Efficient inference of cancer progression models

Daniele Ramazzotti[1], Giulio Caravagna[1], Loes Olde Loohuis[2], Alex Graudenzi[1], Ilya Korsunsky[3], Giancarlo Mauri[1], Marco Antoniotti[1], and Bud Mishra[3]

[1] *Dipartimento di Informatica Sistemistica e Comunicazione, Università degli Studi Milano-Bicocca, Milano, Italy*
[2] *Center for Neurobehavioral Genetics, University of California, Los Angeles, USA*
[3] *Courant Institute of Mathematical Sciences, New York University, New York, USA*

August 18, 2014

Abstract

A tumor is thought to result from successive accumulation of genetic alterations – each resulting population manifesting itself with a novel 'cancer phenotype.' In each such population, clones of higher fitness, contributing to the selection of the cancer phenotype, enjoy a Darwinian selective advantage, thus driving inexorably the tumor progression to metastasis: from abnormal growth, oncogenesis, primary tumors, to metastasis. Evading glutamine deregulation, anoxia/hypoxia, senescence, apoptosis and immune-surveillance are strategies employed by the tumor population along the way to navigate through various Malthusian disruptive pressures resulting from the interplay among the evolutionary forces. Inferring how these processes develop and how altered genomes emerge is a key step in understanding the underlying molecular mechanisms in cancer development, and developing targeted approaches to treat the disease. The problem also poses many interesting challenges for computer and data scientists, primarily spurred by the availability of rapidly growing uninterpreted cancer patient data. We develop an algorithm that seeks to estimate causation by modifying statistical notion of correlation in terms of probability raising (PR) with a frequentist notion of temporal priority (TP) – thus developing a sound probabilistic theory of causation, as originally proposed by Suppes. This reconstruction framework is able to handle the presence of unavoidable noise in the data, which arise primarily from the intrinsic variability of biological processes, but also from experimental or measurement errors. Using synthetic data, we rigorously compare our approach against the state-of-the-art techniques and, for some real cancer datasets, we highlight biologically significant conclusions revealed by our reconstructed progressions.

Keywords: Cancer Progression, Data Science, Algorithms, Causation

Abbreviations: TP, Temporal Priority; PR, Probability Raising; DAG, Directed Acyclic Graph; CNV, Copy-Number Variants; CAPRI, CAncer PRogression Inference.

In the near future, cancer research is likely to become much more data-centric, primarily because of the rapid growth and ready availability of vast amount of cancer patient data, as well as because of advances in single-molecule single-cell technologies. Nonetheless, it remains impossible to track the tumor progression in any single patient over time, though emerging technology for noninvasive analysis of circulating tumor cells and cell free DNA (in blood and urine) is beginning to paint an incomplete, but useful, picture. Motivated by these possibilities, we developed an algorithm for the purpose of analyzing the currently available aggregated data from multiple patients to infer an approximate phenomenological "shape" of cancer progression, which ultimately builds on the similarities among data-points at different scales and analyzes them using tools and algorithms from probabilistic analysis, statistical inference and modal logic. In particular, our algorithm aims to infer causal relations among various mutational events occurring in the course of cancer progression, organizing them as causal networks (*Directed Acyclic Graphs*, DAGs), and ultimately, linking them to our understanding of various intra- and inter-cellular pathways (to be described elsewhere).

The approach described here seeks to understand initiation and progression of cancer in terms of "chronological" and "causal" relations among somatic alterations as they occur in the genomes and manifest as point or indel (insertion-deletion) mutations, structural alterations, DNA methylation and histone modification changes. For example, if through some initial mutations (e.g. in EGFR) a cell acquires the ability to ignore anti-growth signals, this cell-type may enjoy a clonal expansion (modeled as a discrete state of the cancers progression and marked by the acquisition of a set of genetic events). However, such a state of affairs may result in a Malthusian pressure on the population of all the cell-types in terms of deregulation of glutamine metabolism and thus, set the stage for clonal expansion of a new cell-type that can disable G1-S checkpoint (e.g., a "caused" mutation in CDK). Such causal structures is likely to be implicit in the genomic data from multiple patients, some involving tumor populations with just EGFR-mutant-cell-types and some others with a heterogeneous population with EGFR+CDK-mutant-cell-types, etc. This evolutionary (thus, Darwinian) notion of probabilistic causation is similar to that proposed earlier by JBS Haldane ("*The Causes of Evolution*." 1932.)[1]

Resulting structure may be idealized in terms of causal DAG $G = (V, E)$, where the vertices V encode the mutational events and the directed edges E describe the causal relations among the effected vertex and its causal set of parent vertices. When a vertex is connected to multiple parents, the cause may need to be described by a logical relation: e.g., *singular* (only one parent), *conjunctive* (all parent events are necessary), *disjunctive* (any parent event is sufficient), or even more complex relations (but limited to propositional or modal logic expressions, e.g., ones described in CNF, Conjunctive Normal Forms)[2]. Such a graph, of course, ignores the exact metric properties (geometry) of time and only expresses the "temporal priorities" in a topological sense. A causal graph, as described here, can construct a temporal possible-world model, which is amenable to temporal logic analysis (via model checking), thus allowing the data-scientists to propose more complex hypotheses illuminating various evolutionary forces in cancer progression.

Nonetheless, rigorous algorithmic tools to infer such causal and temporal relations from the topology of the patient data have remained largely elusive in case of cancer, which has complex heterogeneity and temporality. The main reason for this state of affairs is that information directly revealed in the data lacks precise temporal measurements but also contains large amount of irrelevant structures, complicated by heterogeneity in cell-types and non-causal passenger mutations. We present a statistical inference algorithm that performs well with a sufficiently large sample of patient (genomic) data, despite the noisy and uninformative measurements. Supp. mat.(SM) proves various correctness, convergence and complexity results for the proposed algorithm.

We base our method on a notion of probabilistic causation, more suitable than correlation in order to infer causal structures. More specifically, we adopt the notion of causation proposed by Suppes [1]. Its basic intuition is simple: event a is a *prima facie cause* of event b if (*i*) a occurs before b (*temporal priority*, TP), (*ii*) the occurrence of a raises the probability of observing b (*probability raising*, PR); when a set of events $\{a_1, \ldots, a_m\}$ are causal parents of b, the same notions are generalized in a natural way, with the introduction of some additional logic operators to combine the causal parents. Such *prima facie causes* are then further classified into *genuine* and *spurious causes*, using various auxiliary principles (e.g., *common cause principle* and controlling for *false-discovery rates*). However, as hinted earlier, since TP properties are to be imputed from the topological structure of the data, the resulting logical assertions must be tested probabilistically and approximately. Furthermore, the problem is complicated by the presence of noise, such as the one provided by the intrinsic variability of biological processes (e.g., *genetic heterogeneity*) and experimental errors. By relying on a formulation encoded in a probabilistic propositional modal logic, it is possible to devise efficient model checking algorithms to infer topologically different causal graphs (e.g., trees, forests and conjunctive DAGs). Algorithms for more complex DAGs are computationally less efficient, but can be tamed by controlling the expressivity of the underlying logic.

1 Literature Survey

There are several competing approaches to modeling cancer progression, some of which incorporate such observed effects as *cancer hallmarks*, *heterogeneity in cell-types*, *drug responses and resistance*

[1] Haldane introduced the ideas of stabilizing and disruptive selection to describe the population dynamics.

[2] See supplementary materials (SM).

development, etc. [2, 3, 4, 5]. Nonetheless, there is a need for models that focus on somatic evolutionary nature of cancer with interplay between positive and negative selective pressures to which the cells in the population respond heterogeneously, stochastically and with cellular machinery that are intricately interconnected: for instance, earlier models (e.g., Vogelstein's path-like progression model for colorectal cancer [6], or oncotree model with singular causes [7]) appear somewhat simplistic and point to the need for a language that is more expressive (e.g., the *probabilistic propositional branching-time modal logic* PCTL, that can describe time, probability as well as Boolean formulas) and for statistical-inference algorithms that are more sophisticated. In addition, because the currently available -omics data lack temporal information, these algorithms must impute the topological structure of time and separate the dynamics in terms of causal (genuine-prima-facie-causal) and chronological (spurious-prima-facie-causal) structure (see supp. mat.(SM)). We note that, among several competing notions of causality, the one that fits naturally into our framework is the one due to Suppes (Probabilistic Causality Framework) [1], and less so to the others: namely, Counter-Factual causality or Intervention-Based causality (developed extensively by Lewis [8], Pearl and students [9, 10] or Spirtes, Glymour, and Scheines [11]) or Mechanistic notion of causality (the Canberra Plan [12]). Nonetheless, our framework does not exist in isolation, as it should be fairly obvious how our hypothesized genuine cause may be verified/falsified using either intervention-based reasoning (in vitro intervention using single-cell xenograft in a mouse-model) or mechanistic analysis using systems biology tools (in silico analysis with known pathways). For instance, if the patient-data, with the imposed imputed temporal structure, points to RAS-mutants in the patient tumor genomes raising the probability of the presence of CDK-mutants in the "subsequent" patient tumor genomes, we may hypothesize and test whether xenograft of RAS-mutant single cells into a mouse would lead to a heterogenous tumor with a colony of dominant CDK-mutants.

The picture that emerges and is exploited in this paper builds on many decades of work, which would be impossible to list exhaustively: in causality: see [1, 10, 11, 13]; in progression models: see [6, 3]; in model building and model checking: see [7, 14, 5].

2 Methods

As described in more details in the supp. mat.(SM), *events* a, b, ... denote Bernoulli random variables modeling the absence or presence (taking binary values $\in \{0, 1\}$, resp.) of a genomic aberration in a population of tumor cells: e.g., point/indel mutation or a copy-number variant in a cancer sample.

Assumptions. Our framework can be derived from the following simplifying assumptions (see supp. mat., SM, for more details): (*i*) All causes involved in tumor progression are expressible as monotone Boolean formulas over events, e.g., "*a and* (*b or c*) cause *d*;" (*ii*) *All events are persistent*, i.e., acquired mutation do not disappear; (*iii*) All causally relevant events in tumor progression are observable, with the observations being able to significantly describe the progressive phenomenon (*closed world*); (*iv*) All the events have non-degenerate (prob $\neq 0$ and $\neq 1$) observed probability; and finally (*v*) All events are *distinguishable*, being neither simultaneously observed nor simultaneously missing.

Only assumption (*ii*) is specific to cancer biology, while others are somewhat technical, but applicable to other domains. Assumption (*iii*) imposes an onerous burden on the experimentalists selecting the events to study, violation of which could increase the number of spuriously inferred causal edges, and may point to new cancer-specific hypotheses that may need to be validated. Notwithstanding this obstacle, the phenomenological model of progression is likely to remain valid and usable in therapy selection. See supp. mat.(SM) for a discussion of the role of each assumption in the derivation of CAPRI's framework.

CAPRI algorithm. CAPRI requires as input a set of n *events*, e.g., mutations, in m *cross-sectional samples*, represented as a dataset in an $m \times n$ binary matrix in which an entry is: 1, if the mutation is observed in the sample; and 0, otherwise. With no other input (i.e., $\Phi = \emptyset$), the algorithm infers at most *conjunctive claims*, (e.g., "EGFR *and* KRAS cause a mutation x,") and it derives more expressive power when it is endowed with other domain-specific causal claims ($\Phi \neq \emptyset$). Thus, when more complex claims are input as logical formulas, e.g. "MYC *xor* ERB cause x", CAPRI can still test them,

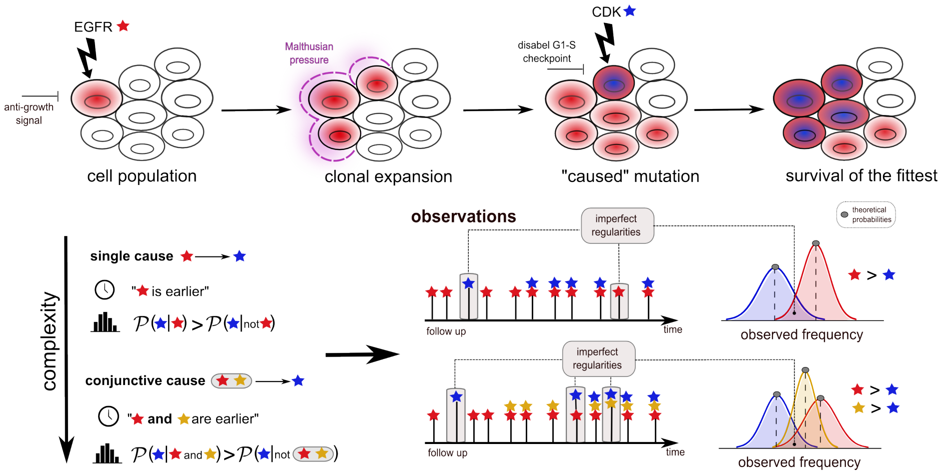

Figure 1: CAPRI algorithm examines cancer patients' genomic data to determine "causal" relationships among the chromosomal aberrations (mutations, copy number fluctuations, epigenetic modifications, etc.) that modulate the somatic evolution of a tumor. When CAPRI concludes that aberration a (say, EGFR) causes aberration b (say, CDK), it implies that the cells with a-mutation initially enjoyed a selective advantage resulting in a clonal expansion, which in turn created a Malthusian pressure (e.g., a micro-environment with deregulated glutamine) that allowed for the cells with b-mutations to emerge with higher fitness (i.e., by disabling G1-S checkpoint). Such causal relations can be succinctly expressed using Suppes' probabilistic causation: which postulates that if a causes b, in the sense described here, then a occurs before b (temporal priority) and occurrences of a raises the probability of emergence of b (probability raising). These properties are checked by CAPRI combining ideas from model checking and Bayes network theory, as illustrated in the bottom panel. Since CAPRI uses model checking it is capable of also testing complex causal claims: e.g., conjunctive causal claims.

though it may fail to generate such claims on its own. We show in the supp. mat. (SM) that CAPRI's execution is polynomial in the number of claims to test, whatever the claim type.

The algorithm derives much of its statistical power via *bootstrap with rejection resampling* [15], and the *Mann-Whitney U test* [16]. CAPRI first builds a *prima facie* directed acyclic graph (DAG), $G = (V, E)$, over events V, where each edge $i \to j \in E$ (with $i, j \in V$) is present if i occurs earlier than j, i.e., $\mathcal{P}(i) > \mathcal{P}(j)$ [3], and the probability of observing i raises the probability of j, i.e., $\mathcal{P}(j \mid i) > \mathcal{P}(j \mid \bar{i})$. CAPRI estimates its confidence in these two measures, expressing it as p-values. These conditions are *necessary but not sufficient* to identify true causal claims, thus such a DAG contains all true claims plus spurious false positives [1]. CAPRI relies upon a likelihood-based approach with the *Bayesian Information Criterion*(BIC) score to remove spurious claims; BIC scores acts like an *Occam's razor* in reducing the model complexity by combining log-likelihood fit with a *penalty criterion* proportional to the log of the DAG size via *Schwarz Information Criterion* [17]. Supp. mat.(SM) proves that CAPRI converges, asymptotically, to a model with only true positives and negatives, even in presence of uniform noise in the input data. Although many other approaches enjoy similar asymptotic properties, it was found that CAPRI, by employing prima facie causation framework, could compute accurate results with surprisingly small sample sizes.

CAPRI is implemented within TRONCO, an open source R package for *translational oncology*; more details of TRONCO appear in the SM.

3 Results

To assess CAPRI's relative accuracy (true-positives and false-negatives) and performance, we used a simulation model to create *synthetic data* (see supp. mat.(SM)), and compared CAPRI against the state-of-the-art techniques for *causal networks inference*. Among the potential competitors of

[3] This is a consequence of assumption (*ii*).

CAPRI we selected: *Incremental Association Markov Blanket* (IAMB, [18]), the *PC algorithm* [11], *Bayesian Information Criterion* (BIC, [17]), *Bayesian Dirichlet with likelihood equivalence* (BDE, [19]) *Conjunctive Bayesian Networks* (CBN, [20]) and *Cancer Progression Inference with Single Edges* (CAPRESE, [21]). These algorithms constitute a rich and lush landscape of structural methods (IAMB and PC), likelihood scores (BIC and BDE) and hybrid approaches (CBN and CAPRESE); their choice is motivated in the supp. mat. (SM).

Also, we applied CAPRI to NGS datasets of somatic mutations in *leukemia* and copy-number variants in *lung cancer* (shown as Supplementary Material).

3.1 Synthetic data

We performed extensive tests using datasets generated by simulating a synthetic progression model with (*i*) branches (a can independently cause either b or c) or (*ii*) confluences (a *and* b must co-occur to cause c), with single or multiple independent progressions to model *heterogenous cancer progressions*, and presence of *false positives* and *negatives* (i.e., noise). These variations confound the inference problem, since samples generated from such topologies will likely contain sets of mutations that are correlated but pair-wise causally irrelevant.

CAPRI's performance was calibrated against the others by *Hamming distance* (HD), *precision* and *recall*. HD measures the *structural similarity* among the reconstructed progression and the data generator (i.e., the smaller a value HD takes, the better); precision and recall measure the rates at which *true positives* and *negatives* causal claims are returned (i.e., the closer to these statistics are to 1, the better). To have a reliable statistics in all the tests 100 distinct progression models per topology are generated and, for various sample size and noise rate, 10 datasets from each topology are sampled; thus, every performance entry is the average of 1000 reconstructions.

Analyses for a variety of settings of input sample size, noise level in the input data, type of generator models and expressivity of causal claims (e.g., *disjunctive* or *mutually exclusive* progressions) are shown in the supp. mat.(SM). Here, we show in Figure 2 CAPRI's performance when 15 events are considered, confluences and a unique progression are present, a few samples are available ($m \leq 250$) and *false positives* and *negatives* are present with rate below 20%. Ranking of CAPRI relative to the state-of-the-art algorithms is shown in the same figure.

3.2 Atypical Chronic Myeloid Leukemia (aCML)

We next evaluated CAPRI's capabilities with a specific set of genomic data from aCML patients. For this purpose we relied on the experiments conducted by Piazza *et al.*, who had used high-throughput *exome sequencing technology* to identity somatically acquired mutations in 64 aCML patients, and had discovered a previously unidentified recurring *missense point mutation* seemingly targeting SETBP1 [23].

By re-sequencing SETBP1 in samples with aCML and other common human cancers, they concluded that around 25% of the aCML patients tested positive for SETBP1, while most of the other types of tumors were negative. They questioned whether it would be possible to determine a causal relationship connecting SETBP1 variants to the mutations in other aCML-specific driver oncogenes such as (e.g., ASXL1, TET2, KRAS, etc.). In particular, since SETBP1 and ASXL1 were frequently mutated together, they further asked what causal (or otherwise) relation connected these two events. The question appears somewhat puzzling, as ASXL1 mutation often presented itself either as a *non-sense point* or as an *indel* mutation.

CAPRI was able to reconstruct an aCML progression model from the datasets provided in [23] with high confidence, and was able to suggest a *potential causal dependency* among mutated SETBP1 and ASXL1. The reconstructed model is depicted in Figure 3, where *indel*, *nonsense indel*, *missense point* and *nonsense point* mutations have been causally interrelated. In particular, the figure shows that CAPRI inferred that SETBP1's missense point mutation can cause a non-sense point mutation in ASXL1, but not an indel.

A more extensive analysis (prospective study or systems biology explanation) is not yet available, but yet this example illustrates how the significance of a mutational associations can be tested using CAPRI. With hypotheses such as these, falsification/validation experiments can shed more light on somatic evolutionary dynamics in cancer.

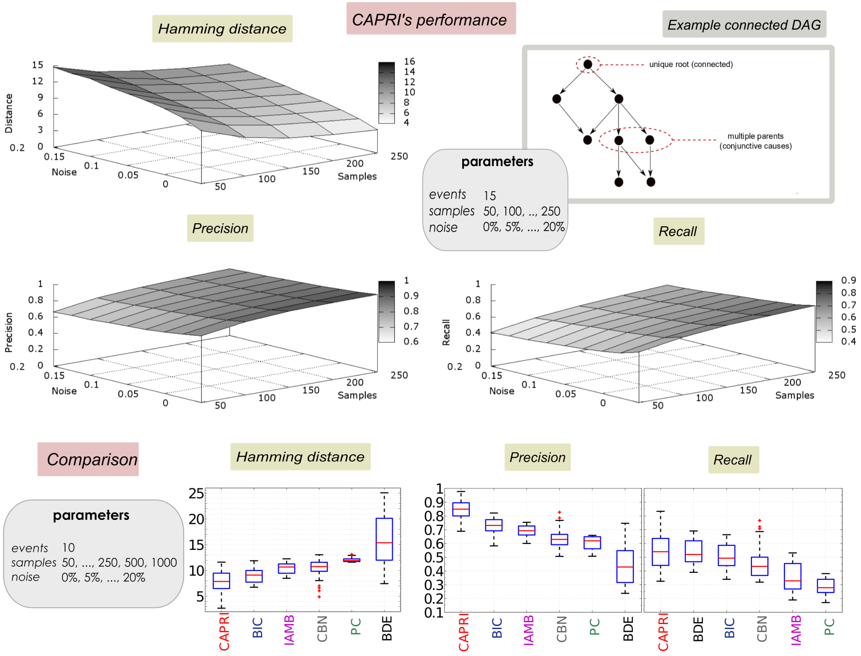

Figure 2: CAPRI's accuracy and performance was calibrated against various competing algorithms via extensive computer simulation. *Top*: *Hamming distance* (HD), *precision* and *recall* of CAPRI were assessed with synthetic data generated by DAGs (confluences, a unique progression and number of samples likely to be found in currently available databases such as TCGA [22], i.e. $m \approx 250$). Lower values of HD implies that the algorithm has mislabeled fewer genuine and spurious causes. Noise accounts for both *false positives* and *negatives*. *Bottom*: Box plot comparison of CAPRI with IAMB, PC, BIC, BDE, CBN and CAPRESE, is presented sorted according to the median performance. Extensive tests on other types of topologies are shown in the supp. mat.

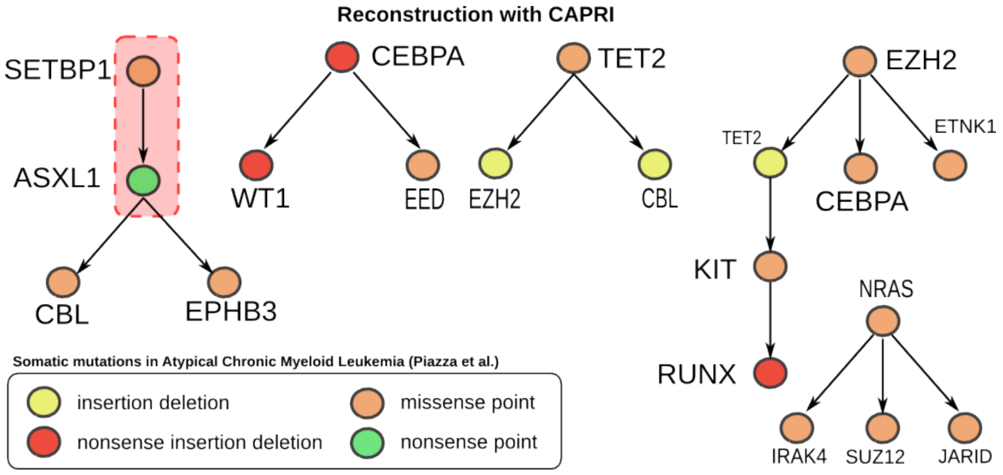

Figure 3: A progression model of *atypical Chronic Myeloid Leukemia* was inferred with CAPRI, suggesting the existence of a potential causal relation connecting SETBP1 (missense point) and ASXL1 (nonsense point), thus partially elucidating a conjectured relationship suggested by Piazza et al. Dataset of somatic mutations is available in [23]. Inference is at high confidece (p-values shown in the supp. mat.).

4 Discussions

CAPRI is a general-purpose approach for type-level causal inference from temporal data, collected over an ensemble. Its analysis is phenomenological, as it is purely data-driven without a mechanistic model to rely on. It is capable of generating hypotheses about causal relations: vast majority of which is expected to be genuine (true positive), but some may be spurious (false negatives). Since these hypotheses are expected to be interpreted and tested by domain experts, their validity (or falsifiability) ultimately depends on further analysis, either experimentally or by means of well-established mechanistic models.

As demonstrated in this paper, CAPRI is theoretically sound, algorithmically efficient and immensely practical, especially in modeling cancer progression. CAPRI is a novel algorithm in its structure and foundations, as it deviates crucially from currently popular methods, which are based on graphical models, building on statistical analysis. In contrast, CAPRI uses a set of rigorously formulated postulates (originally proposed by Suppes to develop a theory of probabilistic causation) and places the causal inference as a model checking problem, using both *temporal logic* and *probabilistic analysis*. Thus CAPRI can be shown to be not only highly expressive, accurate and powerful, but also has a sound basis in the philosophical literature tracing its roots to the antiquities, starting with the work of Avicenna (*circa* 1000 AD), Al Ghazali and Averroes, and continuing further with Francis Bacon, John Stuart Mills, David Hume, and more recently with J.L. Mackie, David Lewis, Hans Reichenbach, Patrick Suppes, Brian Skyrms, John Dupre, Nancy Cartwright, et al.

Note that the theory of causation discussed here focuses primarily on *type-level causality*: it only studies what happens statistically to a population of tumors of a specific type as it progresses in a somatic evolution – however, it is silent as to what takes place in a specific tumor in a specific patient: a question in the domain of *token-level causality*. For instance, a rapidly proliferating/growing tumor (with EGFR and CDK mutations, for instance) will have certain hypoxic inner cells: such a tumor may effect a VEGF mutation with a high probability (signaling angiogenesis), but also with some probability (perhaps to a smaller degree), effect anaerobic glycolysis or EMT (epithelial to mesenchymal transformation). Which particular path a specific tumor takes in a specific patient cannot be answered by the type-level models constructed by CAPRI, since a particular trajectory would be governed by the dynamics of token-level causality, depending on environmental variables, therapeutic history and genetics, which are specific to the patient, but obliterated in the population level statistics.

Regardless of such concerns, CAPRI's contributions are timely and critically important. Recent progress, spurred by large amount of cancer genomic data (e.g., from TCGA) points to the major roles data science plays in cancer research, *albeit* further complicated by *heterogeneity* in cell-types and *temporality* in cell-populations and cell-states in the tumor's somatic evolution. One may thus conclude that the causations in cancer evolution are unlikely to be cell-autonomous and may not even be discernible in a model of single tumor cell (or a clonal population of identical tumor cells). Instead, we may need to seek the answer in a heterogeneous population model, and through an evolutionary model of causation, in which an event that initially infers a selective advantage to a cell-type with a particular mutation becomes causal for subsequent mutations resulting in a different cell-type that thrives under the resulting Malthusian pressure in the micro-environment. Such causal connections, in our framework, are permitted to be even more complex: the "effect" may need more than a single causal event; it may need any one of many possible causal events; or it may even need an exclusively selected causal event out of many possibilities. A particularly interesting example of the last is the causation induced by "synthetic lethality," $a \oplus b \rhd c$ (a and b forming a synthetically-lethal pair). CAPRI, by virtue of being built upon the foundations of probabilistic modal propositional logic, is able to handle this situation.

However, CAPRI's expressivity often exacts a price through computational complexity: namely, an unbridled CAPRI can be computationally intractable. We have tamed CAPRI's complexity by limiting it either to just the conjunctive causal claims, or a *lifting step* that anticipates complex logical causal claims to be polynomially constrained (e.g., based on some domain knowledge). Nonetheless, it is powerful enough to test synthetic lethality, as the results of Figure 4 demonstrate.

Finally, we do not imply that CAPRI, even with unlimited amount of patient genomic data, will be able to enumerate all possible tumor progressions for all possible cancer types. Of course, CAPRI ultimately depends upon the experimentalists in selecting the observable events that are likely to be causal or indicative of cancer progression. We expect the genomic mutations (driver as well as

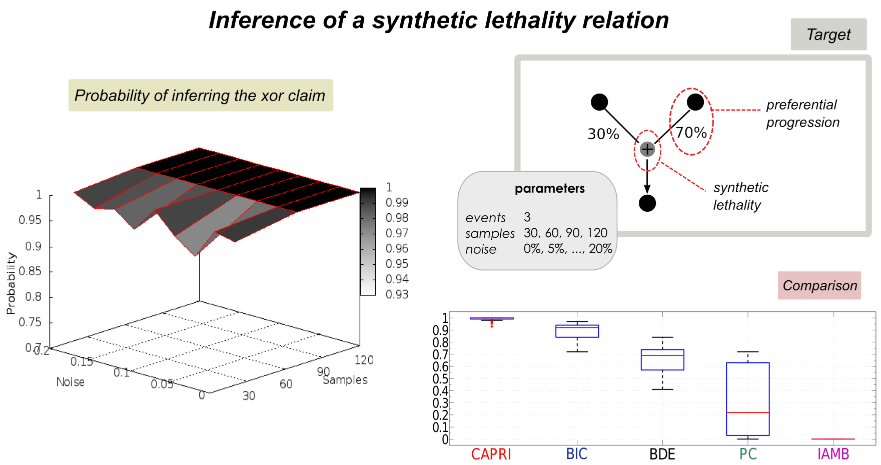

Figure 4: CAPRI was finally applied to a causation model that encodes the phenomenon of *synthetic lethality*: where a- and b-mutation individually may drive tumor's somatic evolution forward, but jointly, they prove to be lethal for tumor. It could be expressed by an *exclusive or* Boolean connective, e.g., $a \oplus b \rhd c$ and evaluated by CAPRI and other competing algorithms. The figure shows the *probability* of inferring a claim $a \oplus b \rhd c$ (*synthetic lethality*), where one progression path is preferential, and where the structure is a priori known to the algorithm. CAPRI's performance surface is shown, as well as its box-plot comparison against BIC, BDE, PC and IAMB. Results suggest that, with $m \geq 60$, CAPRI infers the correct claim almost surely. It is reasonably expected that, for more complex causal structures, comparable performance would be reached only for bigger values of m. See the supp. mat. for comparisons and additional discussions.

passenger) to fall into one or the other class. However, there are many other indicators, some of which we do know (structural, epigenetic, transcriptomic isoforms, microRNA, etc.), but others not yet known. Faced with this unpredictability, CAPRI has settled on operating at a phenomenological level trying to stitch together a causal narrative that could include some spuriousness (e.g., an unobserved common cause of two events, may spuriously lead one to conclude that one of these two causes the other). However, such hypotheses can be experimentally tested and initiate a quest for the missing common cause and its underlying mechanisms. We are continuing to develop algorithms to prioritize such hypotheses and devise experiments to validate them.

Based on such hypotheses-driven analyses, ultimately, CAPRI could lead to creation of very accurate pathway and population based models and an exact mechanical bases of evolutionary causations. Such a research program would not only deepen our understanding of biology, it is also imperative for discovery of novel cancer drugs.

Significance Recent innovations in genomics and computational technologies have made it possible to collect tumor-specific data from vast number of patients, coming from varied genetic backgrounds, wide-ranging variability in life-style and different stages of the disease. However, analysis of this data have been unusually challenging, primarily because of heterogeneity in cell-types and temporality in population structures and cell-states. We propose a rigorous method, firmly built on mathematical foundation of logic and statistics, to infer causal relations among various mutational events in the tumor populations. Resulting causal structure, encoded in a directed acyclic graph (DAG), has obvious applications in understanding therapy design, drug-resitance in cancer, survival prediction and prognosis, and drug discovery.

References

[1] P. Suppes, *A Probabilistic Theory of Causality*. North-Holland Publishing Company, 1970.

[2] R. A. Weinberg, "Coming full circle from endless complexity to simplicity and back again," *Cell*, 2014.

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

[4] M. Wigler, "Broad applications of single-cell nucleic acid analysis in biomedical research," *Genome Medicine*, 2012.

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

[6] 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.

[7] R. Desper, F. Jiang, O.-P. Kallioniemi, H. Moch, C. Papadimitriou, and A. Schäffer, "Inferring tree models for oncogenesis from comparative genome hybridization data," *Journal of Computational Biology*, 1999.

[8] D. Lewis, "Causation," *Journal of Philosophy*, 1973.

[9] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 1988.

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

[11] P. Spirtes, C. N. Glymour, and R. Scheines, *Causation, prediction, and search*, vol. 81. MIT press, 2000.

[12] D. Lewis, *Causation and Counterfactuals*, ch. Causation as Influence. The MIT Press, 2004.

[13] N. Cartwright, *Causal Laws and Effective Strategies*. Noûs, 1979.

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

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

[16] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 50–60, 1947.

[17] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, 1978.

[18] I. Tsamardinos, C. F. Aliferis, A. R. Statnikov, and E. Statnikov, "Algorithms for large scale markov blanket discovery.," in *FLAIRS Conference*, vol. 2003, pp. 376–381, 2003.

[19] D. Heckerman, D. Geiger, and D. Chickering, "Learning bayesian networks: The combination of knowledge and statistical data," *Machine Learning*, 1995.

[20] 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.

[21] L. O. Loohuis, G. Caravagna, A. Graudenzi, D. Ramazzotti, G. Mauri, M. Antoniotti, and B. Mishra, "Inferring tree causal models of cancer progression with probability raising." Submitted for publication (available at arXiv.org)., 2013.

[22] "The cancer genome atlas." `http://cancergenome.nih.gov/`, 2005.

[23] 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*, 2013.