 nodes and where each event is caused by at most 3 *conjunctive events* (randomly assigned). The λ parameter is set to 1/2.

Inference of models with multiple conjunctive parents. CAPRESE is specifically tailored to reconstruct models with independent progressions and a unique cause for each event (i.e., trees or forests), while other approaches such as CBNs can reconstruct models where multiple conjunctive parents co-occur to cause an effect (i.e., $a \wedge b$ cause c). It is thus reasonable to use such conjunctive approaches to infer more complex model, in spite of CAPRESE.

However, it is interesting to asses CAPRESE's performance when (synthetic) data are sampled from a model with multiple parents and noise. By sampling input data from random *directed acyclic graphs* with 10 nodes and where each event is caused by at most 3 conjunctive events (randomly assigned), we assess the number of *false positives* and *false negatives* retrieved in the model reconstructed with CAPRESE. We show the results in Figure 12. Our results indicate that for increasing sample size, the number of false positives approaches 0. Thus, for sufficiently large number of samples, all the causal claims returned by CAPRESE are true. In addition, the number of false negatives is always higher and proportional to the connectivity of the target model. This is to be expected since CAPRESE assigns at most one parent (the cause) to every node.