Chapter 3
Probabilistic Computational Causal Discovery for Systems Biology

Vincenzo Lagani, Sofia Triantafillou, Gordon Ball, Jesper Tegnér and Ioannis Tsamardinos

Abstract Discovering the causal mechanisms of biological systems is necessary to design new drugs and therapies. Computational Causal Discovery (CD) is a field that offers the potential to discover causal relations and causal models under certain conditions with a limited set of interventions/manipulations. This chapter reviews the basic concepts and principles of CD, the nature of the assumptions to enable it, potential pitfalls in its application, and recent advances and directions. Importantly, several success stories in molecular and systems biology are discussed in detail.

Keywords Causality · Causal graphical models · Bayesian networks · Systems biology · Biological networks

3.1 Introduction

The winner of the 2011 ACM Turing Award—the Nobel Prize equivalent in Computing—was Prof. Judea Pearl, a pioneer in probabilistic and causal reasoning. Among many other contributions, the theory of Causal Bayesian Networks that he co-developed is now a standard tool for modeling, inducing, and reasoning with probabilistic causality. Bayesian Networks are at the heart of numerous decision support and expert systems as well as the basis for machine learning algorithms. After

Vincenzo Lagani, Sofia Triantafillou and Gordon Ball have contributed equally to this work.

V. Lagani · S. Triantafillou · I. Tsamardinos
Institute of Computer Science, Foundation for Research and Technology - Hellas,
N. Plastira 100, Vassilika Vouton, GR-700 13 Heraklion, Crete, Greece

G. Ball · J. Tegnér
Unit of Computational Medicine, Karolinska Institutet, Center for Molecular Medicine,
Karolinska University Hospital, L8:05, SE-171 76 Stockholm, Sweden

S. Triantafillou · I. Tsamardinos (✉)
Computer Science Department, University of Crete,
Voutes Campus, GR-700 13 Heraklion, Crete, Greece
e-mail: tsamard.it@gmail.com

L. Geris and D. Gomez-Cabrero (eds.), *Uncertainty in Biology*,
Studies in Mechanobiology, Tissue Engineering and Biomaterials 17,
DOI 10.1007/978-3-319-21296-8_3

several decades of heated debate about the possibility of causal discovery without—or with a limited number of—controlled experiments, it seems that consensus converges towards an affirmative answer.

Knowledge of causal relations is paramount in systems biology. Causal modelling goes beyond traditional statistical predictive modelling by allowing one to *predict the effects of actions and interventions* on a system, e.g., the effects of treating with a drug, knocking out a gene, or inducing a mutation in the genome. In contrast, non-causal, predictive modelling is only valid when the system under study is *observed* under the same experimental conditions and not otherwise manipulated. For example, gene expressions A and B may be correlated: observing the expression levels of A allow us to better predict the observed expression levels of B. But, it does not assure us that A regulates B or the opposite. The difference between observing and intervening on a system is essential for understanding causal modelling. If A is the *only* regulator of B, then the two genes are still correlated in a controlled experiment where A is activated or suppressed; in contrast, the correlation disappears in a control experiment where B is activated and suppressed at will by the experimenter, since the effect of A now becomes irrelevant.

To establish causality, one traditionally needs to perform a manipulation (perturbation, intervention) on the system [29]. In contrast, computational Causal Discovery (CD) methods argue that given certain assumptions about the nature of causality one can sometimes induce causal relations from observational data alone or a limited number of manipulations/interventions. One can then analyse archived data, forgoing expensive, time-consuming, or even impossible experiments, and determine certain aspects of the causal mechanisms. Exactly which aspects of the causal structure can be induced depends on the system under study and the available data. Given the complexity of the cell, performing all the possible experiments to establish all relations among every subset of molecular quantities, under all possible experimental conditions, is impractical. CD may provide an alternative.

Causal models (not necessarily induced through Causal Discovery) are already heavily employed in systems biology: biological pathways are a form of causal models that are indispensable in biological research. Pathways are manually assembled from the literature, where relations are established by performing interventions. However, for the most part, such models are informal and have ambiguous semantics for the edges: an edge may imply a direct or indirect causation; a missing edge may imply lack of direct causation or a yet-to-be established relation. In addition, pathways are largely qualitative; the strength and functional form of the causal relations is not represented (some exceptions exist, such as well-characterized metabolic pathways annotated with flux equations [100]). In contrast, models induced with CD methods have specific formal causal semantics as well as quantitative information that enables quantitative predictions.

In the rest of this chapter we present the basic concepts of CD, focusing on the fundamental underlying assumptions and discussing its limitations and potential pitfalls. We also present selected applications of CD in systems biology, demonstrating the potential of this exciting field.

3.2 The Nature of Causality

3.2.1 Definition of Causality

We use the notation $A \rightarrow B$ to denote our belief that "A causally affects B" (or, "A causes B" for brevity). But, what exactly does this mean and how should it be interpreted? Most of CD employs a probabilistic notion of causality. A and B should denote two well-specified variables (interchangeably: measurable quantities, features) which are measured on a population of objects, such as two protein concentrations in human T-cells. We consider simultaneous measurements of these variables in a random sample of the population, from which we can estimate their joint probability distribution. Thus, for the purposes of this chapter, the data are assumed cross-sectional: snapshots of the state of a cell without regard for the time of measurement.

$A \rightarrow B$ denotes the fact that if an experimenter intervenes and changes the values of A, the distribution of B will also change. This statement is inherently probabilistic: Average-Cigarettes-Smoked-Per-Day causally affects Presence-Of-Cancer-by-Age-60 because the distribution of Presence-Of-Cancer-by-Age-60 changes and the people with value "Yes" become more prevalent. To a single individual, that means that the probability of her getting cancer increases. Yet, causality as presently defined is still deceptively simplistic. A may be causally affecting B only in a given context, e.g., in the presence of another protein C. Thus, a better definition is probably that $A \rightarrow B$ if there is conceivable intervention involving *only* A, and a context of some other variables that are held constant, such that the distribution of B changes (relative to the distribution of B when the context is the same but A is not intervened upon). The "intervention" may be just a thought experiment, technically impossible with present technology. Yet, it has to be theoretically plausible. For example, the statement $Cancer \rightarrow Protein$ is arguably undefined: we cannot intervene on the state of the cell to make it cancerous without affecting anything else in the cell. Such semantically vacuous statements often arise when variables that refer to different abstraction levels are modeled together. In this case Cancer, a quantity that refers to the cell as a whole, and the concentration of a protein are defined on a different time and spatial scale.

Finally, notice that the concept of causation is required to define "intervention", used in the definition of causation; our definition is recursive! To break the vicious cycle, notice that intervention requires defining causality from outside the system (the experimenter) to within the system; causality as defined regards causal effects within the system. In other words, given that we understand what it means for an experimenter to intervene in a population of cells, we can define the causal relations among molecular cell quantities. We can proceed with using causality in an operational way, the same way humanity is doing statistics while still arguing about the philosophical issues of the semantics of probabilities.

3.2.2 Direct Causation

We'll need to distinguish between direct and indirect causality. We'll say A is *directly* causally affecting B relative to variables in the set $\boldsymbol{O}$, if A remains a cause of B even when all other modeled variables' values are held fixed. Direct causation is relative tothe observed variables. A hormone may directly causally affect a gene when nothing else is observed, but indirectly affect it when the status of cell membrane receptors are observed.

3.2.3 Quantitative Causality

Relations $A \rightarrow B$ are qualitative and useful for human inspection and visualization in the form of networks. But, quantitative relations are necessary to make quantitative predictions. If there are two or more direct causes of B ($A \rightarrow B$ and $C \rightarrow B$), then in general we cannot consider the relations independently. This is necessary because A and C jointly determine the values of B. In general, we can model the values of B with the *structural equation*:

$$B = f(Pa(B), U)$$

where $Pa(B)$ are the direct causes of B (or *parents* of B) and U represents all other non-modeled causes. If A and C are the only parents, then $B = f(A, C, U)$. The difference from a non-structural equation is the special role of the left-hand-side: the value of B is set (determined) by the values of PA_B and U and not vice versa: B cannot be moved to the right-hand-side. This special role of the left-hand-side is equivalent to dictating that if we intervene on the values of the right-hand-side, the left-hand-side may change, but not the other way round. The structural equation is not symmetrical. Also notice that the equation is deterministic! However, the presence of unknown values of U introduces uncertainty into the equation and induces a probability distribution of the values of B. The form of function f is important. A few examples follow, where $I(\bullet)$ is the indicator function taking values 1 when the argument holds and 0 otherwise:

- $B = a \cdot A + b + \varepsilon$, B's concentration always increases with the same rate as A increases. This is an example of a linear relation (strictly speaking, if $b \neq 0$ it is called an affine function). The term $\varepsilon = \sum_{i \in U} U_i$ is the effect of all unmeasured causes of B; it is not measurement noise.
- $B = a \cdot I(A > 100 \text{ and } C > 100) + b + \varepsilon$, B's concentration follows a baseline of b, and level $a + b$ when both A and C are larger than 100. Thus, in order to discover this relation one must observe or impose values of A *and C* larger than 100.
- $B = a \cdot (A - 100)^2 + b + \varepsilon$, B's concentration decreases as A increases if $A < 100$, and increases as A increases if $A > 100$. The rate of increase or decrease of B (as

A takes different values) is $2aA$ (the derivative of the equation) and thus it is not constant but depends on the values of A.

Linear relations are perhaps the easiest to discover, step functions (as in the second example) require observing the system in a suitable range of the parameters, and non-linear functions (as in the last example) require more sophisticated modeling approaches.

3.2.4 Necessary/Sufficient/Contributory Causes

Similarly to the distinction between necessary and sufficient conditions in logic, causes are also distinguished among necessary and sufficient, with the additional category of contributory causes:

- *Necessary*: a necessary cause is such that the effect will always imply the cause, but the cause does not imply the effect. For example, passing some course implies that you sat the examination, but sitting the examination does not imply that you will pass.
- *Sufficient*: a sufficient cause is such that the cause always implies the effect, but the effect does not imply the cause. For example, burning hydrogen and oxygen will always result in water, but the presence of water does not imply combustion.
- *Contributory*: a contributory cause is any other cause which may result in an effect, but of itself is neither necessary nor sufficient. For example, an intoxicated driver may result in a crash, but intoxication does not imply a certain crash, and neither does a crash always imply the driver was intoxicated.

The majority of cases for which causal analysis is useful concern contributory causes. Single necessary causes are usually relatively easy to identify: these are the “low-hanging fruit” for which experiment and intuition will readily recover causality without recourse to causal analysis. Conversely, a collection of mildly contributory causes is a harder problem to identify, and one for which causal methods applied to large datasets prove useful.

3.3 Basics of Causal Discovery Algorithms

3.3.1 Causal Graphical Models

A graphical representation is a useful way of quickly seeing the structure of a complex system. Intuitively, a set of causal relationships $A \rightarrow B$ can be readily represented as a graph where nodes represent quantities and directed edges represent causal relationships. Particularly, the formalism of Probabilistic Graphical Models (PGM) helps us define, in a more principled way, the mathematical characteristics of causal models. We will mainly consider the Causal Bayesian Network (CBN) and Bayesian Network (BN) frameworks, as these are some of the most known and widely employed PGMs.

3.3.2 Causal Bayesian Networks

A CBN consists of a graph $G = \{\boldsymbol{V}, \boldsymbol{E}\}$ and a parameterization $\boldsymbol{\theta}$. The set $\boldsymbol{V}$ of nodes represents the observed (modeled) quantities (a.k.a. variables), while $\boldsymbol{E}$ is a set of directed edges $A \rightarrow B$ indicating *direct*[1] causal relationships (where A is the "cause" and B is the "effect"). The graph must consist only of directed edges and contain **no cycles** (Directed Acyclic Graph, DAG). For any node A, any node that can be reached by following directed edges is a *descendant (effect)*, and any node from which A can be reached is an *ancestor (cause)*. The direct causes of a node are named *parents*, while its directed effects are named *children*. A directed path consists of a sequence of nodes where each node, except the first one, is the direct effect of its predecessor in the sequence. An undirected path is a sequence of nodes where each pair of subsequent nodes is connected by an edge without regard to the direction of the edge. Whenever an undirected path $\{A \rightarrow C \leftarrow B\}$ exists with two incoming edges into C, the node C is called *collider on this path*.

The parameterization $\boldsymbol{\theta}$ defines the joint probability distribution of data generated by a system with the causal structure of the network. The parameterization quantifies the functional form of the causal relations among the variables. Adding a parameterization allows us to express whether relationships are linear or not, the effect size of each interaction, and in general to make quantitative inferences. For a discrete joint distribution (all variables being discrete) there is one parameter θ_i for each combination of values of the variables:

$$P(V_1 = v_{i_1}, \ldots, V_n = v_{i_n}) = \theta_i \tag{3.1}$$

A major assumption of the CBN framework is the **Causal Markov Condition**: *each node of V is independent of its non-descendants (non-effects) given its parents (direct causes)*. In other words, the Causal Markov Condition asserts that indirect causes or confounded quantities do not provide additional information for a variable, once the values of the direct causes are known. Notice that, *effects* of V *may* provide additional information, even when all direct causes of V are known. The Causal Markov Condition allows us to connect the causal structure (network) with the distribution parameters. By the chain rule in probabilities we obtain:

$$P(V_1 = v_{i_1}, \ldots, V_n = v_{i_n}) = \prod_j P\left(V_j = v_{i_j} \middle| V_1 = v_{i_1}, \ldots, V_{j-1} = v_{i_{j-1}}\right) \tag{3.2}$$

[1] Direct causation is defined in the context of all other modeled variables, i.e., a causal relation mediated by none of the observed variables.

Let's assume without loss of generality that each variable in the equation above is listed after its parents, i.e., if $i < j$, then V_j cannot be a parent of V_i (in other words, we assume the variables are topologically sorted). Notice that this is always possible for DAGs. By using the Causal Markov Condition in the above equation, we obtain:

$$P(V_1 = v_{i_1}, \ldots, V_n = v_{i_n}) = \prod_j P\left(V_j = v_{i_j} \middle| Pa(V_j) = pa_{qj}\right) \tag{3.3}$$

That is, due to the Markov Condition for each variable V_j, all the variables in the conditioning part have disappeared except the parents of V_j, denoted with $Pa(V_j)$. The quantity pa_{qj} denotes the joint combination of values of the parents of variable V_j. The causal structure now dictates the form of the joint probability distribution, by entering the equation in the form of the parent sets $Pa(V)$ for each variable and imposing a factorization of the joint distribution. Employing Eq. (3.1) to represent an arbitrary distribution with n binary variables requires $2^{\mathrm{n}} - 1$ parameters θ_i to be specified. However, using Eq. (3.3), we only need to represent the distributions $P(V_j = v_{i_j} | Pa(V_j) = pa_{qj})$ for each variable. If a causal system is sparse, e.g., each variable has at most 3 parents, then we need $(2-1) \cdot 2^3$ parameters for each such conditional distribution. So, in total, we need at most $n \cdot (k-1) \cdot k^p$, where n is the number of variables, k the maximum number of values of each variables, and p the maximum number of parents of a variable: *knowledge of the structure of the causal network, assuming it is sparse, allows an exponential reduction in the number of distribution parameters required, and hence the dimension of the parameter space.*

CBNs also assume **Causal Sufficiency**, which corresponds to asserting that there are no external variables which are causes of two or more variables within the model. These common causes are called *confounders*. The Causal Sufficiency assumption implies that the following sub-graph is not present in the causal system under study: $X \leftarrow L \rightarrow Y$, where X and Y are modeled, and L is unobserved and not

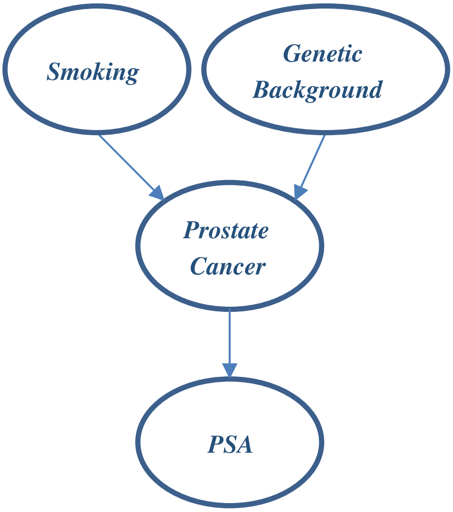


Fig. 3.1 A simple graphical model depicting the (supposed) causal relationships among smoking, genetic background, cancer and Prostate-Specific Antigen (PSA). The parameterization of the distribution associated with this network is described in the text

modeled. A truly causally sufficient model is in practice hard to construct, especially in molecular biology where hundreds of thousands of molecular quantities may be confounding factors. Causal Sufficiency is one of the most restrictive assumptions of causal discovery. Fortunately, there exist more advanced generalizations of CBNs that admit latent confounding variables (see Sect. 3.6).

Let's employ an example in order to better explain the concepts above. Figure 3.1 portrays the DAG of a small CBN. In this example, the probability of developing *Prostate Cancer* is influenced by both *Genetic Background* [94] and *Smoking* [48], while the presence of prostate cancer increases the probability of deregulation in the expression of the Prostate-Specific Antigen (*PSA*, [50]). Let's suppose that all variables are binary, which means each variable can assume a value in the set $\{1, 0\}$. Regarding the *semantics* of the values, "1" means, respectively, *deregulated* for *PSA*, *harmful* for *Genetic Background*, *yes* for *Smoking* and *present* for *Prostate Cancer*. In each variable the value "0" negates the meaning of value "1". We can now parameterize this simple model as follows:

$$P(Smoking = 1) = \pi_{Smoking}$$
$$P(Genetic\ Background = 1) = \pi_{Genetics}$$
$$P(Prostrate\ Cancer = 1|Smoke, Genetics) = a_1 \cdot Smoke + a_2 \cdot Genetics + a_0$$
$$P(PSA = 1|Prostate\ Cancer) = a_4 \cdot prostrate\ Cancer + a_3$$

In this parameterization, having a harmful genetic background and being a smoker are modelled as random events, whose respective probabilities are $\pi_{Smoking}$ and $\pi_{Genetics}$. Coefficients a_0, a_1, a_2 quantify the extent to which *Smoking* and *Genetic Background* affect the probability of developing cancer, while a_3, a_4 quantify how *Prostate Cancer* changes the probability of *PSA* being deregulated.

Notice that all causal relationships are *probabilistic* (non-deterministic), i.e., *Smoking* and *Genetic Background* increase the *probability* of developing *Prostate Cancer*, while the presence of cancer *may* deregulate *PSA* expression. The probabilistic nature of the model is due to the existence of a number of factors $U_i, i \in U$ (e.g., physical activity, diet, medications, etc.) which influence the model's quantities but are not measured. However, recall that the *Causal Sufficiency* assumption requires that no external factor simultaneously influences two or more elements of the model; this means that each variable can be affected by multiple U_i, but each U_i can affect only one variable.

3.3.2.1 Inference in Causal Bayesian Networks

If the CBN is known (this includes both the structure and the parameterization), *any probabilistic inference is possible*. In particular, any predictive or diagnostic query of the form "what is the probability V_i will take/has taken value j given that we observed certain values for other variables" is possible. Without loss of generality let's assume we observed $V_1 = v_1, \ldots, V_k = v_k$ and we would like to compute the conditional probability that $V_{k+1} = v_{k+1}$:

$$P(V_{k+1} = v_{k+1} | V_1 = v_1, \ldots, V_k = v_k) = \frac{P(V_1 = v_1, \ldots, V_k = v_k, V_{k+1} = v_{k+1})}{P(V_1 = v_1, \ldots, V_k = v_k)}$$
$$= \frac{\Sigma_{V_{k+2},\ldots,V_n} P(V_1 = v_1, \ldots, V_k = v_k, V_{k+1} = v_{k+1}, \ldots, V_n = v_n)}{\Sigma_{V_{k+1},\ldots,V_n} P(V_1 = v_1, \ldots, V_k = v_k, V_{k+1} = v_{k+1}, \ldots, V_n = v_n)} \tag{3.4}$$

where each index v_i runs on all values in the domain of variable V_i. Each term in the sums is computed by Eq. (3.3). Let's resume the example presented in Fig. 3.1, and assume that a specific patient has a deregulated *PSA*, is a smoker, and his *Genetic Background* is not harmful. A clinician may be interested in assessing the probability that the patient has *Prostate Cancer*, which can be evaluated by applying in turn Eqs. (3.3) and (3.4):

$$P(Cancer = 1 | Smoke = 1, PSA = 1 Genetics = 0)$$
$$= \frac{P(Cancer = 1, Smoke = 1, PSA = 1, Genetics = 0)}{\Sigma_{pc=0,1} P(Cancer = pc, Smoke = 1, PSA = 1, Genetics = 0)}$$
$$= \frac{P(Genetics = 0) \cdot P(Smoke = 1) \cdot P(Cancer = 1 | Smoke = 1, Genetics = 0) \cdot P(PSA = 1 | Cancer = 1)}{\Sigma_{pc=0,1} P(Genetics = 0) \cdot P(Smoke = 1) \cdot P(Cancer = pc | Smoke = 1, Genetics = 0) \cdot P(PSA = 1 | Cancer = pc)}$$
$$= \frac{(a_1 \cdot 1 + a_2 \cdot 0 + a_0) \cdot (a_4 \cdot 1 + a_3)}{(a_1 \cdot 1 + a_2 \cdot 0 + a_0) \cdot (a_4 \cdot 1 + a_3) + [1 - (a_1 \cdot 1 + a_2 \cdot 0 + a_0)] \cdot a_3}$$

Assuming that *Smoke* sensibly increases the probability of cancer ($a_1 = 0.2, a_0 = 0.01$) and that *PSA* has a high sensitivity ($a_4 = 0.9, a_3 = 0.1$), the patient has a high probability (0.727) of having *Prostate Cancer*. A similar inference would have also been possible in the case information regarding *Genetic Background* was not available, though the sums would have contained more terms (the number of terms grows exponentially with the number of unobserved variables). In general, inference is in the worst case an NP-complete problem, however efficient exact or approximate algorithms do exist [71]. *Thus, a CBN can predict/diagnose any variable or set of variables given the values of any other set of variables*. It is like having trained an exponential number of predictive models, one for each variable subset as predictors. This is a key factor that has made CBNs popular in (clinical) decision support systems where one may have a varying and limited number of observations for each patient.

The graph of a CBN can also provide all the (conditional) independencies implied by the Causal Markov Condition. If *faithfulness* is assumed (see Sect. 3.4.1 for a definition of faithfulness) the graph can also provide all (conditional) dependencies. In other words, by examining the graph, one can determine which variables are conditionally or unconditionally correlated. The property that connects the topology of graphical/causal structure with the concept of conditional (in)dependence is called *d-separation*; two sets of variables $\boldsymbol{A}, \boldsymbol{B}$ (such that $\boldsymbol{A} \neq \boldsymbol{B}$) are conditionally independent given a third set $\boldsymbol{C} \subseteq \boldsymbol{V} \setminus \{\boldsymbol{A}, \boldsymbol{B}\}$ if and only if they are d-separated by $\boldsymbol{C}$ in G. Formally, d-separation is defined as follows: $\boldsymbol{A}, \boldsymbol{B}$ are said to be d-separated given a third set $\boldsymbol{C}$ if there is no undirected path U such that (i) every collider in U has a descendent in $\boldsymbol{C}$ and (ii) no other nodes in $\boldsymbol{C}$ is in U. Intuitively, we can think about d-separation as a criterion that tells us if the "flow of information" between two variables A and B is interrupted or not. For example, variables *Smoking* and *PSA* in Fig. 3.1 are d-connected when conditioned on the empty set (the "flow" of information can freely transfer from *Smoking* to *PSA* through the node *Prostate Cancer*), but they are d-separated when we condition on *Prostate Cancer*, since knowing its value makes the information provided by *Smoking* superfluous in order to predict *PSA*. On the other hand, *Smoking* and *Genetic Background* are d-separated in absence of a conditioning set, but they become d-connected when conditioning on *Prostate Cancer*, *PSA* or both. In fact, in the latter case we condition on all the colliders (or their descendants) in the undirected path between *Smoking* and *Genetic Background*. Note that *Smoking* and *Genetic Background* are independent, but knowing both the values of *Smoking* and *Prostate Cancer* gives us some information on the value of *Genetic Background*, and thus the two variables are not independent anymore. Hereafter we will denote with $dep(A, B|\boldsymbol{C})$ the presence of a conditional dependency between variables A and B given the set $\boldsymbol{C}$, while $indep(A, B|\boldsymbol{C})$ will denote independence.

Finally, CBNs can make inferences unique to causal models: *they can predict the effects of interventions/manipulations/changes in experimental conditions*. Given a CBN we can determine the effect of knocking out a gene, the effect of administering a drug, or the effect of changing any quantity modeled in the network. Conceptually, such inferences are straightforward. The effect of the experimenter on the system that sets the values of a variable V_k to v, removes the effect of any other variable to V_k. This is modeled by removing all incoming causal edges to V_k and setting $P(V_k = v) = 1$ and $P(V_k = v') = 0$, for $v \neq v'$ in the conditional probabilities associated with the graph. The edge removal is called *graph surgery*; in the resulting graph V_k will have no parents. The new joint probability distribution can now be computed with Eq. (3.4), and hence any probabilistic query about the effect of the intervention can also be computed. Interventions that deterministically set the values of some specific variables are called *hard interventions*. When interventions have a chance of not being effective, they are called *soft interventions.* In this case, the intervention does not completely remove the causal effect of all other quantities, and thus, a different treatment is necessary where the probability of effective intervention is also modeled. In addition, when an intervention is not specific to a quantity but may affect other quantities too, the intervention is called a *fat-hand intervention* and also requires different modeling techniques [25].

The main reason for causal modeling and discovery is exactly to enable the prediction of the effect of our actions onto the system. Causal models are the only types of models that enable such inferences. Statistical causal models, such as CBNs perform such inferences without modeling the underlying physical phenomena and mechanisms of causality, while other models such as Ordinary Differential Equations directly model these mechanisms.

3.3.2.2 Dropping the Causal Semantics: Bayesian Networks

It is rarely the case that a CBN can be constructed completely from prior knowledge. Typically, such models have to be learnt from data by algorithms that make numerous assumptions (see Sect. 3.4.1 for a discussion). In cases when the causal assumptions are not to be trusted, and the structure or parameters of the CBN is also not trusted, one may still use the Bayesian Network framework without the causal semantics. Similarly to CBNs, BNs consist of a DAG and a parameterization, but do not make any causal claims or causal predictions. The Causal Markov Condition can be substituted with the Markov Condition: each node is independent of its non-descendant given its parents. Note the substitution of "direct causes" with "parents", since in BNs the term "cause" does not make sense anymore.

An *edge* $X \rightarrow Y$ in a BN should be interpreted strictly from a probabilistic viewpoint: *the edge denotes that* X *provides unique information for* Y *(possibly given some other variables) and vice versa*. The direction of the edges should also not be interpreted causally; the direction is only employed to combine edges into paths and determine implied probabilistic dependencies and independencies from the network with the d-separation criterion. All probabilistic inferences possible with CBNs are also possible with BNs, except for causal inferences: if the causal semantics are dropped, one is not entitled to employ a BN to predict the effect of manipulations into the system. BNs may predict our future observations based on past observations, but not the effects of our actions.

Notice that, Bayesian Networks are only loosely related to some other Bayesian statistical approaches in this volume, for example Bayesian model selection (See Chapters [37, 49, 101]). Bayesian Networks treat probability in a "Bayesian" way, i.e., to represent measures of belief (in contrast to the frequentist interpretation of probabilities). They also make heavy use of the Bayesian Theorem to make inferences. Both of these characteristics justify the term Bayesian. Bayesian model selection also treats probabilities as measure of belief; in particular, Bayesian model selection uses probability distribution on the set of possible models to express the a priori belief on their validity (typically, favoring simpler models). However, Bayesian Networks serve to model and make inferences about joint distributions, while Bayesian model selection aims to select the statistical model that achieve the best trade-off between fitting the data and abiding to the prior beliefs.

3.4 Causal Discovery Approaches

The main goal of Causal Discovery algorithms is reconstructing the network of causal mechanisms underlying a given system, given a dataset D. The dataset D is usually composed of a set of n observations measured over m variables. Such causal learning algorithms have already proven useful to biologists as shown below in Sect. 3.5.

Unfortunately, reconstructing a Causal Model from data is not an easy task. Several algorithms have been proposed in the last few decades, and all of them consist of two stages: firstly, an appropriate causal graph is found, and secondly a parameterization is estimated in accordance with the graph structure. While the second stage is relatively straightforward (given a suitable assumption about the functional form of the causal relationships), identifying the correct causal graph has proven to be NP-hard [14]. So far, two main approaches have been developed for reconstructing the graphs of Causal Models, namely the *Constraint-based* and the *Score-based* (also known as *Search-and-Score*) approach. The basic principles of the two main approaches for learning CBN are the following:

- *Score-based*: first introduced in [78], these methods are based on a score function $S(G|D)$ measuring the fit of the graph to the data, while at the same time favoring simpler structures. A prototypical score function is the Bayesian information Criterion (BIC), defined as $BIC(G|D) = |\boldsymbol{\theta}| \cdot \ln(n) - 2 \cdot \log(\boldsymbol{L}(G|D))$, where $\log(L(G|D))$ is the log-likelihood of the graph given the data, and $|\boldsymbol{\theta}|$ is the number of model parameters [93]. The score function is usually combined with a search-heuristic that explores the space of possible graphs. A typical heuristic is the greedy one: start with the empty graph (no edges) and then add, reverse or delete the edge that maximally increases the score of the network (i.e., the fit to the data) at each step.
- *Constraint-based*: this approach relies on estimating some of the conditional (in)dependencies in the data distribution P from the data D through performing hypothesis tests of conditional independence. Typically, for discrete variables the X^2 or the G^2 tests are employed [63], while for continuous variables testing the partial linear correlation are employed based on the Fisher z-transformation [28]. The results of the hypothesis tests constrain the graph to reconstruct: in the resulting graph G, two variables X, Y should be d-connected given Z if and only if $indep(X; Y|Z)$ in the data. In fact, it can be proven (assuming faithfulness, as defined in Sect. 3.4.1) that two variables are connected by an edge if and only if there is no set of variables Z, such that $indep(X; Y|Z)$. Constraint-based methods usually start with a fully connected, undirected graph and progressively remove edges whenever a new conditional independency is discovered [98].

Typically for a given dataset there will be multiple solutions (i.e., networks) that are Markov equivalent, i.e., they imply the same set of conditional independencies and thus cannot be distinguished based on testing independencies on the data. Under typical scoring functions, these networks receive equivalent scores. Intuitively, each such network provides an equally good causal explanation for the data. *The issue of Markov Equivalence in learning causal structures is a point that an analyst should keep in mind*. The set of equivalent networks has some invariant characteristics, e.g., edges and directions upon which all solutions agree, and some variant characteristics upon which different solutions disagree. Even when all causal assumptions hold, *the analyst is warranted to make claims only about the invariant characteristics*. Fortunately, for CBNs the representation of the set of equivalent networks is compact: they can be represented with another type of network called the Completed Partially DAG (CPDAG) or essential graph [13] and the invariant characteristics can be identified from this graph. Particularly, CPDAGs contain two types of edges, directed and undirected. The first type represents arcs that are similarly (invariantly) oriented in all Markov Equivalent solutions, while the latter represents edges whose orientation varies among equivalent networks. In more complicated causal formalisms discussed in Sect. 3.6, the set of equivalent solutions cannot be compactly represented. See also Chapters [37, 49, 101] for further discussions on model identifiability and (Bayesian) model selection.

Causal Discovery algorithms can also be used for variable selection, i.e., identifying the smallest subset of quantities that can provide the optimal prediction or diagnosis for an outcome variable of interest T (*this is equivalent to molecular signature identification*). Under certain conditions, the set of variables optimally predicting the value of an outcome or molecular quantity T is what is called the Markov Blanket of T: the set of direct causes, direct effects, and direct causes of direct effects [108]. Algorithms that can identify the Markov Blanket of T from data without knowledge of the underlying CBN exist and have proved to be some of the most effective signature identification algorithms from biological data [3]. *Importantly, these theoretical results connect molecular signatures for prediction or diagnosis with the causal structure of the system under study: the most predictive quantities have a specific causal interpretation.*

3.4.1 A Discussion on Some Typical Causal Discovery Assumptions and Practical Issues

We now focus in more detail on the most common assumptions of typical causal discovery algorithms and discuss their implication in the context of causal discovery in biological systems.

(*Causal*) *Markov Condition*: in a DAG G each node is independent from any non-descendant (non-effect) given its parent (direct causes). This condition formalizes our "common belief" about how Causality operates, i.e., indirect causes or confounded effects do not provide additional information, once the direct causes are known. For example, in the network $X \leftarrow W \rightarrow Y \rightarrow Q \rightarrow R$, we expect that once we know Y (the direct cause of Q), neither X (a confounded variable) nor W (an indirect cause) provide additional information for Q. Notice that, observing the effect R of Q still provides additional information for the value of Q. Interestingly, while the Causal Markov Condition is (explicitly or implicitly) accepted and employed "all the time in laboratory, medical and engineering setting" ([98], p. 38), whether it holds in the sub-atomic systems studied by quantum physics it is still under debate [70]. This assumption is what allows the algorithms to discover direct causal relations and drop edges from the causal network being reconstructed. While relatively uncontested in practice, *the Causal Markov Condition may appear to be violated due to measurement error* (*see below*).

Acyclicity: CBN and other PGMs assume that no node in the graph can be a cause of itself, either directly or through other variables. CBNs are not able to represent feedback loops, and in some biological applications this limitation can be quite restrictive. However, some approaches have been developed that do not require this assumption [41]. Typically, formalisms that admit the presence of feedback cycles assume only linear relations. In the presence of both non-linear relations and feedback chaotic phenomena may arise that significantly complicate the problem and the applicable algorithms. Thus, one must substitute one assumption for the other to make causal discovery possible.

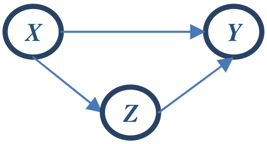


Fig. 3.2 Simple example of a feed-forward network

Causal sufficiency: no pair of nodes shares a common, unmeasured cause. In statistical terms, we assume that there are no latent confounders that may introduce correlations that are not explained by the measured variables. Specifically, consider the network $X \leftarrow L \rightarrow Y$. Because L is a confounder of X and Y, we expect the latter pair of variables to be correlated (dependent). If L is not observed, it is not modeled in the network. There are only two networks with variables X and Y that entail a dependency and fit the data: $X \rightarrow Y$ and $X \leftarrow Y$. Both of them are Markov equivalent and correctly represent the data distribution. But, their causal semantics are wrong: X does not cause Y nor vice versa. The network with the correct causal semantics is "X (no edge) Y". There is no way to correctly simultaneously represent both the probabilistic semantics and the causal semantics of the network without admitting new, unobserved variables in the network. Causal Sufficiency is one of the most restrictive assumptions in CBNs particularly for systems biology where there are millions of possible molecular quantities that may be confounding the observed quantities. For this reason some PGM frameworks have been recently developed (e.g., Maximal Ancestral Graphs [85]) that generalize CBNs to admit and reason over the presence of hidden confounders.

Faithfulness: a distribution P is faithful to a DAG G if it entails all and only the conditional independences implied by G. This assumption turns out to be important particularly for the efficiency of Causal Discovery algorithms, in order to search and identify all solution networks. One interpretation of faithfulness is that the set of conditional independencies is *stable* under infinitesimal perturbations of the data distribution [79]. For example, consider the following feed-forward gene network (Fig. 3.2):

X regulates Y both directly as well as indirectly through Z. The two paths for regulating Y may be competing with each other: X up-regulating Y directly, X up-regulating Z which in turn down-regulates Y. If the causal effects of each regulation are just so finely tuned it is possible that the association between X and Y completely disappears even though X causes Y. Such fine tuning of the parameters of the distribution seems unlikely (and it is infinitely unlikely under certain assumptions, see [98], p. 66) and leads to independences that are unstable: they become dependencies if the parameters of the distribution are slightly perturbed. Faithfulness dictates that this fine tuning is not present in the data distribution. Thus, whenever X causes Y in a network directly or indirectly or through multiple causal paths, we assume the variables are dependent.

Faithfulness seems innocent at first glance, but there are several pitfalls. First, in practice a distribution may be faithful but "close to unfaithfulness"; in the example above, the association between X and Y may not disappear completely but may be too small to be detected with typical sample sizes. Second, while fine-tuning of the parameters seems unlikely to occur by chance, there is evidence that natural selection leads to systems which may be unfaithful; in particular, the presence of negative feedback cycles may lead to associations that disappear [22]. Deterministic relations also violate faithfulness! It seems that randomness (i.e., natural occurring perturbations) is required to allow causal discovery, which is philosophically intriguing to say the least. For example, consider the network $X \rightarrow Y \rightarrow Z$, where X and Y are deterministically related, e.g., they always have equal values. In that case, X provides for Z the same information as Y and so $indep(Y; Z|X)$ holds which is not entailed by the Markov Condition. There are algorithms that do not assume faithfulness for learning CBNs [54]. However, simultaneously dropping the acyclicity and faithfulness assumptions requires sophisticated theory and algorithms [41].

There are some additional assumptions that are often not declared explicitly, but that should be carefully taken into consideration:

No measurement error: the variables are measured without error. This is a subtle assumption that is required to learn CBNs, often not realized by practitioners who apply these techniques. In other words, to allow causal discovery we need to assume that the variance of the measurements of a variable X stems from our uncertainty about (marginalizing over) all other causes of X, and is not due to measurement error. Consider the effect of measurement error: let's assume we measure $X' = X + \mathrm{e}_{\mathrm{X}}$, $Y' = \mathrm{Y} + \mathrm{e}_{\mathrm{Y}}$, $Z' = Z + \mathrm{e}_{\mathrm{Z}}$, where the last terms are the measurement noise terms. Let's assume the true structure is $X \rightarrow Y \rightarrow Z$. Thus, based on the Causal Markov Condition we expect that indep(X; $Z|Y$). However, we observe the noisy versions of the variables, so what we test instead is indep(X'; $Z'|Y'$). If the variance of e_Y is larger than e_X, it may turn out that X' does provide additional information for Z' given Y'. This is equivalent to the Causal Markov Condition being violated. A more relaxed assumption is that all error terms have the same variance, which would lead to noisy versions of the variables that still maintain the same independencies as the true, underlying network involving only the original variables. *This observation is particularly important for measurements by biotechnologies that do have significant measurement error, such as micro-array gene expression data*, *where gene expression may have very different variance of measurement errors.*

Effect of data transformations (discretization, averaging): as above, this issue regards the connection between the actual quantities that we are modeling and the quantities measured and contained in the data. For example, it is common for a practitioner to discretize the data before applying a causal discovery method. However, depending on the discretization, the set of dependencies and independencies in the transformed data distribution may be changed compared to the original [61, 68]. Again, this may appear as a violation of the Causal Markov Condition on the transformed data. *Another important case of potentially harmful transformation is that of averaging*. *Averaging takes place in almost every mass-throughput technology.* For example, in micro-array gene expression data one tries to induce causal relations and networks among *gene expressions in a single cell*, e.g., that $X \rightarrow Y$. However, *what is measured in the data are the average expressions $\bar{X}$ and $\bar{Y}$ of X and Y in millions of cells*. The independencies of a network is $X \rightarrow Y \rightarrow Z$ defined on the single-cell quantities X, Y, Z are not necessarily the same as the independencies on the averages $\bar{X}$, $\bar{Y}$, and $\bar{Z}$ [15]. *This observation favors causal discovery from single cell data* (or in general, measurements that are not averaged) versus other biotechnologies.

No selection bias and case-control studies: A basic assumption for causal discovery is that the samples are not selected for inclusion in the data based on an effect of the modeled variables. Let's consider the case where two genes X, Y regulate the size of the cell Z: when both genes are high the cell is larger with high probability. In addition, we assume the two genes to be independent from each other. Thus, the true network is $X \rightarrow Z \leftarrow Y$. Now, let's imagine that a researcher measures these genes in a collection of cells including mostly in large cells (perhaps because small cells are harder to detect and isolate given the available equipment). In the selected population whenever X is high, Y is also high with large probability: the two gene expressions are correlated in the selected population. *This correlation is an artifact of the data sampling* and not present in the general cell population. *Cytometry data is a particular type of data with possible selection bias as an effect of the gating process and classification to different cell types*. *Another striking example of selection bias is case-control data*. In case-control studies, half the samples (cases) have been selected for inclusion based on the effect (disease) of the modeled variables. In the previous example, let us change the semantics of Z to being the presence or absence of a disease and X, Y two independent causes of disease. In all cases of disease, when X is high, Y is high with high probability, so they appear correlated even though they are not correlated in the general (unselected) population. Epidemiologists try to alleviate these spurious correlations by *matching* cases and controls based on some of the variables (age, gender, race, etc.). If cases and controls are matched in the example above, the spurious association between X and Y would disappear. However, matching cannot be achieved at a molecular level for every variable (e.g., gene expression) that is modeled and so one has to be particularly careful with causal discovery in case-control data. Some methods for learning causal networks [7] try to account for selection bias introduced by unmatched case-control study design.

It should also be noted that standard Causal Discovery algorithms assume that samples are independent and identically distributed (i.i.d.) and that they are all measured under the same experimental conditions and at same the point in time (cross-sectional data). Other algorithms exist for dealing with other types of data and information, e.g., data measured under different experimental conditions [19, 51], in different points in time [30] or for co-analyzing data in the context of prior causal knowledge [8].

Finally, *practical issues* also determine the success of causal discovery:

Statistical errors: statistical errors in the results of the conditional independence tests, or equivalently statistical fluctuations in the score of networks to the data may result in learning networks or relations that are wrong. In fact, in a large network it is almost certain that some parts of the network will be erroneously induced. Robustness against statistical errors and sample sizes depends on the learning method. Employing the most appropriate hypothesis testing procedure or scoring function for the given data is paramount. Inappropriate tests or score functions may introduce systematic reasoning errors. For example, if functional relations are non-linear but linear hypothesis tests are used, some conditional (in)dependencies may not be detectable even with large sample sizes. Methods for assessing the reliability of each feature of the network (e.g., presence of an edge or an edge direction) do exist and should be employed. Some of them employ bootstrapping, i.e., learning with resampled data. However, notice that bootstrapping provides the confidence *given by the method for a given feature* (e.g. edge in the network); bootstrapping does not provide an absolute confidence for the feature. For example, if a method systematically reports a false edge because it employs inappropriate tests for the specific data, bootstrapping will also return high confidence on this edge.

Non-linear relations: non-linear relations in continuous data present particular problems to causal discovery. For example, consider the case when two quantities do not interact, unless a third quantity is present in a sufficient concentration. If the data do not contain samples where this third quantity is indeed in high concentration, the causal relation will be undetected. Equivalently, for discrete data, a correlation may be present only for specific values of the variables that never appear in the dataset and hence will not be detected by any algorithm.

3.5 Causal Discovery in Systems Biology: Success Stories

Despite the philosophical, theoretical, and algorithmic problems described above, CD can work when applied with care, and assumptions, technicalities and limitations are duly taken into consideration. The following success stories from *systems biology* provide evidence for this.

3.5.1 Inferring Causal Relationships Among Genotype and Quantitative Traits

In recent years, computational methods have been introduced for *identifying causal relationships among genetic characteristics and quantitative traits in observational data.* These methods were named differently by their respective authors, e.g., Likelihood-based Causality Model Selection (*LCMS*, [91]) or *Trigger* (Transcriptional Regulation Inference from Genetics of Gene ExpRession, [12]). For simplicity, hereafter we will collectively refer to all these methods as Causal Quantitative Trait Loci (CQTL) algorithms.

Specifically, CQTLs methods attempt to reconstruct the causal interaction between a genome marker L and two quantitative traits, namely T_1 and T_2, all measured in the same segregating population. Each quantitative trait can represent the expression value of a given gene, a quantitative phenotype, or any other continuous measurement on the population of interest. CQTL's cornerstone assumption is that *a statistical association between the genetic marker L and the traits of interest must denote a*

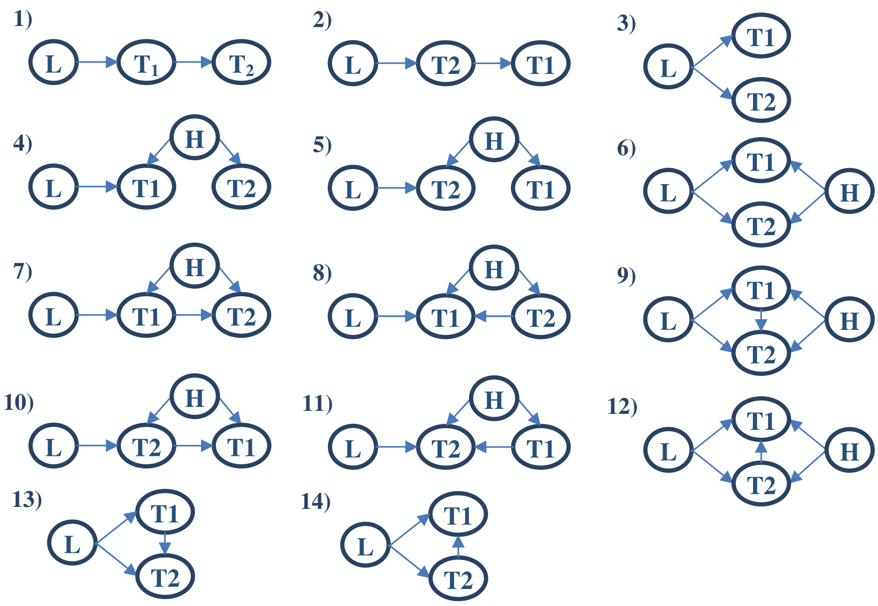


Fig. 3.3 Possible Causal Models among a genetic marker L and two quantitative traits *T1* and *T2* given that the Mendelian Randomization assumptions hold and that all quantities are associated with each other. The node *H* represents one or more hidden confounders

causal effect of L on the associated trait. This assumption is justified by the theory developed in the context of Mendelian Randomization [23, 47]. In a nutshell, Mendelian Randomization methods assume that the random re-composition of the genome during conception can be considered equivalent, from a statistical point of view, to the randomization procedures performed during Randomized Control Trials (RCTs). Consequently, any statistical association between the genetic information and the traits/phenotype of interested cannot be affected by latent confounders, i.e., must denote a causal association.[2] All Quantitative Trait Loci (QTL) studies [66] are based on Mendelian Randomization and its assumptions.

Thus, given that (a) the Mendelian Randomization assumptions hold (i.e., T_1 and T_2 cannot cause L), (b) the Causal Markov and Faithfulness conditions hold as well, and (c) L, T_1 and T_2 are found in the data all statistically associated with each other (i.e., the following dependencies hold: $dep(L, T_1|\emptyset)$, $dep(L, T_2|\emptyset)$, $dep(T_1, T_2|\emptyset)$), then only a very restricted number of causal structures (see Fig. 3.3) are admissible. Each causal model is represented as a CBN, where the node H represents one or more unknown, latent confounders.

Can we further screen out the models presented in Fig. 3.3 and identify the unique, actual causal structure that generated the data at hand? Using the d-separation

[2] Linkage disequilibrium, pleiotropic effects and other factors can invalidate the Mendelian Randomization approach; these issues are better explained later in the text.

criterion above and assuming faithfulness, *if* L *and* T_2 are not statistically associated given T_1 (*i.e.,* $indep(L, T_2|T_1)$ *holds*), *then the true causal model must be* $L \rightarrow T_1 \rightarrow T_2$.

In more detail, each model where L and T_2 are connected through a direct edge is incompatible with $indep(L, T_2|T_1)$, since T_1 cannot d-separate T_2 and L. This leaves models (1), (7) and (8) as the only possible candidates. In the two latter models, T_1 is a collider in the path $L \rightarrow T_1 \leftarrow H \rightarrow T_2$, and thus conditioning on T_1 makes L and T_2 dependent. Thus, the single causal model in agreement with all the assumptions and (in)dependencies encoded in the data is model (1).

Similarly, $indep(L, T_1|T_2)$ holds only when the true underlying causal model is $L \rightarrow T_2 \rightarrow T_1$.

Thus, the causal relationships among a genetic marker and two quantitative traits can be identified, in principle, by assessing whether a limited number of conditional (in)dependencies hold in the data. Particularly, studies focusing on large panels of genomics markers/quantitative traits (e.g., Genome Wide Association Studies) can opportunistically apply CQTL methods on each possible triplet of the form $\{L, T_1, T_2\}$, and potentially discover a large number of causal relationships.

The first theoretically-sound algorithm able to identify, under a well-defined set of assumptions, causal triplets $L \rightarrow T_1 \rightarrow T_2$ where L is known to be "uncaused" was introduced by Cooper in 1997 [18].[3] The first applications of CQTL methods in biology appeared only a decade later: the work presented in Schadt et al. [91] was one of the first studies demonstrating CQTLs effectiveness on a specific biological problem.

Particularly, Schadt and co-authors investigated the causal relationships between a genome-wide panel of markers (L), transcript abundance levels in the liver (T_1) and obesity-related traits (T_2) in mice. They referred to model (1) and (2) in Fig. 3.3 as *Direct Causal model* and *Reactive Causal model*, respectively, while all other models were collectively indicated as the *Independent Causal model.* A model selection procedure, namely LCMS (Likelihood-based Causality Model Selection), was employed for identifying the most plausible causal model for each triplet {genomic marker, transcript abundance level, obesity related trait}. The LCMS procedure belongs to the class of Search-and-Score algorithms, and employs the Akaike Information Criterion (AIC, [1]) as the score metric: $AIC = 2k - 2\ln(\boldsymbol{L})$, where k is the number of parameters of each model and ln ($\boldsymbol{L}$) its log-likelihood.

Chen and co-authors [12] developed a Constraint-based CQTL algorithm. Particularly, they demonstrated the *Causal Equivalence Theorem*, i.e., if the Faithfulness and Causal Markov Condition hold, then:

The causal relationship $L \rightarrow T_1 \rightarrow T_2$ *exists and there are no hidden variables causal for both* T_1 *and* T_2 *if and only if the following three conditions hold:* $dep(L, T_1|\emptyset)$, $dep(L, T_2|\emptyset)$, *and* $indep(L, T_2|T_1)$.[4]

[3] Statistical algorithms for identifying and quantifying mediation effects were known even earlier [58, 97]. However, these algorithms usually assume some particular (linear) distributional model and "fell short of providing a general, causally defensible measure of mediation" [80].

[4] Notably, the "Causal Equivalence Theorem" is identical to the LCD procedure presented in [18].

The authors employ the Causal Equivalence Theorem in order to derive a method, namely Trigger, which provides probability values $\hat{p}_{1,2}$ and $\hat{p}_{2,1}$ for the causal structures $L \rightarrow T_1 \rightarrow T_2$ and $L \rightarrow T_1 \rightarrow T_2$, respectively.

More recently, Millstein and co-authors have proposed another Constraint-based CQTL algorithm, the Causal Inference Test (CIT, [67]), which evaluates a larger set of (conditional) dependencies and independencies than Trigger. Particularly, the following conditions must be satisfied for accepting the Direct Causal Model:

CIT Condition 1: L and T_2 are associated
CIT Condition 2: L and T_1 are associated given T_2
CIT Condition 3: T_1 is associated with T_2 given L
CIT Condition 4: L is independent from T_2 given T_1

A p-value for each of the four CIT conditions can be calculated by applying a suitable statistical test of (conditional) dependency, while the maximum among the four p-values, namely p_{DCM}, is employed as a global statistic for assessing if the four conditions can be jointly accepted. A global p-value p_{RCM} for the Reactive Causal model $L \rightarrow T_2 \rightarrow T_1$ can be derived in a similar way.

Once p_{DCM} and p_{RCM} have been provided, the CIT procedure applies the following rules to distinguish among the possible causal models:

1. If $p_{DCM} < \alpha$ and $p_{RCM} > \alpha$, then the Direct Causal Model is accepted
2. If $p_{DCM} > \alpha$ and $p_{RCM} < \alpha$, then the Reactive Causal Model is accepted
3. If $p_{DCM} > \alpha$ and $p_{RCM} < \alpha$, then the Independent Causal Model is accepted
4. If $p_{DCM} < \alpha$ and $p_{RCM} > \alpha$, then no call is made

where α is a threshold for accepting statistical significance (e.g., $\alpha = 0.05$). Interestingly, CIT does not distinguish among the Independent Causal Model and the case when L is not associated with T_1 or T_2.

CQTL methods have been applied in several studies in order to shade light on specific biological problems. The spread of CQTL methods has also been boosted by the availability of free, open source implementations, whose most notable examples are the R package *cit* (implementing the CIT method), the Network Edge Orienting (NEO) software [4], that implements a score-based CQTLs method, and the R package *qtlnet*, that implements a CQTL algorithm able to take in account and exploit complex correlation structures among multiple traits/phenotypes [72].

A recent example of a successful CQTL study has been presented by Gutierrez-Arcelus et al. [32]: the interaction between DNA sequence, DNA methylation and gene expression was investigated with the CIT method in fibroblasts, T-cells and lymphoblastoid cells extracted from the umbilical cord of 204 babies. This study showed that, when the two alleles of a gene are not equally expressed in a given type of cell, gene expression is mainly regulated by DNA sequence variation, with little or no influence by DNA methylation.

Liu et al. [55] employed the CIT method for disentangling the causal relationships among genome, DNA methylation and Rheumatoid Arthritis. By using the CIT algorithm, the authors found 535 genome—arthritis causal interactions that are mediated by methylation, out of the initially 4016 initially considered associations between genome markers and the rheumatoid arthritis phenotype.

Some controversial CQTL results have been reported in another publication [44]. In this work the authors studied the genome characteristics and expression profile of leukocyte cells from 284 Moroccan individuals. By applying a basic CQTL method, it came out that the SNP rs11987927 seems to trans-regulate the expression of the ZNF71 gene which, in turn, regulates back the transcript abundance of the MYOM2 gene, i.e. the gene where rs11987927 is located. The authors were not able to show whether this counterintuitive result is trustworthy or is instead due to measurement errors [88] or to other causes (e.g. the presence of feedback cycles).

This last example reminds us that the CQTL approach has, obviously, some limitations. Particularly, the limitations affecting Mendelian Randomization [74] affect as well all CQTL studies. Mendelian Randomization assumes that the choice of the mating partner is not affected by the genome. Population stratification is another possible source of bias for Mendelian Randomization and CQTL studies. It can be the case that allelic frequencies and phenotype distributions vary similarly across different populations, even in absence of any causal relations. Consequently, artificial genome-phenotype associations could be detected if the population under study is composed by different sub-populations. Biological redundancy and adaptation to unfavorable genetically-determined phenotypes can hide genome-phenotype causal interactions. Markers that are physically close to each other on the genome tend to be highly associated (a phenomenon known as linkage-disequilibrium) and these associations can lead to the false identification of causal markers that are merely close to the real cause of the phenotype. Highly co-linear (associated, correlated) quantities are close to determinism and violations of Faithfulness (see Sect. 3.4.1 above). Genomic markers can have pleiotropic effects, i.e., simultaneously affecting several traits. If the effect of the pleiotropic marker on each trait is small, it may be necessary to jointly consider all the traits in order to detect the marker-traits causal associations. Furthermore, genomic modifications driven by reverse transcription [9] may ingenerate cases where the observed genomic profiles are actually influenced by the traits under study. Finally, to the best of our knowledge, all CQTL methods developed so far assume that all genome markers follow the same genomic model (usually the additive or co-dominant one), even if assuming the wrong genomic model can lead to a decrease of statistical power [5]. Methodological approaches have been proposed in order to mitigate the effect of some of these limitations, particularly in order to detect causal markers in condition of strong linkage disequilibrium [73] and pleiotropic effects [115].

Despite these limitations, CQTL studies have proven to be able to identify actual casual relationships in a number of different biological context. The main factors enabling CQTL effectiveness are:

Incorporation of prior, biological knowledge: the (apparently) innocuous information that “nothing causes L” is actually pivotal in order to dramatically reduce the number of possible causal models. This means that CQTL methods explore a very small space of possible models thanks to the adoption of Mendelian Randomization assumptions.

Opportunistic approach: CQTL methods are usually applied on a large number of triplets, and whenever a Direct or Reversal causal model cannot be identified, they forgo making a decision. Therefore, the CQTL approach can be thought of as "explorative analysis", useful for discovering novel causal associations which can be subsequently experimentally validated.

"Local" causal discovery: a number of difficulties arise when Causal Discovery methods are applied with the intent to learn a complete network of all direct causal relations, i.e., the CBN among all quantities in the data. Errors in statistical inferences can "propagate", and erroneously orientate edges even in distant regions of the reconstructed network. Conversely, the CQTL approach focuses on a small system formed by solely three quantities, and thus they do not suffer of the issues arising when large networks are induced.

Causal Sufficiency is not assumed: the CQTL approach is "robust" with respect to the presence of latent confounders: no unmeasured variable can affect the association between L and any of the two traits (given the Mendelian Randomization assumption), while if the two traits are both affected by the same latent confounder then the CQTL algorithm will simply forgo making a decision.

Computational feasibility: CQTL algorithms require performing a relatively limited number of statistical (conditional) association tests. Efficient implementations of CQTL algorithm can be easily realized, and CQTL can be applied on hundreds of thousands of triplets in a reasonable time.

Future developments for CQTL methods seem to move in the direction of data integration for network reconstruction. The CIT algorithm was originally proposed as a method for reconstructing causal interaction networks. The QTLnet algorithm [72] tries to reconstruct the interaction network among genome markers and multiple traits. Cai et al. [10] have developed a Structural Equation Model method, namely the Sparsity-aware Maximum Likelihood (SML) algorithm, for reconstructing gene regulatory networks by exploiting genetic perturbations. Finally, in a recent review [90], the author points out that causal triplets provided by CQTL methods can be used for deriving priors for (Causal) Bayesian Network reconstruction algorithms.

3.5.2 Reconstructing Protein Signaling Pathways

Co-ordination of complex cellular activities requires a well-orchestrated propagation of information. In living organisms, this information is transmitted across cells through chemical signals which enter the cell and cause a cascade of chemical, spatial and physical modifications of intracellular compounds. This procedure is broadly described with the term **cell signaling**, and a cascade of responses to a certain extracellular stimulus is generally called a **signaling pathway**, though many argue that presenting a signaling pathway as an isolated set of responses to a specific stimulus may be too simplistic.

Such pathways are typically reconstructed by manually synthesizing pathway components. Each pathway component is discovered through the aggregation of several studies examining the relationship in question under different experimental designs.

Signaling pathways are usually represented as graphs, where the nodes represent participating compounds and the edges represent direct causal links. Different shapes and colors are used to denote different types of participating molecules, and different edges are used to discriminate different types of causal influence.

Bayesian networks, being able to capture both causal and probabilistic relations in multivariate systems, seem fitting to model and quantify signaling pathways. In a ground-breaking paper published in 2005, Sachs et al. [89] applied a Bayesian network learning algorithm to reconstruct a known signaling pathway in T-cells.

The authors used multi-parameter flow cytometry data measuring 11 phosphorylated proteins and phospholipids – all known participants in T-cell signaling – under 9 different experimental conditions in naïve cd4+ T-cells. A score-based algorithm for learning Bayesian networks from a mixture of observational and experimental data [19] was then employed to infer the causal structure and the joint probability distribution of the measured variables.

Each experimental condition included a general or target-specific stimulatory condition, sometimes coupled with a target-specific inhibitor. In total, 5 activators and 5 inhibitors were used. All perturbations were modeled as "ideal" interventions [79] (i.e., hard interventions, not fat hand interventions), where the concentrations of the target molecules are set solely by the manipulation procedure (i.e. the selected inhibitor/activator completely determines the value of the target variable).

The data were discretized into 3 bins, representing "low", "basal" and "high" concentration values, using an algorithm designed to preserve the joint distribution of the variables [33] before being used with the BN learning algorithm. To ensure statistical robustness, the algorithmic process was repeated 500 times with random initial graphs. The output model included only edges present in more than 85 % of the resulting graphs.

The returned network consists of 17 edges and is impressively similar to a consensus signaling pathway manually curated from the literature. Out of the 17 edges identified, 15 edges represent causal links that are well-established in the literature and 2 represent causal links that are not well-established but have been reported at least once. The algorithm failed to discover 3 edges that were expected based on the literature review. However, were they included, these edges would create feedback cycles, which cannot be modeled with Bayesian networks. The causal direction of identified edges was correct, with the exception of a single arc that was found reversed (Fig. 3.4).

To further evaluate the validity of the predicted relations, the authors performed an experiment to test one of the causal links that was found by the algorithm but was not sufficiently backed up by the literature. Specifically, the model included a direct causal link from Erk to Akt, a connection previously reported only in colon cancer cells [31, 114]. The model entails that a perturbation of Erk will influence the abundance of Akt, while it will have no effect in the abundance of PKA.

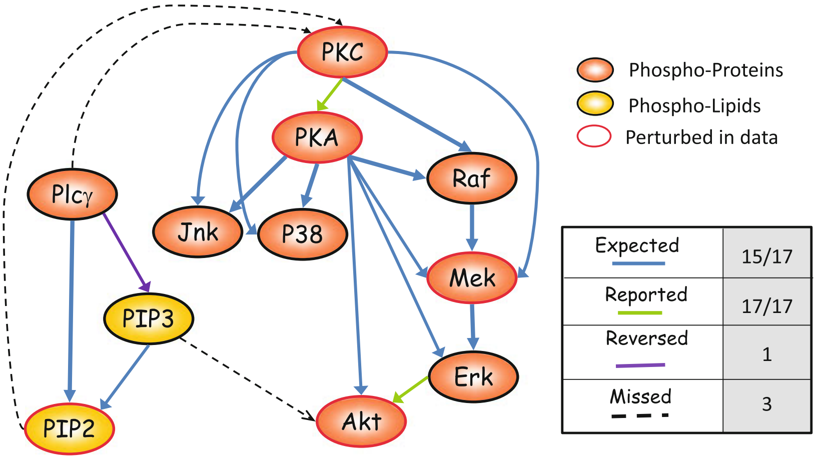


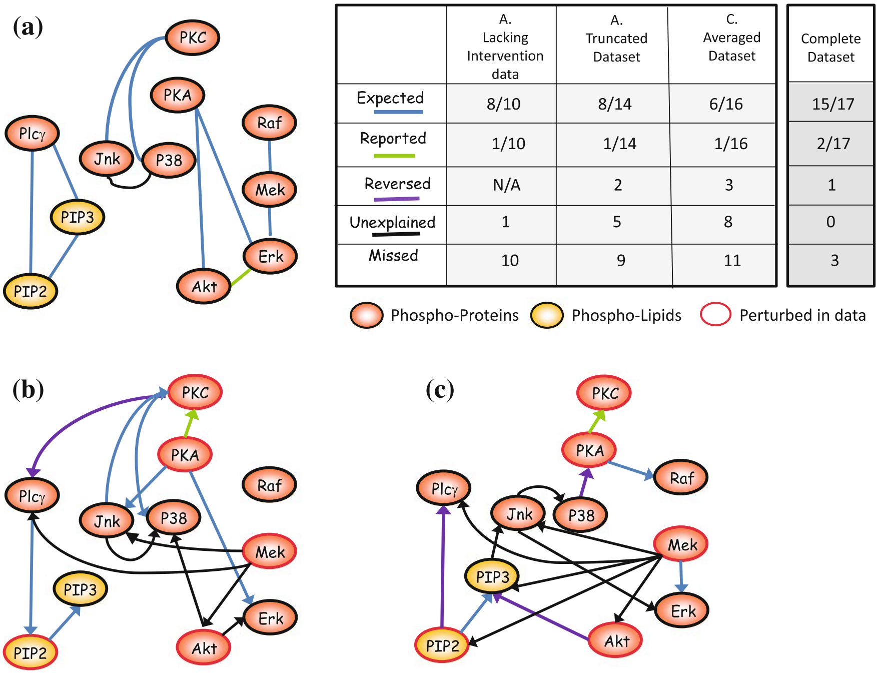
Fig. 3.4 Network inferred from flow cytometry data. The network is a consensus average of 500 high-scoring networks. Only edges present in more than 85 % of the networks are included. Out of 17 edges, 15 are well established in the literature and 2 are reported but not well established. One of the edges is found reversed. The resulting network missed three edges that were expected based on the literature review. Figure from [89]

The authors validated this by inhibiting Erk with a suitable siRNA. True to the model's prediction, Akt activity was reduced ($p < 9.4 \times 10^{-5}$), while PKA activity remained uninfluenced ($p < 0.28$).

Despite the impressively accurate pathway reconstruction and the experimental validation of a previously unknown predicted arc, to the best of our knowledge, this paper remains the only case study of Bayesian network learning for automatic network reconstruction. To understand the reasons automatic causal discovery is still sparsely used in bioinformatics, let us discuss the main factors enabling causal discovery in flow cytometry data:

Network perturbations. An important factor in the success of this method is the inclusion of network perturbations, which are particularly important for correctly identifying the directionality of arcs. To test the significance of including experimental data sets, the authors test the algorithm on a data set consisting of 1200 samples measured without intervention. The resulting network contains only 8 out of the 18 expected edges (compared to 15 when the complete data set is used). In addition, all identified edges are undirected, demonstrating the significance of experiments in identifying causal relations. Nevertheless, we do note that the set of perturbations is still quite limited compared to the full set of experiments required to fully generate the structure without the use of CBN methodology.

Large sample size. Bayesian network learning methods require large sample sizes, while typical experimental designs in molecular biology are usually limited to producing just enough samples to ensure the technical soundness of the procedure. Flow cytometry, measuring the abundance of proteins in single cells, results in hundreds



	A. Lacking Intervention data	A. Truncated Dataset	C. Averaged Dataset	Complete Dataset
Expected	8/10	8/14	6/16	15/17
Reported	1/10	1/14	1/16	2/17
Reversed	N/A	2	3	1
Unexplained	1	5	8	0
Missed	10	9	11	3

● Phospho-Proteins ● Phospho-Lipids ○ Perturbed in data

Fig. 3.5 Networks inferred from: **a** A data set consisting from observations alone. **b** A data set consisting of 420 randomly selected samples from the original data set. **c** A data set consisting of 420 data points, each of which is an average of 20 randomly selected samples from the original data set. In all three cases, the resulting network is far less accurate compared to the one resulting from the complete data set. We can therefore infer that the inclusion of experiments, the large sample size and the lack of averaging effects are crucial for accurate network reconstruction. Figure from [89]

of data points in each experiment, enabling the detection of causal relations in noisy multivariate data. The authors show the importance of large sample sizes by applying the same algorithmic procedure on a truncated version of the original data consisting of 420 randomly selected samples. The resulting network is shown in Fig. 3.5b. It consists of 14 edges, out of which only 8 are expected and only 1 is reported.

Single cell measurements. A key obstacle in applying Bayesian networks in molecular biology data is that the measurements are usually averages of quantities in cell tissues. Using averaged measurements for Bayesian network learning is known to be problematic [15], since the correlation structure of measured quantities may not be preserved. Flow cytometry measurements are single cell measurements, and are therefore suitable for this type of inference. To illustrate this point, the authors simulate a western blot data set over the same variables by selecting at random 20 data points at a time and averaging them, creating a data set of 420 samples in total. The resulting network, shown in Fig. 3.5c, displays a further decline in accuracy: Out of 16 edges, only 6 belong to the expected ones.

Causal sufficiency. In this work, the authors aim to provide a proof-of-concept of the use of Bayesian network in analyzing multivariate flow-cytometry data by reconstructing a well-studied pathway in mammalian T-cells. In doing so, the authors pick 11 compounds in the cell that are *not* confounded (in the context of the set of measured compounds), thus satisfying one of the most difficult assumptions of Bayesian networks, that of causal sufficiency. While the authors do not test how decisive this factor is for the success of the method, it is well known that violation of the causal sufficiency assumption causes errors that propagate through the network. Over the past few years, there has been a growing body of work on causal models for causally insufficient systems, some of which are discussed in Sect. 3.6. However, these models are for the most part developed and disseminated in the machine learning community, and remain fairly unknown in the field of molecular biology.

Overall, several attractive features of the flow cytometry technology render it an ideal test-bed for causal Bayesian network learning. Compared to other high-throughput molecular biology techniques, flow cytometry data have vast sample sizes, do not suffer the unwelcome effects of averaging, and samples can easily be perturbed with in-vitro, close-to-ideal interventions. Unfortunately, flow cytometry technology can only measure up to approximately 20 variables simultaneously, limited by the number of distinguishable fluorescents. This number prevents the measurement of all variables participating in known pathways, let alone the numerous cellular compounds for novel pathways. However, the recently developed technique of **mass cytometry**, where antibodies are tagged with rare isotopes instead of fluorescents, allows measuring up to 30 variables, with a theoretical limit of circa 60 variables [75]. Moreover, the demonstration of the problematic effects of using averaged data along with the development of novel technologies has resulted in growing availability of single-cell genomic data [83, 111], promising a bright future for automatic causal discovery in Bioinformatics.

3.5.3 Estimating Causal Effects in High-Dimensional, Observational Data: The Intervention Calculus when the DAG Is Absent Approach

Identifying cause-effect relationships is one of the main goals of Causal Discovery methods. However, in some cases assessing whether a causal relationship holds is not sufficient, and one also desires to quantify the size of the causal effect. For example, once it has been established that a gene regulates a particular protein, it may also be relevant to know what variation should be expected in the level of the protein's abundance (effect) for a given variation in the level of the expression of the gene (cause).

Estimating the size of a causal effect is not a trivial task, although it becomes feasible when the true causal structure is known. Pearl [78] proposed a technique, named “do-calculus”,[5] which, given a DAG and a suitable parameterization, allows estimating the magnitude of the causal effect between any pair of variables X, Y modeled in the DAG. Unfortunately, in almost all biology-related, real-world problems the actual underlying causal structure is not known, and its reconstruction is often prohibitive, as discussed in Sect. 3.4.

Recently, Maathuis and co-authors [56] proposed a method for estimating a lower bound on the size of the causal effect between two quantities by using a worst-case analysis. Their method, namely *IDA* (Intervention calculus when the DAG is Absent), has at least two appealing features: (a) it is able to estimate causal effects' lower bounds solely on the basis of observational data, i.e., without requiring data from experimental perturbations, and (b) can scale up to high-dimensional settings involving thousands of variables.

The basic idea underlying the IDA algorithm is the following: first, let's assume that the underlying causal mechanism that has generated the data can be represented as a DAG, and that no latent confounders are present (i.e., we assume causal sufficiency). Then, the size of the causal effect $X \rightarrow Y$ between any pair of quantities included in the data can be estimated with the following steps:

1. Identify the CPDAG P that best fits the distribution of the data at hand. Recall from Sect. 3.4 that a CPDAG is a compact representation of the set of DAGs that are Markov equivalent, i.e., the set of DAGs that cannot be distinguished among each other solely on the basis of the available (observational) data. P can be identified by applying any suitable Causal Discovery method, e.g., the PC algorithm [98].
2. Calculate the effect size $ES(X \rightarrow Y)$ for the causal relationship $X \rightarrow Y$ separately for each DAG represented by P. The minimum absolute value among these effect sizes is the lower bound for the effect size of the causal relationship $X \rightarrow Y$.

The apparent simplicity of the IDA algorithm hides an insidious technical issue: the number of DAGs included in P can become intractable even in the case of small systems (e.g., a few tens of measured quantities). For this reason, IDA exploits some sophisticated theoretical results in order to avoid a complete enumeration of the DAGs included in P, while ensuring the correctness of the final results. Moreover, IDA assumes that the data follow a multivariate normal distribution. This assumption is not strictly necessary for the general soundness of the algorithm, but leads to a great simplification of the calculations, since multivariate normality implies linearity of the causal effects. Under the multivariate normal distribution assumption the effect size *ES* of the causal relationship $X \rightarrow Y$ does not depend by the specific value of X and can be expressed as:

$$ES(X \rightarrow Y) = E\left(Y|do(X = x + 1)\right) - E\left(Y|do(X = x)\right)$$

[5] Explaining the details of the do-calculus is beyond the scope of this chapter. Interested readers can refer to Pearl's original publication.

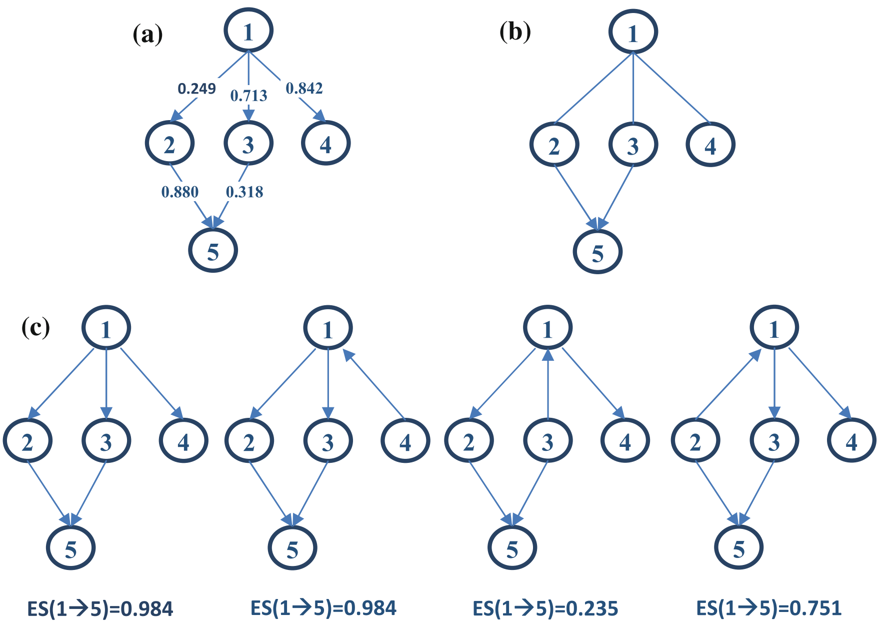


Fig. 3.6 Graphical representation of IDA operation. **a** Example causal network involving five nodes. Causal effects are assumed to be linear, with weights specified on each edge. **b** CPDAG reconstructed by the PC algorithm from 1000 samples simulated from the example causal network. Undirected edges denote arcs that are reversible. **c** DAGs corresponding to the reconstructed CPDAG. For each DAG the effect size of the causal relationship 1 → 5 is reported, as calculated with the do-calculus. The minimum among these values (0.235) is a lower bound of the real effect size. All simulations were performed with the R package *pcalg* [46]

where $E\left(Y|do(X=x)\right)$, in the language of the do-calculus, represents the expected value of the random continuous variable Y if the value of X is forcefully set, through an external intervention, to a fixed value x over the whole population. If all quantities are scaled in order to have zero mean and unitary standard deviation, $ES(X \rightarrow Y)$ would represent the expected variation of Y for a variation of X equal to its standard deviation.

It should also be noted that IDA can be considered a conservative algorithm, performing a “worst case scenario” analysis, since it returns the minimum absolute value among the calculated size effects. Figure 3.6 shows a graphical representation of the operation of IDA.[6]

The main drawback of the IDA algorithm is that it is based on a set of assumptions that are unlikely to hold in real settings, particularly Causal Sufficiency and multivariate normality. Overall, it is not well understood how the results of the algorithm may change whenever one or more of these assumptions is violated.

[6] An implementation of the IDA algorithm is available in the R package *pcalg* [46].

Despite these limitations, IDA has proved to be effective in identifying and quantifying causal relationships from observational data. In a subsequent, ground-breaking publication [57], Maathuis and co-authors applied IDA on two different sets of observational data: a compendium of expression profiles of Saccharomyces Cerevisiae, and the set of simulated gene expression data from the DREAM4 competition *In Silico Network Challenge* [60] For both sets of data, m "true" causal relationships were estimated and quantified through gene knock-out experiments, while q "predicted" causal relationships were obtained by applying the IDA algorithm *on the observational data*. For both the Saccharomyces Cerevisiae and the DREAM4 data the sets of true and predicted causal effects had an overlap statistically significantly larger than the one that can be expected by random guessing. The overlap was statistically significant for different values of q and m. Moreover, when contrasted against two state-of-the-art correlation-based algorithms, (the Lasso and Elastic Net regressions [116]), the IDA algorithm largely outperformed both methods in correctly ranking putative causal relationships; *in fact, the correlation-based algorithms' predictions were only as good as random guessing*. The importance of these results was highlighted in an editorial in the same issue of Nature Methods [11].

An additional application of the IDA algorithm on another real-world problem was also reported [45]. In this work the researchers employed a slightly modified version of IDA (able to deal with binary variables) in order to identify the factors causally influencing the level of general health perception in a sample of spinal cord injury patients. The results of the study confirmed, once more, the capability of IDA in identifying and quantifying causal relationships from observational data.

The factors enabling effective causal discovery with the IDA approach are the following:

Worst case analysis: IDA provides a "worst-case" estimation of the causal effects. This means that only causal relationships strongly supported by the data will be retrieved.

Opportunistic approach: similarly to the CQTL algorithms, IDA is an explorative analysis whose main scope is identifying novel causal relationships, rather than confirming existing ones.

Ranking of putative causal associations: causal associations discovered by IDA are associated with their respective effect size. This means that researchers can rank the putative causal relationships provided by the IDA algorithm according to their estimated effect sizes, and eventually retain/experimentally validate only the top ones.

Finally, it is worth noting that some extensions of IDA were recently published. Le and co-authors presented a version of IDA modified to detect and quantify microRNA/mRNA causal relationships [53]. The Causal Stability Ranking (*CStaR*) method [99] employs the IDA algorithm and a re-sampling based stability selection method [65] to identify, out of a list of possible candidates, the factors that causally influence a given outcome.

3.6 Future Directions

In the previous sections, we presented some introductory concepts related to causality and causal discovery. We also briefly presented (Causal) Bayesian networks, which are one of the main tools for causal discovery without randomized control experiments. Finally, we discussed some of the most prominent and successful applications of causal discovery in the field of molecular biology. Despite years of research in the field of causal discovery and the increasing availability of public data, the applications remain limited and are often contrived examples rather than methods of global applicability. In this section, we explain some of the reasons thereof, and discuss recent developments in causal discovery that may help tackle some of the problems in applied causal discovery, and present some future directions for a unified, robust and integrative approach in causal discovery.

Admitting Latent Confounding Factors: The theory of Bayesian networks relies on the assumption of causal sufficiency, i.e. that no two variables included in the model shares an unobserved common cause (latent confounder). In most real scenarios, this assumption is somewhat arbitrary, since the possibility of a latent confounder can rarely be excluded [87]. The presence of latent confounders is a common source of error in the output of Bayesian network learning algorithms, and an even more common source of criticism and mistrust for causal discovery.

Over the past few years, however, several causal models that do not rely on the assumption of causal discovery have been developed. **Semi Markov causal models** (SMCMs, [103]) are causal models that implicitly model hidden confounders using bi-directed edges. Like Bayesian networks, SMCMs consist of a joint probability distribution over a set of variables and a causal graph over the same set of variables. The graph is an **acyclic directed mixed graph**, where nodes represent variables and edges represent causal relations: A directed edge ($\rightarrow$) denotes a direct causal relation (in the context of variables included in the model), while a bi-directed edge ($\leftrightarrow$) denotes that the variables in question share a latent common cause. Two variables can share both a directed and a bi-directed edge. Under the causal Markov condition and faithfulness, conditional (in)dependencies entailed in the distribution correspond to graph properties of the graph according to the criterion of m-separation, an extension of d-separation in BNs. While obtaining a parameterization of a mixed graph is possible for discrete variables [27, 86] there exists no algorithm that can reverse-engineer a semi-Markov causal model from data.

Maximal ancestral graphs (MAGs, [85]) constitute a different approach in modeling causality in causally insufficient systems. Maximal ancestral graphs are ancestral mixed graphs, meaning they contain no directed or almost directed cycles: An almost directed cycle occurs when $A \leftrightarrow B$ and there exists a directed path from A to B. Every pair of variables A, B in an ancestral graph is joined by at most one edge. The orientation of this edge represents (non) causal ancestry: A bi-directed edge $A \leftrightarrow B$ denotes that A does not cause B and B does not cause A, but (under the faithfulness assumption) the two share a latent confounder. A directed edge $A \rightarrow B$ denotes causal ancestry: A is a causal ancestor of B. Thus, if A causes B (not

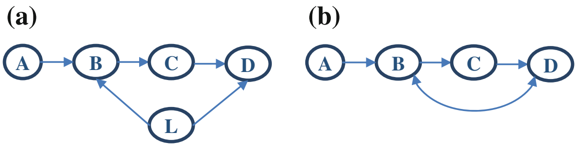


Fig. 3.7 Causal insufficiency. **a** A causal Bayesian network over variables A, B, C, D, L. **b** The semi-Markov causal model over the (causally insufficient) set of variables, A, B, C, D. **c** The maximal ancestral graph over the same variables

necessarily directly in the context of causal variables) and the two are also confounded, there is an edge $A \rightarrow B$ in the corresponding MAG. Figure 3.7 illustrates an example of a marginal SMCM and MAG for the same underlying causal Bayesian network. Some features of a MAG are not identifiable from the joint probability distribution alone. Classes of MAGs that correspond to the same probability distribution form a Markov equivalence class. The FCI algorithm [98, 112] is a constraint-based algorithm that can learn all the invariant features of Markov equivalent MAGs from passive observational data. The algorithm is shown to be sound and complete.

Admitting Feedback Cycles: Another long debated assumption of causal Bayesian networks is acyclicity; i.e., the lack of feedback loops in the system under study. While some may argue that causal processes are acyclic *over time*, in many practical settings we only have cross-sectional, non-temporal data, hopefully having reached equilibrium. Particularly in molecular biology feedback is a well-known regulatory mechanism and thus, acyclicity a problematic assumption.

To address this shortcoming of causal Bayesian networks, several approaches have been introduced, most of which resort to the parametric assumption of linearity. Richardson and Spirtes are the authors of the first general constraint-based algorithm for learning linear cyclic models, the Cyclic Causal Discovery algorithm [84]. The algorithm however is not complete. Schmidt and Murphy present a method for learning discrete cyclic models [92], but their method heavily relies on experimental data. Moreover, the authors present no theoretical results for their algorithms completeness and identifiability status. Itani et al. introduce generalized Bayesian networks [96], an extension of Bayesian networks for cyclic systems with discrete variables, and present a learning algorithm. The method relies on experimental data to both identify data and to apply BN learning algorithms in data where the cycles are broken by perturbations. All of the methods above employ the assumption of causal sufficiency. Hyttinen et al. present a method for learning linear cyclic model from a series of experiments in causally insufficient systems [41], along with sufficient and necessary conditions for identifiability. Unfortunately, this method also relies on linearity, which is generally known not to hold in biological systems.

Local and Opportunistic Learning: Given the limitations, difficulties, and pitfalls of CD, learning complete large networks may degrade quality of learning and present large computational demands. *Local Causal Discovery* takes a different approach. There are at least two types of causal discovery. The first, pioneered by Cooper and colleagues attempts to identify (all) *marginal* graphs (i.e., representing the distribution of a subset of the variables) of some special interest. For example, in [18] all triplets leading to a CBN of the form $L \rightarrow T_1 \rightarrow T_2$ when L is known not to have any causes within the system under study are identified. As discussed in Sect. 3.5.1 this work preceded the CQTL studies and was re-discovered independently later. This is the smallest graph that postulates a new causal relation $T_1 \rightarrow T_2$ *without assuming Causal Sufficiency*, due to the prior knowledge that nothing causes L (we do not consider $L \rightarrow T_1$ or $L \rightarrow T_2$ as new interesting causal postulates since if nothing causes L and L is correlated with T_1 or T_2 then the causal relation should hold trivially). When prior knowledge is not available, the smallest marginal graph that postulates a causal relationship without assuming Causal Sufficiency is called a Y-structure and is of the form $\mathrm{X} \rightarrow Q \leftarrow \mathrm{Z}$, $\mathrm{Q} \rightarrow \mathrm{W}$. If this CBN is induced from the data, then $Q \rightarrow W$ even if Causal Sufficiency is violated (the CBN of course also claims $\mathrm{X} \rightarrow \mathrm{Q}$ but this may not be the case if Causal Sufficiency is violated). Algorithms to identify Y-structures appeared in [59]. Another type of Local Causal Discovery is the reconstruction of focused regions of the underlying causal graph around a variable of interest, e.g., a specific gene, without the need to reconstruct the complete network. The first such method was [62], later receiving more attention in [81, 110]. Such local CD algorithms are closely related to variable selection as the Markov Blanket of a variable is the part of the network relevant for variable selection [3]. The difference between the two types of local causal discovery is that the first learns marginal networks, while the second learns sub-networks. For example, if the true network is $\mathrm{X} \rightarrow \mathrm{Y} \rightarrow Z \rightarrow W$, and nothing causes X, then the method by Cooper [18] will return 3 triplets: $\mathrm{X} \rightarrow \mathrm{Y} \rightarrow W$, $\mathrm{X} \rightarrow Z \rightarrow W$, and $\mathrm{X} \rightarrow \mathrm{Y} \rightarrow W$ corresponding to marginalizing (treating as latent) one variable at a time. The method learning regions in [110] with target Z will return the network $\mathrm{Y} \rightarrow Z \rightarrow W$ (if the region is restricted to be only the nodes adjacent to Z). The latter is a sub-graph of the original graph (in general, local discovery may not orient the same edges as global discovery). *Local Causal Discovery forgoes learning complete networks to save computational time or to make more robust inferences with fewer assumptions*. We also use the term *opportunistic* learning to denote all methods that perform a reliability, confidence, robustness estimation of findings and focus only on the findings for which the method is confident on. The CQTL methods presented above heavily use these ideas.

Integrative Causal Analysis: in recent years, the proliferation of publicly-available, on-line data repositories allow the possibility of co-analyzing large amounts of data and information. This is particularly evident in some fields, for example System Biology, where on-line data repositories are well-established [2, 6] and researchers are encouraged to share their raw data along with their results and findings. Typically, however, data from different studies cannot be pooled together naively and be jointly analyzed, even when all studies examine the same biological system. Any difference in recording conventions, study design or experimental procedures requires sophisticated statistical approaches in order to be addressed. A non-exhaustive list of approaches that attempt to address these issues includes Meta-Analysis [76], Transfer Learning [77], Statistical Matching (also called Data Fusion) [24] and Batch-Effect removal [52]. Each of these approaches is characterized by its own scope, advantages and weaknesses. In general the integrative analysis of heterogeneous datasets is still an open problem and a field of active research.

Integrative analysis from a causal perspective takes a specific form. The key observation in this approach is that all data measuring the same biological system stem from a single causal mechanism. Each study maybe measuring different quantities, under different experimental conditions or sampling methodologies, yet there should exist a causal model that can produce all these datasets. Thus, to co-analyze a collection of datasets coming from heterogeneous studies one searches for a causal model (or all causal models) that simultaneously fit and can explain all data. Over the past few years, several methods for extending causal analysis to the integrative analysis of heterogeneous datasets have been introduced. We collectively refer to these methods as Integrative Causal Analysis (INCA). A major advantage of INCA is that it can model the effect of interventions, e.g. the knock-out of a gene in one dataset and treatment with a hormone in a second one, to enable the co-analysis of datasets over different experimental conditions.

INCA methods can address different types of heterogeneity. Several INCA works have focused on the problem of overlapping variable sets, i.e., co-analyzing data sets that have only a subset of the included variables in common. In this setting the scope of the analysis is usually to infer information regarding the causal mechanism defined over the union of all measured variables. A first pioneering work was published in 2002 by Danks [20], who proposed a two-stage approach consisting in separately learning a Bayesian Network from each study and then using a set of rules for extracting information about the underlying causal structure. Successive methods generally follow a similar two-stage approach, but use more expressive causal models in the first stage (e.g. MAGs) and employ more sophisticated rules that are able deal with conflicts arising from inconsistencies among the models [17, 21, 104, 106].

Studies often differ because they were conducted under different experimental conditions. In this setting, naively pooling data from different studies together can lead to the creation of spurious correlations or to the disappearance of present associations among the measured variables [51]. Several works propose modifications of Search-and-Score and Constraint-based algorithms able to deal with mixtures of observational and experimental data. Cooper and Yoo [19] propose a Bayesian score able to incorporate information about the different experimental settings, while Hauser and Bühlmann [34] investigate the concept of Markov Equivalence in the presence of experimental interventions and propose a learning algorithm on that basis. Eaton and Murphy [25] model interventions as special nodes of the network, and proposed an algorithm that attempts to infer the actual effects of each intervention directly from the data. Constraint–based algorithms for mixtures of experimental data are proposed in [16, 102], but they are limited to specific experimental settings. Sufficient conditions for checking conditional (in)dependencies in data coming from different experiments were proposed in [26, 51]. Other approaches assume specific functional forms for all interactions among variables [39, 41]. These approaches are even able to deal with hidden confounders, but their application is limited by their strict assumptions regarding functional forms among variables. Finally, some algorithms first learn a provisional causal structure from observational data, and then employ experimental information in order to refine the learned model [35, 64].

A particular type of heterogeneity is obtained when the same information is recorded with different encodings, for example smoking information may be recorded as a binary (yes/no) or a continuous variable (number of packets a day). When a direct conversion is not possible, more sophisticated approaches must be employed [107].

Recent developments in Integrative Causal Analysis focus on co-addressing multiple sources of heterogeneity at the same time: several works attempt to integrate data measured over overlapping variable sets *and* in different experimental conditions [40, 42, 43, 105].

One of the main, unresolved issues in the INCA field is the efficacy of the current methods on real data. While significant efforts have been spent on laying down the theoretical foundations of this field, several algorithmic and methodological improvements are necessary before applying these methods on real data analysis tasks. A first attempt in applying INCA methods on real-world, large datasets has produced evidence that INCA methods can actually provide meaningful results and even outperform current statistical methods [109]. Bridging the gap between theory and practice is crucial for the future of integrative causal analysis.

CD Based on Functional-Form Analysis: So far we have mainly discussed causal discovery methods based on the analysis of conditional (in)dependencies. These methods query the joint probability distribution for (in)dependencies either directly (constraint-based methods) or indirectly (search-and-score methods) to identify all causal structure that fit the data. Recently, a different approach on causal discovery has been developed, one that is based on the exploiting possible asymmetries of causal relations. The methods assume Causal Sufficiency and acyclicity, thus if X and Y are correlated, either X → Y or $Y \rightarrow X$. Expressed as structural equations, either $X = f(Y, \varepsilon)$ or $Y = f(X, \varepsilon)$, where the disturbance term ε is the effect of all other factors. It turns out that one can distinguish between the two possibilities if either ε is non-Gaussian, or f is non-linear [36]. While the assumptions of linear relations and Gaussian residual term ε is probably the most common set of assumptions in statistics, it turns out that any departure from these assumptions allows the discovery of the directionality of causation!

A specific case follows. Assume that X and Y are variables and X causes Y in a linear manner, thus $Y = \beta_{YX}X + \epsilon_Y$, where ϵ_Y follows a non-Gaussian distribution. Also assume that we have obtained a set of measurements of both X and Y and we want to identify the causal structure of the variables. By assuming linearity, additive disturbance terms ε, and causal sufficiency, we can fit both models using simple linear regression, and obtain estimates for both $\widehat{\beta_{XY}}$ and $\widehat{\beta_{YX}}$. Based on these estimates, we can then calculate the disturbances ϵ_X and ϵ_Y for both models. These disturbances will be independent with each other if the fitted model is the correct one, and dependent if the fitted model is the reverse one. A graphical depiction of this principle for uniform distributions of disturbances is shown in Fig. 3.8. LiNGAM [95] automates this procedure, inferring a unique causal model from observational data. LiNGAM is limited to linear relations, but this assumption has been relaxed in

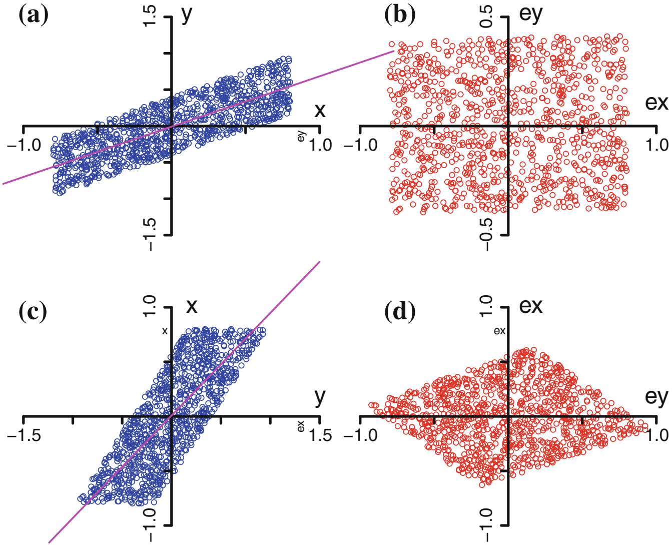


Fig. 3.8 The key idea for LiNGAM and similar algorithms: true structural equation $Y = \beta_{YX} X + \epsilon_Y$, where ϵ_Y follows a uniform distribution. **a** Regression with Y as the dependent variable (true model). **b** Regression with X as the dependent variable (reverse model). **c** Estimated $\widehat{\epsilon_Y}$ versus $\widehat{\epsilon_X}$ based on the model shown in (**a**). The disturbances are and independent. **d** Estimated $\widehat{\epsilon_Y}$ versus $\widehat{\epsilon_X}$ based on the model shown in (**b**). The disturbances are dependent. Figure from [38]

a subsequent body of work [36, 82, 113] to include non-linear relations. However learning such relations requires non-linear optimization techniques and appropriate independence measures [69].

This class of methods is more powerful than traditional causal discovery methods, in the sense that with the functional form assumptions (e.g., linear relations, additive disturbances, non-Gaussian disturbances) causal models are fully identifiable (no statistical indistinguishability). Moreover, the methods also work under unfaithfulness. On the other hand, all methods in this category require large sample sizes and rely on some kind of parametric assumption, and have been shown to be unreliable when this assumption is violated. Nevertheless, these ideas add a new direction and dimension to the way we think about, treat, model, and induce causality and could soon lead to practical results.

3.7 Discussion

Inducing causal models or relations from data is necessary to fully understand biological mechanisms and design new drugs and therapies. Traditional means for such inferences rely on performing interventional experiments. Causal Discovery methods attempt to make such inferences from observational data alone or with a limited set of such interventions by making assumptions that connect the notion of causality with quantities estimable from the data. The analyst should be aware and conscious of the explicit and implicit assumptions employed by the tools and algorithms that are used and whether they are appropriate for the type of biological data at hand. Despite the inherent theoretical and practical difficulties of the task, there are several successful applications of Causal Discovery methods in systems biology that demonstrate the potential of the field. In addition, recent theoretical and algorithmic breakthroughs promise to further improve the successful application of causal discovery on systems biology.

Acknowledgments This work was partially funded by STATegra EU FP7 project, No 306000 and EPILOGEAS GSRT ARISTEIA II project, No 3446.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

1. Akaike, H.: A new look at the statistical model identification. IEEE Trans. Automat. Contr. **19**(6), 716–723 (1974)
2. Akbani, R., Ng, P.K.S., Werner, H.M.J., Shahmoradgoli, M., Zhang, F., Ju, Z., Liu, W., Yang, J.-Y., Yoshihara, K., Li, J., Ling, S., Seviour, E.G., Ram, P.T., Minna, J.D., Diao, L., Tong, P., Heymach, J.V., Hill, S.M., Dondelinger, F., Städler, N., Byers, L., Meric-Bernstam, F., Weinstein, J.N., Broom, B.M., Verhaak, R.G.W., Liang, H., Mukherjee, S., Lu, Y., Mills, G.B.: A pan-cancer proteomic perspective on The Cancer Genome Atlas. Nat. Commun. **5**, 3887 (2014)
3. Aliferis, C.F.: Local causal and Markov blanket induction for causal discovery and feature selection for classification Part I?: algorithms and empirical evaluation. J. Mach. Learn. Res. **11**, 171–234 (2010)
4. Aten, J.E., Fuller, T.F., Lusis, A.J., Horvath, S.: Using genetic markers to orient the edges in quantitative trait networks: the NEO software. BMC Syst. Biol. **2**, 34 (2008)
5. Balding, D.J.: A tutorial on statistical methods for population association studies. Nat. Rev. Genet. **7**, 781–791 (2006)
6. Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C.L., Serova, N., Davis, S., Soboleva, A.: NCBI GEO: Archive for functional genomics data sets—Update. Nucleic Acids Res. **41** (2013)
7. Borboudakis, G., Tsamardinos, I.: Bayesian network learning with discrete case-control data. In: Uncertainty in Artificial Intelligence (UAI), 2015
8. Borboudakis, G., Tsamardinos, I.: Incorporating causal prior knowledge as path-constraints in Bayesian networks and maximal ancestral graphs. In: Proceedings of the 29th International Conference on Machine Learning (ICML-12), 2012, pp. 1799–1806
9. Burns, M.B., Temiz, N., Harris, R.S.: Evidence for APOBEC3B mutagenesis in multiple human cancers. Nat. Genet. **45**(9), 977–983 (2013)

10. Cai, X., Bazerque, J.A., Giannakis, G.B.: Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations. Plos Comput. Biol. **9**(5), e1003068 (2013)
11. Cause and effect. Nat. Methods **7**, 243 (2010)
12. Chen, L.S., Emmert-Streib, F., Storey, J.D.: Harnessing naturally randomized transcription to infer regulatory relationships among genes. Genome Biol. **8**(10), R219 (2007)
13. Chickering, D.M.: Learning equivalence classes of Bayesian-network structures. J. Mach. Learn. Res. **2**, 445–498 (2002)
14. Chickering, D.M., Heckerman, D., Meek, C.: Large-sample learning of Bayesian networks is NP-hard. J. Mach. Learn. Res. **5**, 1287–1330 (2004)
15. Chu, T., Glymour, C., Scheines, R., Spirtes, P.: A statistical problem for inference to regulatory structure from associations of gene expression measurements with microarrays. Bioinformatics **19**(9), 1147–1152 (2003)
16. Claassen, T., Heskes, T.: Causal discovery in multiple models from different experiments. In: Advances in Neural Information Processing Systems (NIPS 2010), 2010, pp. 1–9
17. Claassen, T., Heskes,T.: Learning causal network structure from multiple (in) dependence models. In: Proceedings of the Fifth European Workshop on Probabilistic Graphical Models (PGM), pp. 81–88
18. Cooper, G.F.: A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. Data Min. Knowl. Discov. **1**, 203–224 (1997)
19. Cooper, G., Yoo, C.: Causal discovery from a mixture of experimental and observational data. In: Proceedings of the Fifthteenth Conference on Uncertainty in Artificial Intelligence (UAI'99), 1999, pp. 116–125
20. Danks, D.: Learning the causal structure of overlapping variable sets. In: Discovery Science: Proceedings of the 5th International Conference, 2002, pp. 178–191
21. Danks, D., Glymour, C., Tillman, R.E.: Integrating locally learned causal structures with overlapping variables. In: Advances in Neural Information Processing Systems, pp. 1665–1672. MIT Press, Cambridge (2009)
22. Dash, D., Druzdel, M.: Caveats for causal reasoning with equilibrium models. ECSQUARU **2143**, 192–203 (2001)
23. Davey Smith, G., Ebrahim, S.: Mendelian randomization: can genetic epidemiology contribute to understanding environmental determinants of disease?. Int. J. Epidemiol. **32**(1), 1–22 (2003)
24. D'Orazio, M., Di Zio, M., Scanu, M.: Statistical Matching: Theory and Practice, p. 268. Wiley, New York (2006)
25. Eaton, D., Murphy, K.: Exact Bayesian structure learning from uncertain interventions. In: AISTATS (2007)
26. Eberhardt, F.: Sufficient condition for pooling data from different distributions. Error (2006)
27. Evans, R.J., Richardson, T.S.: Marginal log-linear parameters for graphical Markov models. J. R. Stat. Soc. Ser. B. Stat. Methodol. **75**(4), 743–768 (2013)
28. Fisher, R.A.: The distribution of the partial correlation coefficient. Metron **3**(3–4), 329–332 (1923)
29. Fisher, R.A.: The Design of Experiments. Hafner Publishing, New York (1935)
30. Friedman, N., Murphy, K., Russell, S.: Learning the structure of dynamic probabilistic networks. In: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, 1998, pp. 139–147
31. Fukuda, R., Kelly, B., Semenza, G.L.: Vascular endothelial growth factor gene expression in colon cancer cells exposed to prostaglandin E2 is mediated by hypoxia-inducible factor 1. Cancer Res. **63**(9), 2330–2334 (2003)
32. Gutierrez-Arcelus, M., Lappalainen, T., Montgomery, S.B., Buil, A., Ongen, H., Yurovsky, A., Bryois, J., Giger, T., Romano, L., Planchon, A., Falconnet, E., Bielser, D., Gagnebin, M., Padioleau, I., Borel, C., Letourneau, A., Makrythanasis, P., Guipponi, M., Gehrig, C., Antonarakis, S.E., Dermitzakis, E.T.: Passive and active DNA methylation and the interplay with genetic variation in gene regulation. Elife **2**, e00523 (2013)

33. Hartemink, A.J., Gifford, D.K., Jaakkola, T.S., Young, R.A.: Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. In: The Pacific Symposium on Biocomputing, 2001, pp. 422–433
34. Hauser, A., Bühlmann, P.: Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. J. Mach. Learn. Res. **13**, 2409–2464 (2012)
35. He, Y.-B.: Active learning of causal networks with intervention experiments and optimal designs. J. Mach. Learn. Res. **9**, 2523–2547 (2008)
36. Hoyer, P.O., Janzing, D., Joris, M., Peters, J., Schölkopf, B.: Nonlinear causal discovery with additive noise models. In: NIPS, 2008
37. Hug, S., Schmidl, D., Li, W.B., Greiter, M.B., Theis, F.J.: Bayesian model selection methods and their application to biological ODE systems. In: Uncertainty in Biology, A Computational Modeling Approach. Springer, Cham (2016, this volume)
38. Hyttinen, A.: Discovering Causal Relations in the Presence of Latent Confounders. University of Helsinki, Helsinki (2013)
39. Hyttinen, A., Eberhardt, F., Hoyer, P.O.: Noisy-OR models with latent confounding. In: Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence, Barcelona, 2011
40. Hyttinen, A., Eberhardt, F., Hoyer, P.O.: Causal discovery of linear cyclic models from multiple experimental data sets with overlapping variables. In: Proceedings of the Uncertainty in Artificial Intelligence, 2012
41. Hyttinen, A., Eberhardt, F., Hoyer, P.O.: Learning linear cyclic causal models with latent variables. J. Mach. Learn. Res. **2013**(3387–3439), 3387–3439 (2012). Jan
42. Hyttinen, A., Eberhardt, F., Jarvisalo, M.: Constraint-Based Causal Discovery: Conflict Resolution with Answer Set Programming. In: Proceedings of the Uncertainty in Artificial Intelligence, 2014
43. Hyttinen, A., Hoyer, P.O., Eberhardt, F., Järvisalo, M.: Discovering cyclic causal models with latent variables: a general sat-based procedure. In: Proceedings of the Uncertainty in Artificial Intelligence, 2013
44. Idaghdour, Y., Czika, W., Shianna, K.V., Lee, S.H., Visscher, P.M., Martin, H.C., Miclaus, K., Jadallah, S.J., Goldstein, D.B., Wolfinger, R.D., Gibson, G.: Geographical genomics of human leukocyte gene expression variation in southern Morocco. Nat. Genet. **42**(1), 62–67 (2010)
45. Kalisch, M., Fellinghauer, B.A.G., Grill, E., Maathuis, M.H., Mansmann, U., Buhlmann, P., Stucki, G.: Understanding human functioning using graphical models. BMC Med. Res. Methodol. **10**, 14 (2010)
46. Kalisch, M., Maechler, M., Colombo, D., Maathuis, M.H., Buehlmann, P.: Causal inference using graphical models with the R package pcalg. J. Stat. Softw. **47**(11), 1–26 (2012)
47. Katan, M.B.: Apolipoprotein E isoforms, serum cholesterol, and cancer. Lancet **1**(8479), 507–508 (1986)
48. Kenfield, S.A., Stampfer, M.J., Chan, J.M., Giovannucci, E.: Smoking and prostate cancer survival and recurrence. JAMA **305**(24), 2548–2555 (2011)
49. Kirk, P., Silk, D., Stumpf, M.P.H.: Reverse engineering under uncertainty. In: Uncertainty in Biology, A Computational Modeling Approach. Springer, Cham (2016, this volume)
50. Labrie, F., Dupont, A., Suburu, R., Cusan, L., Tremblay, M., Gomez, J.L., Emond, J.: Serum prostate specific antigen as pre-screening test for prostate cancer. J. Urol. **147**(3 Pt 2), 846–851 (discussion 851–852) (1992)
51. Lagani, V., Tsamardinos, I., Triantafillou, S.: Learning from mixture of experimental data: a constraint—based approach. In: SETN'12 Proceedings of the 7th Hellenic Conference on Artificial Intelligence: Theories and Applications, 2012, vol. 7297, pp. 124–131
52. Lazar, C., Meganck, S., Taminau, J., Steenhoff, D., Coletta, A., Molter, C., Weiss-Solís, D.Y., Duque, R., Bersini, H., Nowé, A.: Batch effect removal methods for microarray gene expression data integration: a survey. Brief. Bioinform. **14**(4), 469–490 (2013)
53. Le, T.D., Liu, L., Tsykin, A., Goodall, G.J., Liu, B., Sun, B.-Y., Li, J.: Inferring microRNA-mRNA causal regulatory relationships from expression data. Bioinformatics **29**(6), 765–771 (2013)

54. Lemeire, J., Janzing, D.: Replacing causal faithfulness with algorithmic independence of conditionals. Minds Mach. (2012)
55. Liu, Y., Aryee, M.J., Padyukov, L., Fallin, M.D., Hesselberg, E., Runarsson, A., Reinius, L., Acevedo, N., Taub, M., Ronninger, M., Shchetynsky, K., Scheynius, A., Kere, J., Alfredsson, L., Klareskog, L., Ekström, T.J., Feinberg, A.P.: Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nat. Biotechnol. **31**(2), 142–7 (2013)
56. Maathuis, M.H., Kalisch, M., Bühlmann, P.: Estimating high-dimensional intervention effects from observational data. Ann. Stat. **37**(6A), 3133–3164 (2009)
57. Maathuis, M.H., Colombo, D., Kalisch, M., Bühlmann, P.: Predicting causal effects in large-scale systems from observational data. Nat. Methods **7**(4), 247–248 (2010)
58. MacKinnon, D.P.: Introduction to Statistical Mediation Analysis (Multivariate Applications Series), p. 488. Routledge, New York (2008)
59. Mani, S., Cooper, G.F.: Causal discovery using a Bayesian local causal discovery algorithm. Stud. Health Technol. Inform. **107**(Pt 1), 731–735 (2004)
60. Marbach, D., Schaffter, T., Mattiussi, C., Dario, F.: Generating realistic in silico gene networks for performance assessment of reverse engineering methods. J. Comput. Biol. **16**(2), 229–239 (2009)
61. Margaritis, D.: Distribution-free learning of Bayesian network structure in continuous domains. In: AAAI'05 Proceedings of the 20th National Conference on Artificial Intelligence—Volume 2, 2005, pp. 825–830
62. Margaritis, D., Thrun, S.: Bayesian network induction via local neighborhoods. Adv. Neural Inf. Process. Syst. **12**, 505–511 (2000)
63. McDonald, J.H.: Handbook of Biological Statistics, p. 291. Sparky House Publishing, Baltimore (2009)
64. Meganck, S., Maes, S., Leray, P., Manderick, B.: Learning semi-markovian models using experiments. In: Third European Workshop on Probabilistic Graphical Models (PGM), 2006
65. Meinshausen, N., Buhlmann, P.: Stability selection. J. R. Stat. Soc. Ser. B **72**(4), 417–473 (2010)
66. Miles, C., Wayne, M.: Quantitative trait locus (QTL) analysis. Nat. Educ. **1**(1) (2008)
67. Millstein, J., Zhang, B., Zhu, J., Schadt, E.E.: Disentangling molecular relationships with a causal inference test. BMC Genet. **10**, 23 (2009)
68. Monti, S., Cooper, G.F.: A multivariate discretization method for learning Bayesian networks from mixed data. In: UAI'98 Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, 1998, pp. 404–413
69. Mooij, J., Janzing, D., Peters, J., Schölkopf, B.: Regression by dependence minimization and its application to causal inference in additive noise models. In: Proceedings of the 26th Annual International Conference on Machine Learning—ICML '09, 2009, pp. 745–752
70. Näger, P.M.: Causal graphs for EPR experiments. In: Foundations of Physics 2013: The 17th UK and European Meeting on the Foundations of Physics, 2013
71. Neapolitan, R.E.: Learning Bayesian Networks. Pearson Prentice Hall, New York (2004)
72. Neto, E.C., Keller, M.P., Attie, A.D., Yandell, B.S.: Causal graphical models in systems genetics: a unified framework for joint inference of causal network and genetic architecture for correlated phenotypes. Ann. Appl. Stat. **4**(1), 320–339 (2010). Mar
73. Nica, A.C., Montgomery, S.B., Dimas, A.S., Stranger, B.E., Beazley, C., Barroso, I., Dermitzakis, E.T.: Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. PLoS Genet. **6**(4), e1000895 (2010)
74. Nitsch, D., Molokhia, M., Smeeth, L., DeStavola, B.L., Whittaker, J.C., Leon, D.A.: Limits to causal inference based on Mendelian randomization: a comparison with randomized controlled trials. Am. J. Epidemiol. **163**(5), 397–403 (2006)
75. Ornatsky, O., Bandura, D., Baranov, V., Nitz, M., Winnik, M.A., Tanner, S.: Highly multiparametric analysis by mass cytometry. J. Immunol. Methods **361**(1–2), 1–20 (2010)
76. O'Rourke, K.: An historical perspective on meta-analysis: dealing quantitatively with varying study results. J. R. Soc. Med. **100**(12), 579–82 (2007)

77. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**, 1345–1359 (2010)
78. Pearl, J.: Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann, San Francisco (1988)
79. Pearl, J.: Causality: Models, Reasoning and Inference. Cambridge University Press, Cambridge (2009)
80. Pearl, J.: Interpretation and Identification of Causal Mediation. University of California, Los Angeles (2013)
81. Peña, J.M., Björkegren, J., Tegnér, J.: Growing Bayesian network models of gene networks from seed genes. Bioinformatics **21**(Suppl 2), ii224–i229 (2005)
82. Peters, J., Mooij, J., Janzing, D., Schoelkopf, B.: Identifiability of Causal Graphs Using Functional Models. arXiv.org (2012)
83. Petretto, E.: Single cell expression quantitative trait loci and complex traits. Genome Med. **5**(8), 72 (2013)
84. Richardson, T., Spirtes, P.: Automated causal discovery in linear feedback models. In: Glymour, C., Cooper, G. (eds.) Computation, Causation and Discovery, pp. 253–302. AAAI press, Cambridge (1999)
85. Richardson, T., Spirtes, P.: Ancestral graph Markov models. Ann. Stat. **30**(4), 962–1030 (2002)
86. Richardson, T., Evans, R., Robins, J.: Transparent parametrizations of models for potential outcomes. Bayesian Stat. **9**, 569–610 (2011)
87. Robins, J.M., Wasserman, L.: On the impossibility of inferring causation from association without background knowledge. In: Glymour, C., Cooper, G.F. (eds.) Computation, Causation, and Discovery, pp. 305–321. AAAI Press/The MIT Press, Menlo Park, CA, Cambridge, MA (1999)
88. Rockman, M.V.: Reverse engineering the genotype-phenotype map with natural genetic variation. Nature **456**, 738–744 (2008)
89. Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D.A., Nolan, G.P.: Causal protein-signaling networks derived from multiparameter single-cell data. Science **308**(5721), 523–529 (2005)
90. Schadt, E.E.: Causal inference and the construction of predictive network models in biology. In: Handbook of Systems Biology Concept and Insights, pp. 499–514. Elsevier Inc. (2013)
91. Schadt, E.E., Lamb, J., Yang, X., Zhu, J., Edwards, S., Guhathakurta, D., Sieberts, S.K., Monks, S., Reitman, M., Zhang, C., Lum, P.Y., Leonardson, A., Thieringer, R., Metzger, J.M., Yang, L., Castle, J., Zhu, H., Kash, S.F., Drake, T.A., Sachs, A., Lusis, A.J.: An integrative genomics approach to infer causal associations between gene expression and disease. Nat. Genet. **37**(7), 710–717 (2005)
92. Schmidt, M., Murphy, K.: Modeling discrete interventional data using directed cyclic graphical models. In: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI '09), 2009, pp. 487–495
93. Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**(2), 461–464 (1978)
94. Shang, Z., Zhu, S., Zhang, H., Li, L., Niu, Y.: Germline homeobox B13 (HOXB13) G84E mutation and prostate cancer risk in European descendants: a meta-analysis of 24,213 cases and 73,631 controls. Eur. Urol. **64**(1), 173–176 (2013)
95. Shimizu, S., Hoyer, P.O., Hyvärinen, A., Kerminen, A.: A linear non-Gaussian acyclic model for causal discovery. J. Mach. Learn. Res. **7**(2), 2003–2030 (2006)
96. Sleiman Itani, B.S., Ohannessian, M., Sachs, K., Nolan, G.P., Dahleh, M.A., Guyon, I., Janzing, D.: Structure learning in causal cyclic networks. In: NIPS, 2008, pp. 165–176
97. Sobel, M.E.: Asymptotic confidence intervals for indirect effects in structural equation models. Sociol. Methodol. **13**(1982), 290–312 (1982)
98. Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction, and Search, vol. 81. Springer, New York (1993)
99. Stekhoven, D.J., Moraes, I., Sveinbjornsson, G., Hennig, L., Maathuis, M.H., Buhlmann, P.: Causal stability ranking. Bioinformatics **28**(21), 2819–2823 (2012)

100. Stelling, J., Klamt, S., Bettenbrock, K., Schuster, S., Gilles, E.D.: Metabolic network structure determines key aspects of functionality and regulation. Nature **420**(6912), 190–193 (2002)
101. Sunnåker, M., Stelling, J.: Model extension and model selection. In: Uncertainty in Biology, A Computational Modeling Approach. Springer, Cham (2016, this volume)
102. Tian, J., Pearl, J.: Causal discovery from changes. In: UAI'01 Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence, 2001, pp. 512–521
103. Tian, J., Pearl, J.: On the identication of causal effects. Technical Report R-290-L, 2003
104. Tillman, R.E., Spirtes, P.: Learning equivalence classes of acyclic models with latent and selection variables from multiple datasets with overlapping variables. J. Mach. Learn. Res. Proc. Track **15**, 3–15 (2011)
105. Triantafillou, S., Tsamardinos, I.: Constraint-Based Causal Discovery from Multiple Interventions over Overlapping Variable Sets. JMLR, to appear
106. Triantafillou, S., Tsamardinos, I., Tollis, I.G.: Learning causal structure from overlapping variable sets. In: Proceedings of Artificial Intelligence and Statistics, 2010
107. Tsamardinos, I., Borboudakis, G.: Permutation testing improves Bayesian network learning. In: ECML PKDD'10 Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases: Part III, 2010, pp. 322–337
108. Tsamardinos, I., Brown, L.E., Aliferis, C.F.: The max-min hill-climbing Bayesian network structure learning algorithm. Mach. Learn. **65**(1), 31–78 (2006)
109. Tsamardinos, I., Triantafillou, S., Lagani, V.: Towards integrative causal analysis of heterogeneous data sets and studies. J. Mach. Learn. Res. **13**(1), 1097–1157 (2012)
110. Tsamardinos, I., Aliferis, C.F., Statnikov, A., Brown, L.E.: Scaling-Up Bayesian Network Learning to Thousands of Variables Using Local Learning Techniques
111. Wills, Q.F., Livak, K.J., Tipping, A.J., Enver, T., Goldson, A.J., Sexton, D.W., Holmes, C.: Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. Nat. Biotechnol. **31**(8), 748–752 (2013)
112. Zhang, J.: On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. Artif. Intell. **172**(16–17), 1873–1896 (2008)
113. Zhang, K., Hyvärinen, A.: On the identifiability of the post-nonlinear causal model. In: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, 2009, pp. 647–655
114. Zhang, W.M., Wong, T.M.: Suppression of cAMP by phosphoinositol/Ca2+ pathway in the cardiac kappa-opioid receptor. Am. J. Physiol. **274**(1 Pt 1), C82–C87 (1998)
115. Zhang, W., Zhu, J., Schadt, E.E., Liu, J.S.: A Bayesian partition method for detecting pleiotropic and epistatic eQTL modules. PLoS Comput. Biol. **6**(1), e1000642 (2010)
116. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B **67**(2), 301–320 (2005)