Knowledge-Based Systems 76 (2015) 79–95

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

SemCaDo: A serendipitous strategy for causal discovery and ontology evolution

Montassar Ben Messaoud [a,b,*], Philippe Leray [a], Nahla Ben Amor [b]

[a] Laboratoire d'Informatique de Nantes Atlantique (LINA) UMR 6241, Ecole Polytechnique de l'Université de Nantes, France
[b] LARODEC, Institut Supérieur de Gestion Tunis, 41, Avenue de la liberté, 2000 Le Bardo, Tunisia

ARTICLE INFO

Article history:
Received 25 May 2013
Received in revised form 4 December 2014
Accepted 5 December 2014
Available online 17 December 2014

Keywords:
Causal Bayesian Networks
Domain ontologies
Serendipity
Decisional guidance
Causal discovery
Experimental data
Ontology evolution

ABSTRACT

Within the last years, probabilistic causality has become a very active research topic in artificial intelligence and statistics communities. Due to its high impact in various applications involving reasoning tasks, machine learning researchers have proposed a number of techniques to learn Causal Bayesian Networks. Within the existing works in this direction, few studies have explicitly considered the role that decisional guidance might play to alternate between observational and experimental data processing. In this paper, we go further by introducing a serendipitous strategy to elucidate semantic background knowledge provided by the domain ontology to learn the causal structure of Bayesian Networks. We also complement our contribution with an enrichment process by which it will be possible to reuse these causal discoveries, support the evolving character of the semantic background and make an ontology evolution. Finally, the proposed method will be validated through simulations and real data analysis.

1. Introduction

Recently, the Machine Learning community has become increasingly aware of the need for developing methods that unify statistical and relational aspects of a wide variety of challenging machine learning problems. In this context, Probabilistic Relational Models, a range of Statistical Relational Learning formalisms, seem to be well placed to reason about uncertainty and provide relational, structured representations. Because of their elegant way of dealing with variables as well as the relationships that hold amongst them, Probabilistic Relational Models have been successfully applied in a wide variety of domains such as social network analysis, biological systems, pattern recognition and other domains that involve relational data.

Probabilistic Graphical Models [33] are a class of Probabilistic Relational Models that can represent rich dependency structures and capture the causal process by which the data was generated. Well-known examples of Probabilistic Graphical Models include Bayesian Networks [48] which are compact probabilistic graphs able to model domains with intrinsic uncertainty.

One of the important properties relative to the Bayesian Networks is the *Markov equivalence property*, which can be illustrated by the fact that the two networks $X \rightarrow Y$ and $X \leftarrow Y$ are equivalent (i.e., encode the same joint probability distribution). Nevertheless, only one of them is correct from causal point of view. In fact, in the first network, X causes Y, thus, manipulating the value of X affects the value of Y contrary to the second one where Y is a cause of X meaning that manipulating X will not affect Y.

One way to cope with this limitation is to use Causal Bayesian Networks, which are probabilistic models able to capture stochastic causal processes in a rigorous and compact way [50]. The main feature of Causal Bayesian Networks is their ability to perform both probabilistic and causal inference. This means that given a Causal Bayesian Network, one can use it either to determine how the observation of specific values (evidence) affects the probabilities of query variable(s) or to predict the effect of an intervention on the remaining variables.

Contrary to the non-Gaussian models (also called LiNGAM [55]), whose causal graph can be fully identified from observational data, usual Causal Bayesian Network formalisms need interventional data in order to recover the fully causal structure. In this work, we do not make use of LiNGAM methods since no suitable parametrization of the joint distribution can be established when working under the non-gaussianity assumption. This is the key reason for restricting our focus to Causal Bayesian Networks only.

* Corresponding author at: Laboratoire d'Informatique de Nantes Atlantique (LINA) UMR 6241, Ecole Polytechnique de l'Université de Nantes, France.
E-mail addresses: benmessaoud.montassar@hotmail.fr (M. Ben Messaoud), philippe.leray@univ-nantes.fr (P. Leray), nahla.benamor@gmx.fr (N. Ben Amor).

http://dx.doi.org/10.1016/j.knosys.2014.12.006
0950-7051/

In the standard Causal Bayesian Network learning procedure, an experimentation phase must be conducted on certain variables to identify the true causal links connecting them to their neighborhood. However, experiments are often difficult to conduct, greedy in terms of resources, costly or even impossible. In this context, the aim of this paper is to propose a decisional strategy for allowing more efficient causal discovery, where experiments are chosen with a great care. On the other hand, it should be noted that most of the recent knowledge-based systems are supplemented and enhanced with structured background knowledge representation such as ontologies. At first glance, it seems that Bayesian networks and ontologies have almost nothing in common but this does not preclude that some recent studies have addressed some issues related to the integration of the two formalisms.

This article provides a substantially extended version of our previous works [7–10] in which we introduce the preliminary findings for integrating a semantic distance calculus to choose the appropriate interventions. Further developments along this direction have been made in order to deploy more efficient strategies to integrate the semantic prior knowledge, improve the causal discovery process and reuse the new discovered information.

The remainder of this paper is arranged as follows: Section 2 gives the necessary background for both Causal Bayesian Networks and ontologies and discusses some related works that combine the two formalisms. Section 3 sets out how to use the ontological knowledge to enhance the causal discovery and vice versa. In Section 4, we give simulation and experimental results that prove the usefulness of the proposed algorithm. Concluding remarks and future works are given in Section 5.

2. State of the art

This section surveys some elementary notions from probabilistic graphical models and semantic knowledge-based systems, both of which are prerequisites for understanding the rest of the paper.

2.1. Probabilistic graphical models

Probabilistic graphical models (PGMs) [33] provide a framework within which applications from various domains of inherit uncertainty such as artificial intelligence, machine learning, data mining, and cognitive science will take advantage. Their popularity essentially comes from the fruitful marriage between graph theory and probability theory.

A PGM is essentially a graph $G = (V, E)$, where V denotes a set of random variables and E a set of edges. These variables may be discrete (i.e., counted within a predefined set of values) or continuous (i.e. can take on any value on a range). The presence of an edge between a pair of vertices implies a possible dependency between their corresponding variables. PGMs also provide a general-purpose modeling language for exploiting conditional independence relationships between the random variables to achieve compactness and computational efficiency.

Depending on the specific nature of the pairwise interactions among variables, there are basically three popular classes of PGMs:

- Directed ones such as Bayesian Networks [47,48,29] and Causal Bayesian Networks [26,50,58] are popular alternatives in artificial intelligence and machine learning applications. These models are more consistent in revealing unidirectional causality.
- Undirected Markov networks [36,48] are more adapted to statistical physics and computer vision. They are often used to capture the spatial correlation or mutual dependencies between random variables.
- Chain graphs [35] (hybrid graphs combining directed and undirected edges) are most useful when there are both causal-explanatory and symmetric association relations among variables, while Bayesian Networks specifically deal with the former and Markov networks focus on the later.

In this paper, we will especially focus on directed PGMs and more especially on (Causal) Bayesian Networks.

2.1.1. Bayesian networks

At the qualitative level, a Bayesian Network (BN) is a directed acyclic graph (DAG), in which directed cycles are not allowed. The (missing) edges encode conditional (in) dependence relations among the variables. At the numerical level, each node has a conditional probability table (CPT) that quantifies each of its states given every possible combination of the parent states (i.e., $P(X_i|Pa(X_i))$). This leads to express a global factorization of the joint probability distribution (JPD) over the set of random variables in the graph. When the JPD is known, it is possible to answer all possible inference queries by marginalization:

$$P(V) = \prod_{X_i \in V} P(X_i|Pa(X_i)). \quad (1)$$

However, for problems with a large number of variables, this direct approach is not practical. Fortunately, at least when all variables are discrete, we can use the locality of CPTs to make probabilistic inference (i.e., calculate the posterior probability $P(X_i|e)$ of a variable X_i assuming certain values given some observations on evidence nodes e) more efficient.

Note that several BNs can model the same probability distribution. In this case, the DAGs are called Markov equivalent [61] since they can be used to represent the same set of probability distributions. More specifically, Markov equivalent DAGs share the same skeletons and the same sets of v-structures, that is, two converging arrows whose tails are not connected by an arrow (e.g. $X_i \rightarrow X_j \leftarrow X_k$).

There are various proposals in the literature to provide a representative class for Markov equivalent BNs. One of them is the essential graph or complete partially directed acyclic graph (CPDAG) [3]. An illustration of those Markov equivalent representations is given in Fig. 1.

Definition 2.1. A PDAG (Partially Directed Acyclic Graph) is used to represent an equivalent class of DAGs. A PDAG contains the same skeleton and v-structures as the original DAGs but possesses both directed and undirected edges. Several PDAGs representing the same equivalence class of DAGs can be found.

Definition 2.2. A CPDAG (Complete Partially Directed Acyclic Graph) is a PDAG except that a CPDAG is unique for an equivalence class of DAGs. Every directed edge $X_i \rightarrow X_j$ of a CPDAG denotes that all DAGs of this class contain this edge, while every undirected edge $X_i - X_j$ in this CPDAG-representation denotes that some DAGs contain the directed edge $X_i \rightarrow X_j$, while others contain the opposite edge $X_i \leftarrow X_j$.

To extract from a DAG its corresponding CPDAG, Meek [41] proposed an implementation for the DAG-to-CPDAG algorithm. The idea of this rule-based strategy is as follows. First, we make all edges in the DAG undirected, except for those participating in a v-structure. Then, we repeatedly apply the Meek rules [41] that transform undirected edges into directed edges. Those transformation rules have been formally proven sound and complete [42]. When no rule matches on the current graph, that graph must be a completed PDAG.

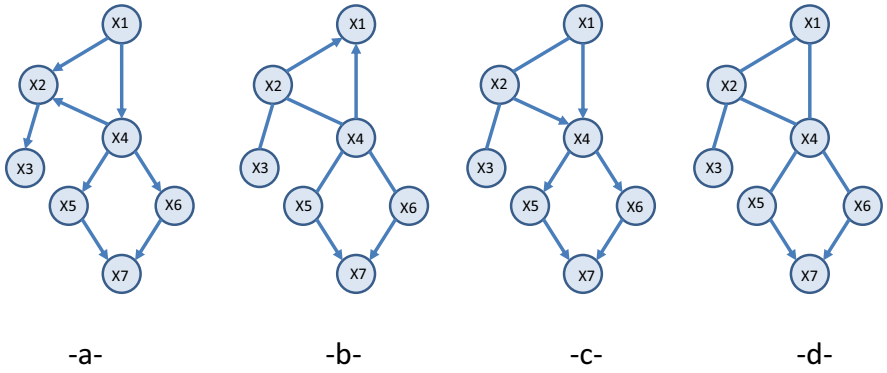

Fig. 1. (a) Original DAG, (b) and (c) two PDAGs corresponding to the same equivalence class and (d) the associated CPDAG.

Learning BN structure is known to be an NP-hard task [14] since the space of DAGs grows exponentially in the number of variables. Up to now several heuristics have been proposed for learning the structure of BNs using *observational data*. These techniques can be divided into two main categories, namely *score-based* and *constraint-based* algorithms.

*Score-based*algorithms [16,18] attempt to identify the network that maximizes a scoring function evaluating how well the network fits the data while *constraint-based* algorithms [38,57] look for (in) dependencies in the data and try to model that information directly into the graphical structure. However, these approaches often admit multiple candidate causal structures within equivalence and local optima.

2.1.2. Causal Bayesian networks

The biggest problem when learning BNs from observational data is that we simply do not observe causal relationships. What we really observe is the cause, the effect and the fact that they occur in a fixed pattern. This correlation implies an unresolved causal structure. This empirical point of view makes BNs inappropriate for recovering fully causal structures. That is why an extension of traditional BNs, where directed edges represent correct causal influences, is needed. To learn Causal Bayesian Networks (CBNs), we have to collect further information on causality via interventions (i.e., actions tentatively adopted without being sure of the outcome). In this context, Eberhardt et al. [21] showed that $N-1$ experiments are sufficient and, in the worst case, necessary to discover the causal structure among a set of N variables if at most one variable can be subjected to an intervention in any one experiment.

Using this form of representation, we can exploit the causal inference, the major asset of CBNs, to ask intervention queries. Such queries arise in situations where we want to know the effect of a given perturbation $do(X_i = x_i)$ (i.e., fixing the value of X_i to x_i) on the outcome Y. Here, we should note that intervening on a system may be very expensive, time-consuming or even impossible to perform. For this reason, the choice of variables to experiment on can be vital when the number of interventions is restricted. All those distinguishing features have motivated many researchers to develop a variety of techniques and algorithms to learn such models [28,43,44,46].

2.1.3. MyCaDo sketch

The MyCaDo [43,44] algorithm is one of the active techniques suggested to discover causal structures further from a Markov equivalence class using a design strategy for planning experiments. In what follows we give a brief sketch of each of the principle phases of MyCaDo algorithm (refer to the right-hand part of Fig. 4) for later purposes. Given a perfect observational dataset, the first phase consists in learning the CPDAG using traditional structure learning techniques. Meganck [43] and Meganck et al. [44] especially used PC algorithm with modified orientation rules that were proved complete and sound by [41].

The second phase of MyCaDo approach employs some techniques from decision theory to select the best experiments to perform and, hence, direct the maximum number of edges in the partially directed structure. So before performing an experimentation on X_{cand} (the term *cand* stands for "*candidate*"), MyCaDo measures all neighboring variables and, accordingly to the result, directs all edges connecting X_{cand} and its neighbors $Nei(X_{cand})$. This edge orientation represents one graph instantiation ($inst(A_{X_{cand}})$) among all possible instantiations. It is then possible to continue the edge orientation by using the Meek rules [41] to infer new causal relations. Let $inferred(inst(A_{X_{cand}}))$ be the number of inferred edges based on $inst(A_{X_{cand}})$. MyCaDo proposes that the utility of an experiment is related to the number of edges directly oriented or inferred, weighted by the cost of experiment ($cost(A_{X_{cand}})$) and measurement ($cost(M_{X_{cand}})$):

$$U(X_{cand}) = \frac{Card(Nei(X_{cand})) + Card(inferred(inst(A_{X_{cand}})))}{\alpha.cost(A_{X_{cand}}) + \beta.cost(M_{X_{cand}})} \quad (2)$$

where measures of importance α and $\beta \in [0,1]$ and Card() represents the cardinality of any set.

2.2. Knowledge-based systems

A Knowledge-based system (KBS) provides a consistent reasoning framework doted with an inference engine that deductively reason over a logical language. Ontology is one such kind of knowledge representation framework designed to support the reasoning task of a knowledge-based system. There are different definitions in the literature of what should be an ontology. The most popular one was given by [27], stipulating that an ontology is an *explicit* specification of a *conceptualization*. The "conceptualization", here, refers to an abstract model of some phenomena having real by identifying its relevant concepts. The word "explicit" means that all concepts used and the constraints on their use are explicitly defined.

An ontology usually contains modeling primitives such as:

- a set of concepts or classes $C = \{C_1, \ldots, C_n\}$ structured by means of taxonomic (is-a) and partonomic (part-of) hierarchy $\mathcal{H}$,
- concept properties or attributes,
- semantic relations between concepts ($\mathcal{R}_c : C_i \times C_j$),
- a set of concept (resp. relation) instances $\mathcal{I}$ (i.e., occurrences of classes and semantic relations),

- a set of formal axioms $\mathcal{A} = \langle c_{ik}, c_{jm}, v_n \rangle$ with $c_{ik}, c_{jm} \in \mathcal{I}$ and $v_n \in \mathcal{V}$ (i.e., a set of constraint-relationships like must, must not, should, and should not).

All of these components are depicted in Fig. 2, where concepts are tagged by yellow circles and instances are marked with blue diamonds. The is-a relations concern inter-related concepts and the non-labeled ones indicate instantiation relationships. We distinguish between two types of causal relations in the ontology. The first ones which are indicated in solid lines build causal connections between the ontology concepts. The other type in dashed lines consider more specific causal relations that exist between concept instances. We restrict the use of semantic relations to only causal ones between concepts since they are the main relations recovered in our approach.

With the influence of reasoning systems, standardized languages, such as RDF [34] and OWL [40], have played a role of particular importance in providing data modeling specification to represent machine-interpretable semantics of information. These languages are based on descriptive logic (DL), which enables them to infer implicit knowledge from the ontology and check its consistency.

Simultaneously, there are several reasons to opt for continuous changes in the ontology to make semantic inference more robust and reliable. Such proposals can take different forms such as a change in the domain, the diffusion of new discoveries or just an information received by some external source [22]. There are many ways to change an ontology in response to the fast-changing environment. One possible direction is the ontology evolution, which consists in taking the ontology from one consistent state to another by updating (adding or modifying) the concepts, their properties and the associated relations [31]. The ontology evolution can be of two types:

- *Ontology population*: when new concept instances are added, the ontology is said to be populated.
- *Ontology enrichment*: which consists in updating (adding or modifying) concepts, properties and relations.

In the rest of this paper, we will especially focus on the second type of ontology evolution. Following the ontology evolution cycle (see Fig. 3), we can distinguish six phases employed to guide ontology change validation in a systematic and optimized way. Each of these phases will be covered in more details in the context it occurs in our algorithm. In order to establish the context in which the ontology evolution takes place, the principle of ontology continuity should be fulfilled. It supposes that the ontology evolution

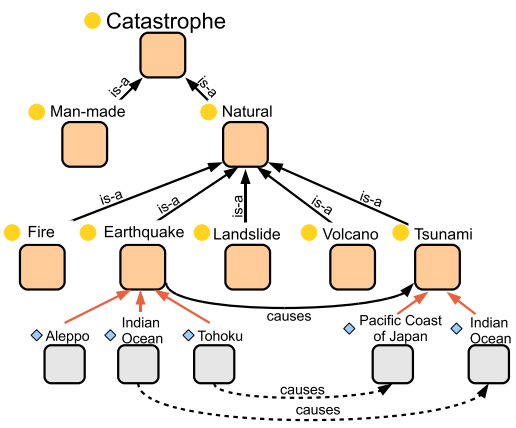

Fig. 2. An illustrative example of Risk and Catastrophe Ontology.

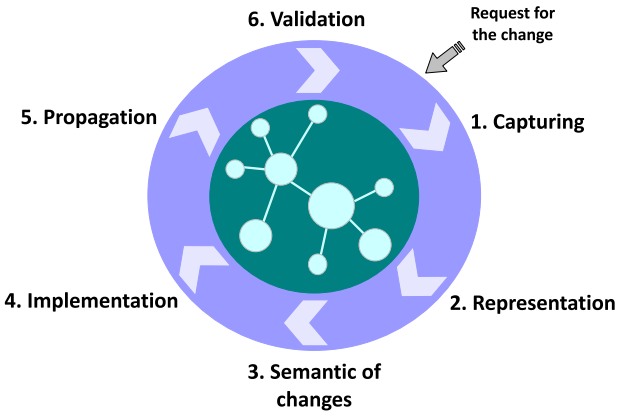

Fig. 3. The six-phases of ontology evolution process [37].

should not make false an axiom that was previously true. When changes do not fulfill the requirement of ontological continuity, it is no more an evolution, it is rather an ontology revolution [63].

The knowledge acquisition during the ontology evolution is done by applying automatic (or semi-automatic) knowledge extraction processes such as text mining approaches. This provides support for continual ontology improvement over time as documents are added or removed. This practice is particularly prevalent in the biomedical field where large quantities of scientific articles are used to discover new relationships between known concepts (e.g., symptoms, diseases, proteins, etc.) [17]. Nevertheless, these knowledge extraction tools merely reproduce the shared knowledge without making any real new discoveries.

2.3. PGM and ontology cooperation: Review of relevant literature

Recent studies have suggested some ways to combine both ontologies and PGMs. We especially investigate contributions that formalized the cooperation between the two formalisms.

2.3.1. Modeling uncertainty in ontologies

The first line of research focused on how to integrate the power of PGMs to enhance the potential of ontologies by supplementing it with the principal means of modeling uncertainty. In this context, Albagli et al. [1] have introduced iMatch, a probabilistic scheme for ontology matching based on Markov networks, Wu and Weld [62] have also presented the KOG system that builds a rich ontology by combining Wikipedia infoboxes with WordNet using Markov logic networks, and [64] have proposed the OntoBayes approach, an ontology-driven Bayesian model for uncertain knowledge representation, to extend ontologies to probability-annotated OWL in decision making systems. This approach was later complemented with a concrete service oriented framework for decision support systems [65].

Some other approaches have investigated how to provide languages to capture uncertainty. Zhang et al. [67] have proposed a prototype implementation for BayesOWL [20], a probabilistic framework that augments and supplements OWL for representing and reasoning with uncertainty based on BNs. Similarly, Settas et al. [54] have presented the BNTab Protégé[1] plugin, another tool for quantifying uncertainty in the antipattern ontology by capturing and visualizing probabilistic causal relationships between antipattern variables.

[1] Ontology-building editor and API: http://protege.stanford.edu.

2.3.2. Ontology driven construction of PGMs

Other researches have enhanced the PGM construction by integrating ontologies. For example, Jeon and Ko [30] have developed a semi-automatic BN construction system based on e-health ontologies. Their framework enables probabilistic inferencing capabilities for various E-health applications and contributes to reduce the complexity of BN construction. A similar approach was proposed by [19], in which they present an automatic solution for BN construction, implemented in the context of an adaptive, self-configuring network management system in the telecommunication domain. More recently, Anantharam et al. [2] developed an approach to extract the structure of a PGM from observations of a cyber-physical system and declarative knowledge about the domain in the form of ontologies and linked open data.

2.3.3. PGM-ontology combination

Other applications have exploited the combination of the two formalisms to cumulate the advantages they provide. For instance, Zheng et al. [68] have proposed an ontology-based BN approach to represent the uncertainty in clinical practice guidelines. In the biological field, GO-Bayes, a tool proposed by [66], performs overrepresentation analysis in the Gene ontology using BNs by investigating whether gene annotations are statistically overrepresented in identified Gene ontology terms. Tagliasacchi and Masseroli [60] have implemented a post-processing method that uses a BN to eliminate predictions of anomalous annotations that are inconsistent with the Gene ontology structure. And finally, McGarry et al. [39] have provided a framework for integrating high level biological knowledge obtained from Gene ontology with low level information extracted from a BN.

In our previous work [8], we focused on only one facet of the CBN–ontology combination by integrating the ontological knowledge to learn CBNs. Taking a further step in the same direction, this paper proposes to reuse discovered causal relationships to enrich the ontologies. A similar two-way approach [5,6] have been conducted to learn object-oriented Bayesian networks from ontologies and vice versa.

3. SemCaDo: A serendipitous causal discovery algorithm for ontology evolution

Generally, in the research area, scientific discoveries represent a payoff for years of well-planned works with clear objectives. This affirmation did not exclude the case of other important discoveries that are made while researchers were conducting their works in totally unrelated fields and the examples are abundant from Nobel's flash of inspiration while testing the effect of dynamite to Pasteur brainstorm when he accidentally discovered the role of attenuated microbes in immunization.

In this way, we propose a new causal discovery algorithm which stimulates serendipitous discoveries when performing the experimentations using the following CBN–Ontology correspondences.

3.1. CBN–ontology correspondences

The question that emerges at this stage is how to manage the duality between serendipity and causality in order to provide an innovative design for assessing causal discoveries. Before proceeding with the correspondences among CBNs and ontologies, let us first investigate the constraints that we impose all along this paper:

- Only a single domain ontology should be specified for each causal discovery task.
- The ontology evolution should be realized without introducing inconsistencies or admitting axiom violations.

Taking these criteria and constraints into account, we first propose to assign a single concept (C_i) from the domain ontology to each CBN node (X_i). Accordingly, the causal dependencies ($\mathcal{E}$) represented by directed links in the CBN will be depicted by the specific causal relations ($\mathcal{R}_c$) between the appropriate concepts in the ontology. More precisely, the present study focuses on semantic causal relations among which only one concept attribute is considered in the ontology. On a more fine-grained level, we can associate both observational and experimental data ($D_{obs/exp}$) to the state instantiations ($\mathcal{I}$) of the ontology concepts.

3.2. SemCaDo sketch

Based on a domain ontology, our approach relies on extending the MyCaDo algorithm [43,44] (cf. Section 2.1.3) in order to incorporate available prior causal knowledge. The original characteristic of our algorithm named Semantic Causal Discovery (SemCaDo) is essentially its ability to discover and reuse the capitalized knowledge in CBNs to make an ontology evolution. The general overview of the SemCaDo algorithm is given in Fig. 4 and Algorithm 1. SemCaDo inputs an observational dataset and a corresponding domain ontology. As outputs, SemCaDo returns a CBN and a list of new discovered causal relationships. The whole process takes place through three main phases.

Algorithm 1. SemCaDo

Input: *Observational dataset* D_{obs}, *Ontology* O, *System* $S, cost(A_{X_i}), cost(M_{X_i}), \alpha, \beta$

1. $Const = ExtractConstraints(O)$
2. $PDAG = LearnInitialStructure(D_{obs}, Const)$
3. **for** all X_{cand} with undirected edges **do** $U(X_{cand}) = CalculateUtility(O, cost(A_{X_{cand}}), cost(M_{X_{cand}}), \alpha, \beta)$ **end for**
4. $X_{best} = argmax(U(X))$
5. $D_{exp} = PerformExperiment(S, X_{best})$
6. $PDAG = InstantiateUndirectedEdges(X_{best}, D_{obs}, D_{exp})$
7. $PDAG = ApplyMeekRules(PDAG)$
8. Return to step (3) until all links are directed or no additional experiment can be performed
9. $G := PDAG$
10. $newK = ValidateResultsWithExperts(G, O)$
11. $O' = EnsureOntologyEvolution(O, newK)$

Output: *CBN* G, *enriched ontology* O'

3.2.1. Learning the initial structure using causal prior knowledge

The ontology in input may contain some causal relations in addition to hierarchical and semantic relations. Those causal relations should be integrated into the structure learning process from the beginning in order to reduce the task complexity and to get better final output. Therefore, such direct cause-to-effect relations will be incorporated as constraints when applying structure learning algorithms (Refer to step 1 in Algorithm 1). Our main objective is to narrow the corresponding search space by introducing some restrictions that all elements in this space must satisfy.

In our context, the only constraint that will be defined is edge existence. But we could also imagine in future work that some axioms in the ontology also give us some information about forbidden edges. All these edge constraints can easily be incorporated in usual BN structure learning algorithm [13]. Under some condition of consistency, these existence restrictions shall be fulfilled, in the sense that they are assumed to be true for the CBN representing

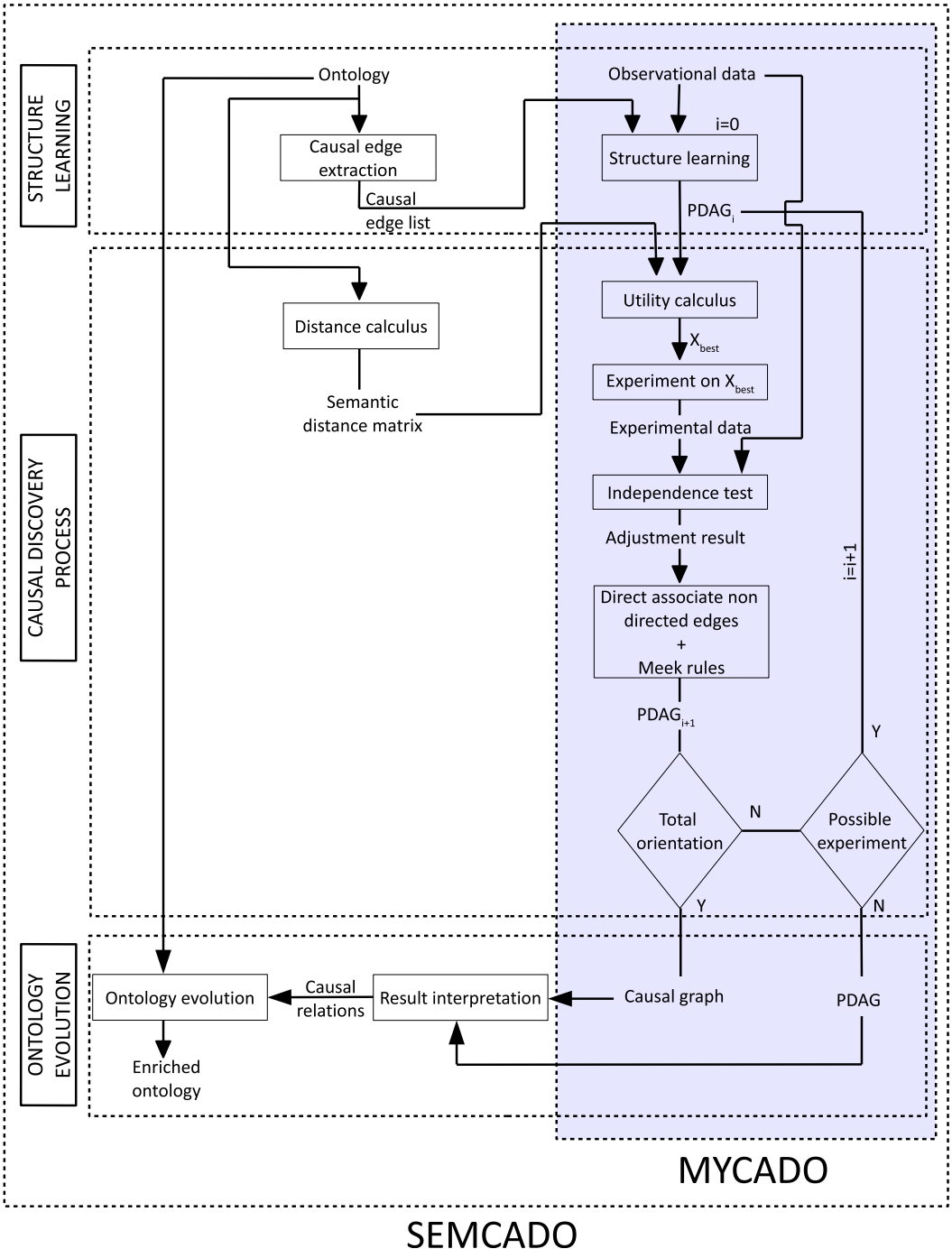

Fig. 4. SemCaDo: Extending MyCaDo to allow CBN–ontology interactions.

the domain knowledge, and therefore all potential Partially Directed Acyclic Graph (PDAG) must necessarily satisfy them (cf. Definition 2.1).

Definition 3.1. Given a domain ontology O, let G=(C, R_c) be the DAG where $R_c : C_i \times C_j$ represents the subset of semantic causal relations extracted from O. This subset included both direct and logically derivable semantic causal relations. Let H=(X, E_h) be a PDAG, where X is the set of the corresponding random variables and E_h corresponds to the causal dependencies between them. H is consistent with the existence restrictions in G if and only if:

$$\forall C_i, \quad C_j \in C, \quad if\ C_i \rightarrow C_j \in R_c\ then\ X_i \rightarrow X_j \in E_h.$$

When we are specifying the set of existence restrictions to be used, it is necessary to make sure that these restrictions can indeed be satisfied. In fact, such causal integration may lead to possible conflicts between the two models. When this occurs, we have to maintain the initial causal information in the PDAG since we are supposed to use perfect observational data. On the other hand, we should ensure the consistency of the existence restrictions in such a way that no directed cycles are created in G.

3.2.2. Semantic-based causal discovery process

It seems obvious that the gained information of the utility function employed in MyCaDo is essentially the node connectivity (i.e., the number of undirected edges and those susceptible to be

inferred) which serves to orient the maximal number of edges but not necessarily the most informative ones. To cope with this limitation, the strategy we propose in our approach makes use of a semantic distance calculus [12] provided by the ontology structure. So, for each node in the graph, SemCaDo gives a generalization of the node connectivity by introducing the semantic inertia, denoted by $SemIn(X_i)$ and expressed as follows:

$$SemIn(X_i) = \frac{\sum_{X_j \in Nei(X_i) \cup X_i} dist_{Sem}(mscs(Nei^*(X_i) \cup X_i^*), X_j^*)}{Card(Nei(X_i) \cup X_i)} \quad (3)$$

where L^* represents the set of concepts relative to the set of nodes L, $mscs(L^*)$ is the most specific common subsumer of the set of concepts L^* and $dist_{Sem}(C_i, C_j)$ evaluates the disaffection between C_i and C_j.

Moreover, the semantic inertia $SemIn(X_i)$ presents three major properties:

- When the experimented variable and all its neighbors lie at the same level in the concept hierarchy, the semantic inertia will be equal to the number of hierarchical levels needed to reach the *mscs*.
- If the corresponding concepts have the same parent in $\mathcal{H}$, then *SemIn* will be proportional to $Card(Nei(.))$.
- It essentially depends on semantic distance between the studied concepts. This means that the more important the distance is, the more maximized the *SemIn* will be.

In this way, we will accentuate the serendipitous aspect of the proposed strategy and investigate new and unexpected causal relations on the graph. Further, we also integrate a semantic cumulus relative to the inferred edges denoted by *Inferred_Gain* in the utility function. For this purpose, we use $I^*(X_i)$ to denote the set of concepts corresponding to nodes attached by inferred edges after performing an experimentation on X_i. So, the *Inferred_Gain* formula is expressed as follows:

$$Inferred\_Gain(X_i) = \frac{\sum_{X_j \in I(X_i)} dist_{Sem}(mscs(I^*(X_i)), X_j^*)}{Card(I(X_i))} \quad (4)$$

Inferred_Gain also represents a generalization of $Card(inferred(inst(.)))$ in Eq. (2) and depends on the semantic distance between the studied concepts.

When using the two proposed terms, the utility function will be as follows:

$$U(X_i) = \frac{SemIn(X_i) + Inferred\_Gain(X_i)}{\alpha.cost(A_{X_i}) + \beta.cost(M_{X_i})} \quad (5)$$

where the two measures of importance α, β are usually chosen proportionally ($\alpha, \beta \in [0,1]$ and $\max(\alpha, \beta) \neq 0$).

This utility function will be of great importance to highlight the serendipitous character of SemCaDo algorithm by guiding the causal discovery process to investigate unexplored areas and conduct more informative experiments (Refer to steps 3 and 4 in Algorithm 1).

3.2.3. Edge orientation and ontology evolution

Once the specified intervention has been performed, we follow the same edge orientation strategy (step 5 in Algorithm 1) as in MyCaDo [43,44]. So if there are still some non-directed edges in the PDAG, we re-iterate over the second phase and so on, until no more causal discoveries can be made. Since certain experimentation cannot be performed, either because of ethical reasons or simply because it is impossible to do it, the final causal graph can be either a CBN or a partially causal graph.

In both cases, the causal knowledge will be extracted and interpreted for an eventual ontology evolution (Refer to step 10 in Algorithm 1). In this way, the causal relations will be translated into semantic causal relations between the corresponding ontology concepts (Refer to step 11 in Algorithm 1). We also note that only causal relations ensuring semantic consistency will be retained for the ontology evolution process. For this purpose, SemCaDo algorithm uses the six-phases evolution process (previously shown in Fig. 3):

- *Change capturing:* the aim of this initial step in the ontology evolution process is to capture the new discovered causal relations on the current causal graph which are not actually modeled. It starts after finishing the structure learning step to treat all changes in a consistent and unified manner.
- *Change representation:* these causal changes need to be represented formally, explicitly and in a suitable format to be correctly implemented. In the context of SemCaDo algorithm, we only handle elementary changes [59] (i.e., restricted to adding semantic causal relations) that cannot be decomposed into simpler ones.
- *Semantics of change:* the semantics of change is the phase that enables the resolution of ontology changes in a systematic manner by ensuring the consistency of the ontology. In our case, conflicting knowledge is highly possible to occur when deducing causal conclusions from the ontology axioms. Such inconsistencies should be handled by automated reasoning. This step also prevents the creation of new cycles in the ontology when integrating the causal discoveries. This consistency rule is maintained since the causal discovery step in SemCaDo avoids the creation of cycles during the structure learning.
- *Implementation:* to avoid unwanted changes, a list of all consequences in the ontology and dependent artifacts should be generated and presented to the ontology engineer, who should then be able to accept the change or reject it. If the implementer agrees to add the new causal relationships, all actions to apply the change have to be performed.
- *Propagation:* pursuing and adopting the new causal discoveries can generate additional changes in other parts of the ontology. These changes are called derived changes. That is why, during this step, it is necessary to determine the direct and indirect types of changes to be applied. In case of ambiguity, the ontology expert decides the action to perform. A human intervention at this level is essential to remove ambiguity and to make the final decision.
- *Validation:* change validation enables justification of performed changes and undoing them at user's request. If the output of SemCaDo causal discovery step is a partially directed graph, it is possible to restart the cycle when there is sufficient budget to make further discoveries.

4. Experimental study

SemCaDo and MyCaDo [43,44] have both been implemented in our local platform dedicated to probabilistic graphical modeling and learning. This platform in C++ uses Boost Graph API[2] for graph manipulation and ProBT API[3] for BN modelling.

To evaluate the effectiveness of our approach, we first resort to simulations in order to give accurate results on the performance of SemCaDo compared to MyCaDo in various situations. MyCaDo will serve as comparison reference to SemCaDo since both of them share the same assumptions and use the same observational input data. Then, an experimental evaluation using real cellular network

[2] http://www.boost.org/.
[3] http://www.probayes.com/.

[11] will be undertaken to confirm the effectiveness of the proposed technique.

4.1. Validation on synthetic data

We first proceed through simulations to achieve a SemCaDo-MyCaDo performance comparison and validate the causal discovery process in both strategies. As a starting point, we randomly create a set of synthetic 50 and 200 node graphs. For each simulated graph, we automatically generate a corresponding ontology, defined by a hierarchy of concepts and some causal relations. Specifically, we integrate a varying percentage (10–40%) of the initial causal relations. Still, as long as we do not dispose of a real system to intervene upon, we decide to simulate the experimentations directly in the previously generated CBNs as in [43,44] and choose equal measures of importance when calculating the expected utilities (i.e., $\alpha = \beta = 1$ in Eq. (2)).

To perform the experimentation, we propose to mutilate (i.e., disconnect) the node X_{best} from $Pa(X_{best})$ in the DAG such that the manipulated variable become totally independent of its parents in the post-intervention distribution [49]. We force X_{best} to take on random values and then sample the post-intervention distribution to get the experimental data.

The obtained dataset as well as the initially supplied observations will be transferred to the conditional independence χ^2 test in order to determine if the variable under experiment is the cause or the effect of its neighboring variables. Now that the experimental setup has been described, we will detail the three main blocks of our strategy (refer to Fig. 4 in Section 3.2 to follow the cycle in more details):

- *Structure learning:* We apply a DAG-to-CPDAG algorithm [15] on those CBNs in order to simulate the result of a structure learning algorithm working with a perfect infinite dataset (refer to Section 2.1.1).
- *Causal discovery process:* To evaluate the reconstruction results with both MyCaDo and SemCaDo, we propose to compute the semantic gain of each experiment X_i as follows:

$$Semantic\_Gain(X_i) = \sum_{X_j \in O(X_i)} dist_{Sem}(mscs(O^*(X_i)), X_j^*) \quad (6)$$

where O^* represents the set of concepts relative to the set of nodes O that become linked by oriented edges after performing an experiment on X_i, $mscs(O^*)$ is the most specific common subsumer of the set of concepts O^*, and $dist_{Sem}(C_i, C_j)$ is the semantic distance between C_i and C_j.
Throughout the causal discovery phase, we will use the Rada's distance [52] as the specific distance when calculating the semantic gain. The choice of this distance is motivated by its extensive use in the literature (see [12] for a comparative study between semantic distances).

Definition 4.1. Let C_i and C_j be two concepts in an is-a (resp. part-of) hierarchy. A measure of the conceptual Rada's distance is defined as the minimum number of edges separating the given two concepts.

The intent here is to update a semantic cumulus after each SemCaDo (resp. MyCaDo) iteration as shown in Fig. 5. In both modeling strategies, the two corresponding curves are increasing in, meaning that the higher the number of experimented variables, the higher the value of the semantic gain. Nevertheless the faster the curve is increasing, the more impressive and better experiments the approach is converging to in term of semantic interpretation. The combined results indicate that our approach comfortably outperforms MyCaDo in term of semantic gain. This is essentially due to the initial causal knowledge integration and the causal discovery strategy when performing the experimentations. However, the two curves reach a common semantic maxima when obtaining a fully directed graph. This is always the case since, without using the same experimentations, the two strategies orient the same number of edges when finishing the experimental process. In this regard we should remember that we are approaching a decision problem which is subject to the experimentation cost and the budget allocation. Taking this constraint into account, the SemCaDo's advantage will be extremely beneficial when the number of experiments is limited.

- *Ontology evolution:* This step is not significant in the current experimental design since we are generating the ontology from the simulated graphs.

4.2. Validation on Sacharomyces cerevisiae cell cycle microarray data

Discovering and modeling gene regulatory circuitry from both observational and experimental data is one of the most challenging problems facing biologists today. This is essentially due to the non negligible number, duration and cost of experiments and the lack of facilities for conducting genetic[4] (resp. environmental[5]) perturbations. In such circumstances, it would be far better to propose an experimental design to cope with the lack of data and provide maximal expected information. In this context, we propose to validate our approach using *S.* cerevisiae cell cycle microarray data [56] and the corresponding Gene Ontology annotations.

4.2.1. Data description

The experimentation using real biological systems requires the use of gene-expression microarray data, the Gene Ontology and causal pathway repositories.

- *Gene expression dataset:* We consider the Yeast *S.* cerevisiae cell cycle microarray data [56] since the Yeast genome is relatively small compared to more complex eukaryote organisms and highly annotated with Gene Ontology functions. In this dataset, the mRNA concentrations of nearly 6178 genes were measured with three independent fluorescence measurement methods. Overall, the dataset contains 73 sampling points for all genes. Each of them is measured in different phases of the yeast cell cycle. According to [56], about 800 of these genes have been reported with varying transcripts over the cell cycle stages.
- *Gene ontology:* Most of the *S.* cerevisiae genes are annotated with specific biological functions from the Gene Ontology[6] (GO), which remains the most popular initiative aiming at providing a structured, precisely defined, and dynamic controlled vocabulary to facilitate the description of gene roles and gene product attributes in the eukaryotic genome. The GO structure is in the form of a rooted DAG where nearly 30000 concepts are formalized into three related (sub-) ontologies, referred to as molecular function, cellular component and biological process. According to the GO consortium, these GO domains represent three separate ontologies which are unrelated by a common parent node.
  The directed edges between concept nodes represent either subsumption links ("is-a") or composition relationships

[4] Gene knockout (deletion of the gene), or overexpression (setting the expression level higher than its usual level).
[5] Change in one or more non-genetic factors, such as a change in environment, nutrition, pressure or temperature.
[6] http://www.geneontology.org/.

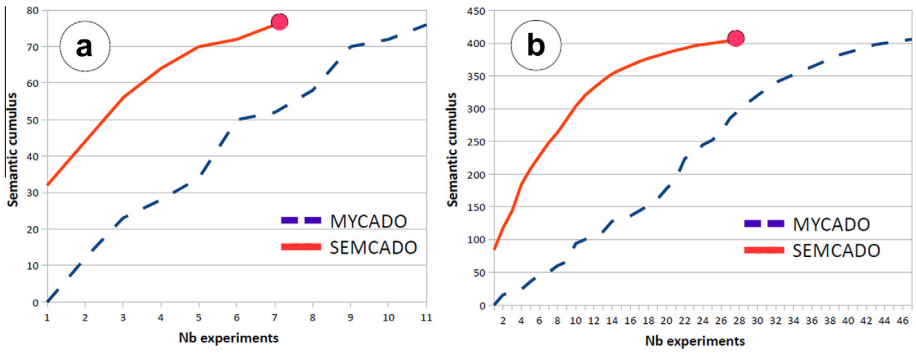

Fig. 5. The semantic gain given the number of experiments using MyCaDo and SemCaDo on relatively small graphs (a) and bigger ones (b).

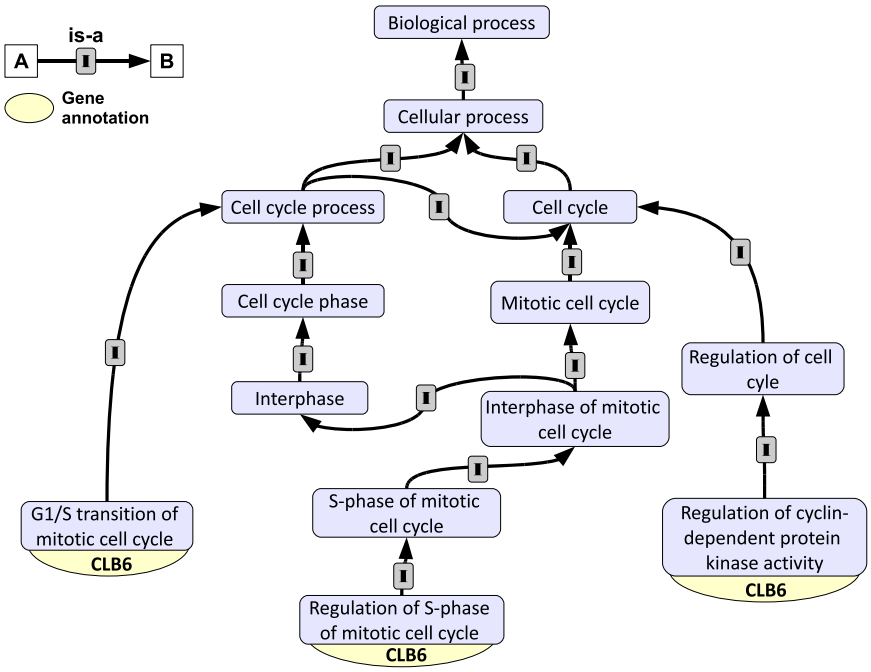

Fig. 6. CLB6 multiple localizations in GO.

("part-of"). The GO concepts are given a unique ID number in the form of GO:N (where N is a natural number) to identify and characterize some biological properties. This GO structure (concepts + relationships) reflects the current representation of biological knowledge as well as serving as a guide for classifying new data.

According to the existing biomedical literature's assertions, the gene products may be annotated to as many GO concepts as needed and at the most specific levels. For instance, as shown in Fig. 6, the gene CLB6 is involved in:

1. regulation of cyclin-dependent protein kinase activity (GO:0000079),
2. regulation of S phase of mitotic cell cycle (GO:0007090),
3. G1/S transition of mitotic cell cycle (GO:0000082).

Such a classification will provide a higher-level understanding of how tissue-specific genes are regulated and expressed.

Given two other genes NPL3 and UFE1, which are respectively annotated with the cell nucleus (GO:0005634) and the SNARE complex (GO:0031201), we show in Fig. 7 the multiple paths that can be found between them. As discussed above, we set the cell part term (colored in red) as the *mscs* of the two studied concepts. If there are multiple paths between any two concepts and their *mscs*, only the shortest one is considered. The red dashed lines indicate, in our case, the optimal path according to the GO structure. We note that the best GO-distance between two genes can be equal to 0 when both of them are annotated to the same GO concept.

- *Causal pathway repositories:* However, since the GO structure consists essentially of hierarchical classification, we will be unable to extract or enrich the GO with regulatory pathways. An alternative source of causal relations is the Biochemical Pathway Repositories where regulatory information could be available. Fueled by the availability of experimentally determined pairwise gene interactions, different datasets for delineating the biochemical pathways and reactions have been merged. Most of these scientific databases, such as Data Repository of Yeast Genetic Interactions (DRYGIN)[7] [32], enable a convenient access to genes in terms of the biological pathways in which they intervene [4]. Fig. 8 shows a DRYGIN

[7] http://drygin.ccbr.utoronto.ca/.

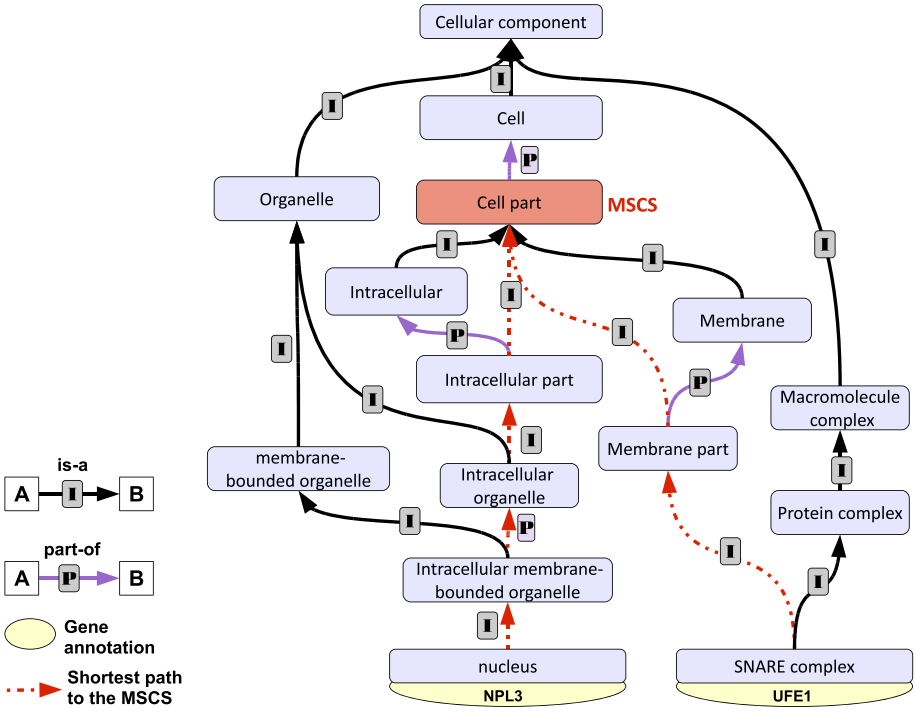

Fig. 7. The shortest path between NPL3 and UFE1 in GO.

screen capture for the top regulatory pathways involving the gene CLB6.

4.2.2. Experimental design

As in Section 4.1, we continue to simulate the experimentations directly on the Gene Regulatory Network (GRN). Here, we have to adapt some of the initial CBN–ontology correspondences that we provide in Section 3.1. Henceforth, the GRN nodes which correspond to genes will be assigned to the most specific level of the Gene Ontology using term annotations (i.e., instances). The rest of the correspondences remain unchanged. Given these correspondences, our analysis will be shaped by two design choices:

First, we will evaluate our approach using all available annotations in a way such that multiple annotations can correspond to a single gene. Then, we propose to modify our benchmark data by sorting each studied gene to one of its multiple localizations in the GO. In fact, when allowing the studied genes to be fully annotated with all GO terms describing their biological functions, we may be left with no alternative than treating regulatory relations between "semantically" adjacent genes. Such a factor would introduce even more surprising relations among the GO terms. In Fig. 9, two histograms are used to plot the distributions of the semantic distance between concepts by using both single and multiple annotations. When comparing the two histograms, we notice that the frequency distribution in the blue colored histogram is heavily skewed to left, giving credence to the claim that original GO includes more proximal genes than the biased one.

In order to overcome the lack of real experimental data needed to build the causal model, we propose to use the GRN of [24] as a starting causal model and the GO structure as a source for calculating semantic distances between genes. From a modeling standpoint, a GRN can be thought of as a DAG $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ where $\mathcal{V}$ is the set of n gene nodes (resp. protein concentrations and other experimental conditions) and $\mathcal{E}$ is the set of directed edges among the nodes belonging to $\mathcal{V}$. Such models are well suited for representing cellular processes (i.e., metabolism, signal transduction and transport).

- *Structure learning:* Using the Yeast *S.* cerevisiae cell cycle microarray data [56,24] proved that they were able to extract a finer structure of regulatory interactions between genes. Their heuristic approach aimed at focusing on a pair of features that are common to high-scoring networks. The first type of features they identified is the high confidence Markov relations[8] which assume that a gene interaction exists between two genes if no variable in the model mediates the dependence between them. The second feature is synonymous to causality in the model since, out of all 800 genes they treat, only a few seem to dominate the order (i.e., appear before many other genes) in the overall networks of a given equivalence class. The intuition is that precedence over the ordering is indicative of potential cause-to-effect relationships on the cell-cycle process. Screen capture of the considered causal graph is shown in Fig. 10.
  The main reason for choosing the GRN of [24] as starting model is that it is free from assumptions and do not reuse any prior knowledge. We also note that interactions between genes other than causal relationships are not considered.
  We initially need this causal graph to sample normally distributed data that will be fed to the Greedy search (GS) algorithm [23] in order to re-discover the original network. Here we consider the BIC as our model selection criterion since this provides a good approximation to the marginal likelihoods. Simultaneously, we slightly modify GS in order to permit the experimenter to integrate prior knowledge about the direction of some edges in the final graph. These "hard" restrictions are assumed to be true for the resulting BN, and therefore all the candidate BNs must satisfy them. Three cases were considered in which we incorporate 20% (resp. 40% and 60 %) of the edges being present in the initial graph. For proper sampling, we have chosen the causal priors uniformly at random from the set of all edges. Then, once the BN is con-

[8] Are pairwise Markov relations the only conditional independence relations encoded by the graph.

Query Gene			Array Gene		
Common Name	ORF	Aliases	Common Name	ORF	Aliases
lcb1-5	YMR296C	TSC2|END8	CLB6	YGR109C	Aliases
YDL133W	YDL133W		CLB6	YGR109C	Aliases
MIA40_damp	YKL195W	TIM40|FMP15	CLB6	YGR109C	Aliases
CSM1	YCR086W		CLB6	YGR109C	Aliases
MNE1	YOR350C		CLB6	YGR109C	Aliases
ERS1	YCR075C		CLB6	YGR109C	Aliases
COX10	YPL172C		CLB6	YGR109C	Aliases
cdc11-5	YJR076C	PSL9	CLB6	YGR109C	Aliases
ARO1	YDR127W		CLB6	YGR109C	Aliases
YFH1_damp	YDL120W		CLB6	YGR109C	Aliases

Fig. 8. Screen capture of the top DRYGIN regulatory pathways involving the gene CLB6.

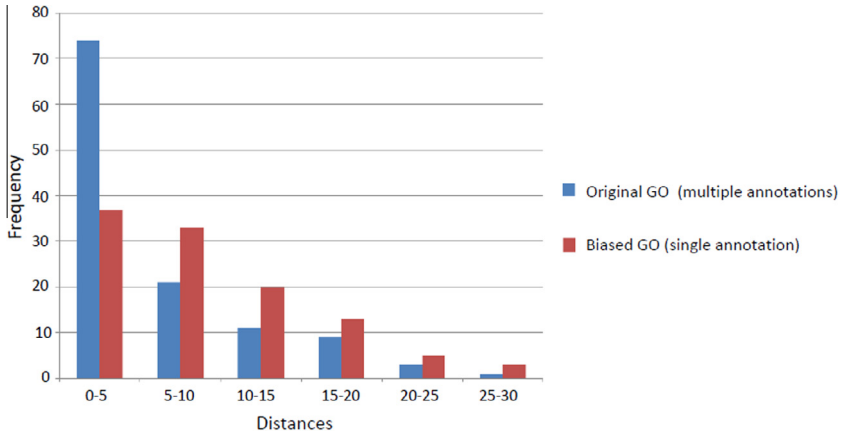

Fig. 9. Distance frequency distributions using multiple (resp. single) gene annotations in the GO.

structed, we need to execute the DAG-to-CPDAG algorithm [15] in order to keep only the edge directions which are statistically significant.

- *Causal discovery process:* When calculating the SemCaDo utilities, we were also forced to add a "fake" term (GO root) as a parent of the three existing root nodes in the GO (i.e., molecular function, cellular component and biological process) to perform semantic distance calculations on one unique ontology. This GO root will be then associated with a dozen of *S. cerevisiae* gene products which are not yet annotated with any GO term. The rest of the experimental process remains unchanged from that used in Section 4.1.
- *Ontology evolution:* Although, to make the experimental design more realistic in the context of biological resource management, we need to modify the third phase of our algorithm by updating the biological pathway datasets (e.g., DRYGIN repository) instead of making the GO enrichment. Metabolic pathways in such databases are computationally predicted using automated literature mining and then manually reviewed to ensure higher accuracy. This new dimension ensures optimal reuse of causal discoveries obtained from experimentations by submitting missing gene pairwise interactions. Unfortunately, since we are not intervening on a real system, we are unable to provide the dataset curators with any suggestions or corrections. We therefore content ourselves with a brief outline of the principle.

4.2.3. Results and interpretation

The present experimental study employed two key steps to evaluate our strategy, in both quantitative and qualitative way. In the first one, we evaluate the SemCaDo performance in terms of both the recovering of the expected structure and the total time required for execution. In the second context of analysis, we will proceed through a comprehensive comparison between the SemCaDo and MyCaDo algorithm. The goal of this experimental study is to assess the quality of the two algorithms, and, more importantly, to understand how theoretically predicted properties manifest themselves in a practical setting. To this end, we carry out a quantitative and qualitative comparison between the two algorithms.

- *Quantitative comparison between SemCaDo and MyCaDo (using multiple annotations):* Four scenarios have been analyzed, as shown in the two plots in Figs. 11 and 12.
  First, we apply the structure learning algorithm without considering any prior knowledge. Then we proceed through three series of tests with varying degrees of prior knowledge (20%, 40% and 60% of the edges contained in the initial

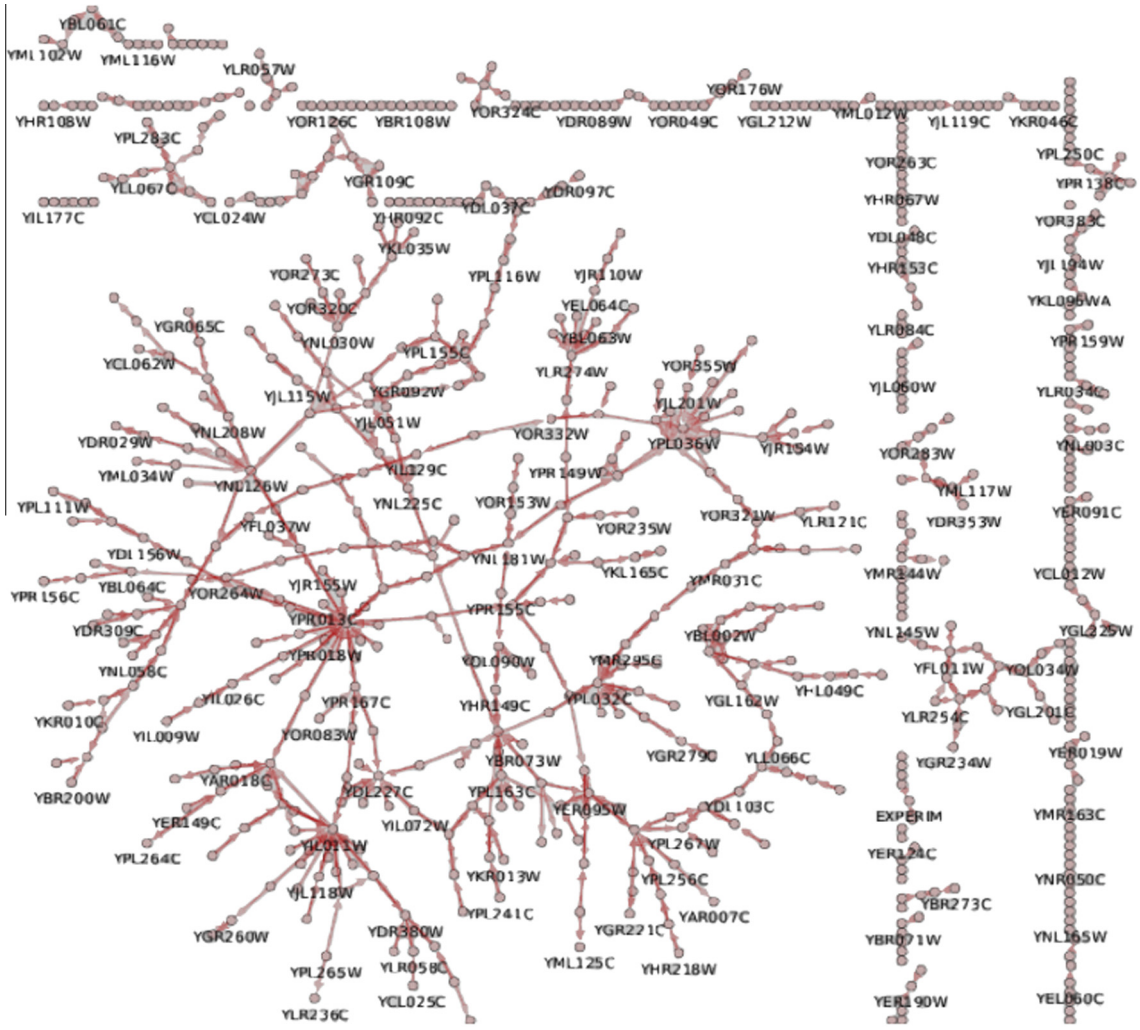

Fig. 10. Graphical representation of the entire GRN [24] employed for the experimentations.

graph). After each test run, we counted the number of correctly (resp. inversely) directed edges and non-existent edges obtained with SemCaDo (Refer to Fig. 11) and measure the time execution needed to reconstruct the full structure using GS (resp. MyCaDo and SemCaDo) algorithm (Refer to Fig. 12). In order to properly compare and contrast the empirical results, the above-mentioned algorithms must be applied using the same initial network and test conditions. Accordingly, it was so reasonable that the MyCaDo also integrates the same causal prior knowledge in a way to not penalize its performance when compared to SemCaDo.

We first set out to investigate the impact of injecting causal prior knowledge on the quality of the learning reconstruction. Fig. 11 illustrates the averaged results obtained with both SemCaDo (solid lines) and MyCaDo (dashed lines). We also plotted the proportion of edge orientation relative to applying MyCaDo without integrating priors in dashed horizontal line so that we could compare the SemCaDo results to those obtained with the original (resp. modified) version of MyCaDo.

As it could be verified by the empirical results, the edge direction accuracy in the first scenario is not sufficiently satisfactory for both MyCaDo and SemCaDo since the number of inversely and non-existent edges are still quite large. In the next three scenarios, with the introduction of causal priors, we realize a progressive improvement in the accuracy and efficiency of the learning process. For example, when proceeding with SemCaDo, the integration of 20% of the initial causal edges has adjusted the total number of correctly directed ones by about 10%. This leads to a relative decrease of both inversely and non-existent edges. We note that all the associated means and standard deviations are equals when dealing with MyCaDo and SemCaDo, indicating the close correlation between the studied curves. Thus, it appears that while MyCaDo is powered by causal priors, there was no significant improvement over the SemCaDo strategy.

Let us now reexamine the same four scenarios from the perspective of time execution. As we stated earlier in Section 3.2, the integration of the priors is restricted to the initial step of SemCaDo (resp. MyCaDo). This implies that GS is intended to be the first to benefit from this external guidance.

So, as shown in Fig. 12, the green line in the plot corresponds to the time execution of GS. The other curves in red and blue display the runtimes relative to both SemCaDo and MyCaDo. All runtimes are normalized and computed as the ratio of the actual time execution to the GS without priors runtime.

Fig. 11 shows the execution time relative to applying MyCaDo without integrating priors in dotted horizontal line. With

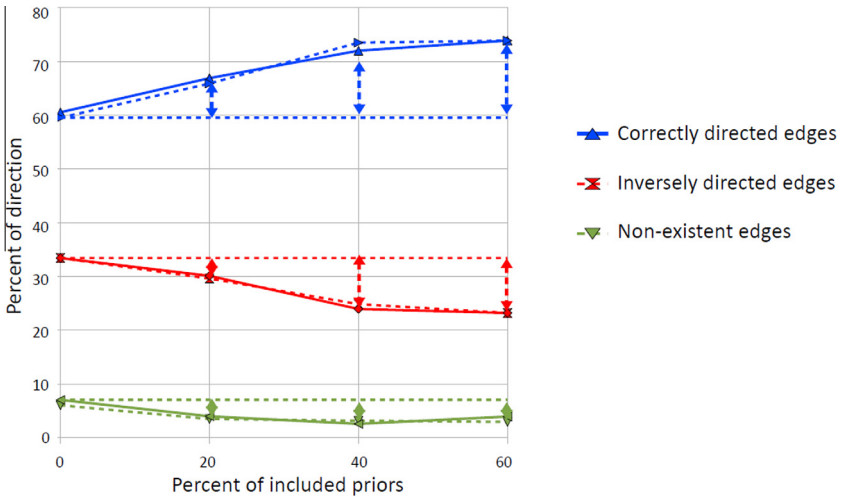

Fig. 11. Enhancing the structure learning performance of SemCaDo (resp. MyCaDo) by exploiting causal prior knowledge.

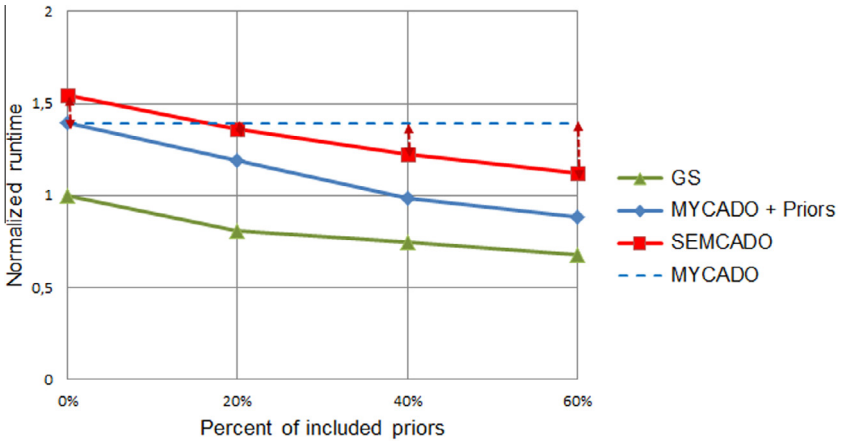

Fig. 12. Computational time for SemCaDo (resp. MyCaDo and GS) algorithms.

the actual parameter settings, SemCaDo requires the integration of 13% of causal prior knowledge to yield the same execution time as MyCaDo. This decreasing trend continues with the further introduction of causal priors. For example, with the introduction of 20% of causal priors, we improve the SemCaDo runtime by about 14%.

Similarly, we gain further improvement reducing overall execution time by about 24% (resp. 30%) when integrating 40% (resp. 60%) of causal priors.

Overall, the SemCaDo appears to be slightly more complex and time consuming compared to MyCaDo + priors. This is essentially due to the fact that SemCaDo returned each time to the ontology to perform the required distance calculations and determine which variable to alter next. However, further improvements can be achieved through the use of cache system performance.

- *Qualitative comparison between SemCaDo and MyCaDo (using multiple annotations):*

Given the above parameter settings, our analysis here will focus on a qualitative comparison between MyCaDo and SemCaDo.

The corresponding results are reported in Fig. 13 under the same four test conditions as in the previous analysis except that now we will display the evolution of the semantic cumulus along the experimental process for both MyCaDo (resp. MyCaDo + priors) and SemCaDo. As in Section 4.1, we also continue to assume the use of original GO and Rada distance. Table 1 can be used in conjunction with Fig. 13 to obtain additional numerical information relative to the gain in cumulus margin, the difference between the curves areas and the number of experiments that we saved when applying SemCaDo instead of MyCaDo.

For ease of interpretation, the curves analysis would be relatively more straightforward if we compare them in pairs. We thus need to shift the focus of our study in comparing SemCaDo to the original version of MyCaDo (solid curve) and then move to a comparison with the adapted version (dashed curve).

First of all, we apply both MyCaDo and SemCaDo without any prior knowledge (see Fig. 13A). The difference in areas between the two curves was about 7% and around one hundred experiments have been realized with the two algorithms.

When we integrate 20% of the initial causal relations before starting the learning process (Fig. 13B), we earn a cumulus margin of about 15% from the beginning. The difference in areas passes to 21% and we save nearly twenty experiments. This increasing trend continues when incorporating 40% of the initial causal relations (Fig. 13C) to obtain 31% as a cumulus margin, 30% as total difference in areas between the two

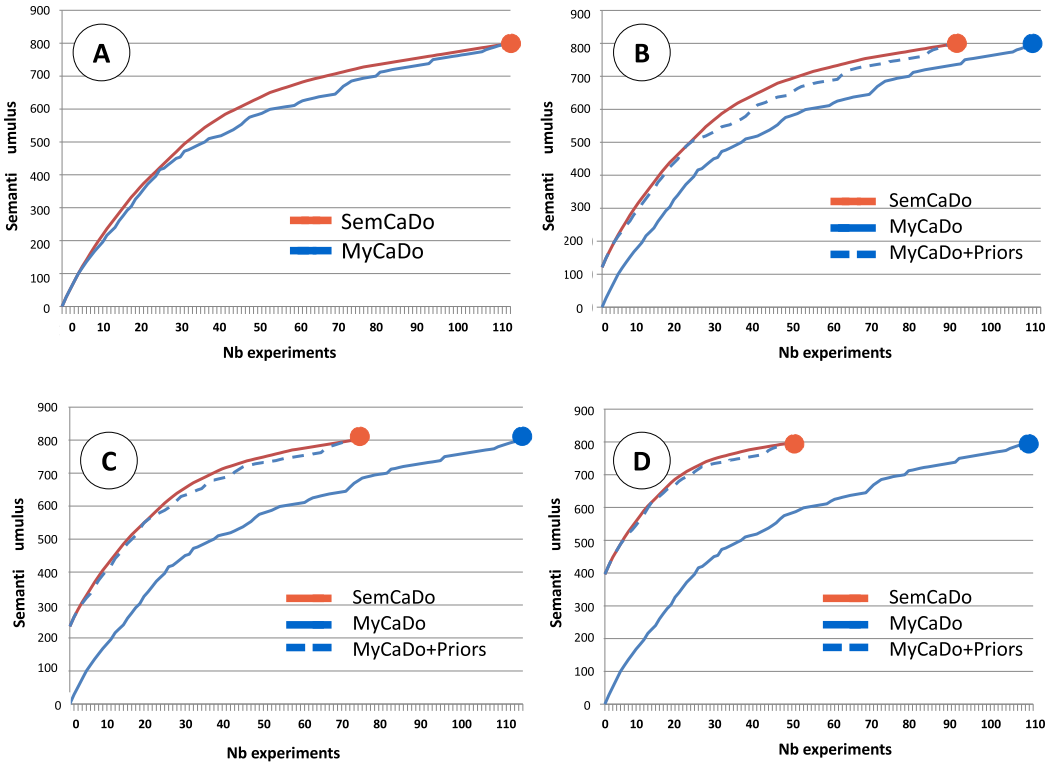

Fig. 13. Comparison between MyCaDo and SemCaDo (using multiple gene annotations) without any prior knowledge (A) and after integrating 20%, resp. 40% and 60% (B–D).

Table 1
Numerical comparison between SemCaDo and MyCaDo (resp. MyCaDo + priors) curves in Fig. 13.

	Causal integration (%)	Cumulus margin gain (%)	Difference between SemCaDo and MyCaDo curves areas (%)	Difference between SemCaDo and MyCaDo + priors curves areas	Saved experiments
Fig. 13A	0	0	7	–	0
Fig. 13B	20	15	21	3%	19
Fig. 13C	40	31	30	1%	35
Fig. 13D	60	50	37	0.5%	57

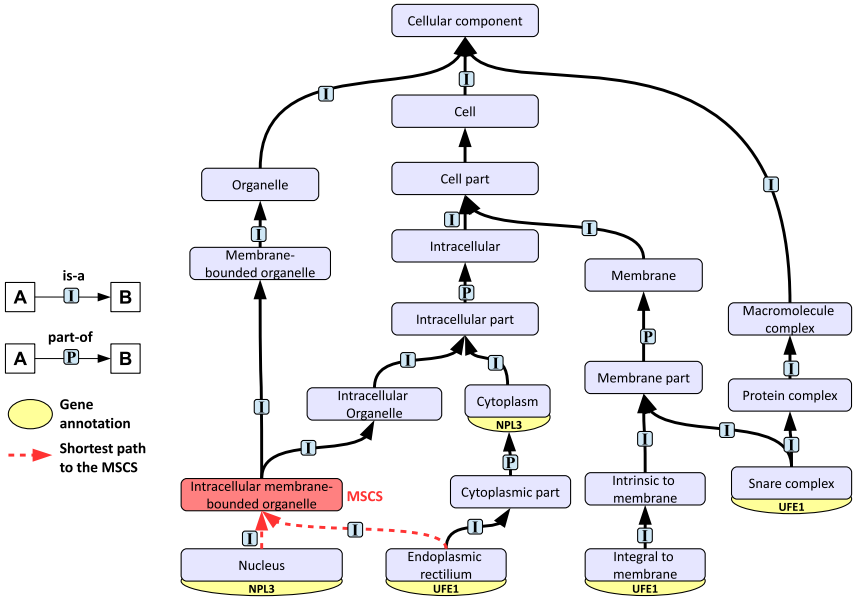

Fig. 14. The shortest path between two genes with multiple localizations in the GO.

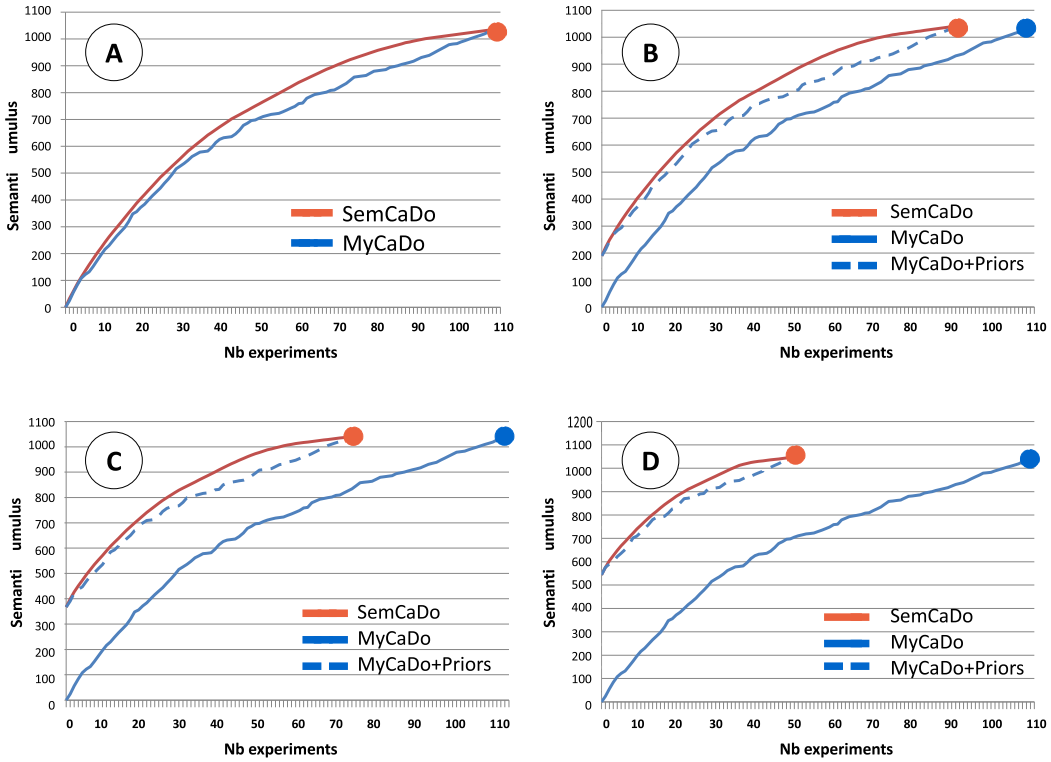

Fig. 15. Comparison between MyCaDo and SemCaDo (using single gene annotation) without any prior knowledge (A) and after integrating 20%, resp. 40% and 60% (B–D).

Table 2
Numerical comparison between SemCaDo and MyCaDo (resp. MyCaDo + priors) curves in Fig. 15.

	Causal integration (%)	Cumulus margin gain (%)	Difference between SemCaDo and MyCaDo curves areas (%)	Difference between SemCaDo and MyCaDo + priors curves areas	Saved experiments
Fig. 13A	0	0	10	–	0
Fig. 13B	20	18	29	8%	17
Fig. 13C	40	34	38	5%	34
Fig. 13D	60	51	47	3.5%	56

curves and 35 less experiments. We finish with the integration of 60% of the initial causal relations (Fig. 13D) to reach a cumulus margin of about 50%, a total difference in areas between the two curves exceeding the 37% and save more than 50% of unnecessary experiments.

- *Qualitative comparison between SemCaDo and MyCaDo (using single annotation):* Nevertheless, there is no obvious disparity among the curve of SemCaDo and that of MyCaDo + priors. At this point, it seems pertinent to discuss the effects of using genes showing multiple localizations in the GO. In Fig. 14, we utilize the same example as in Fig. 7 except that we now consider multiple annotations for NPL3 and UFE1.
  Our main conclusion is that MSCS localization (i.e., Intracellular membrane-bounded Organelle) has completely changed, and the shortest path to the MSCS has decreased from 6 to 2. Consequently, this seeming lack of semantically highly weighted edges has significantly reduced the performance of SemCaDo to nearly the same level as MyCaDo + priors.
  We then follow exactly the same experimental setup as in the previous test experimentation with original GO. Fig. 15 and Table 2 illustrate the obtained results with both SemCaDo and MyCaDo when using biased GO. Interestingly, the difference between the semantic cumulus curves becomes more significant. In fact, under the four test conditions, the difference between the SemCaDo and MyCaDo (resp. SemCaDo and MyCaDo + priors) curves areas is increased by about 8% (resp. 5%) on average compared to the previous tests. However, the cumulus margin gain and the number of saved experiments remain almost unchanged in spite of the significant increase in the semantic cumulus obtained after each experimentation.

In this way, we experimentally proved that SemCaDo gives better results than MyCaDo + priors when the domain ontology can provide discriminative information about serendipitous relationships. SemCaDo and MyCaDo + priors give similar results when the ontology cannot provide such information.

5. Conclusion and future works

In this paper, we outlined our serendipitous and active learning approach [7] which aims to (i) integrate the causal prior knowledge contained in the corresponding ontology when learning the initial structure from observational data, (ii) use a semantic distance calculus to guide the iterative causal discovery process to the more surprising relationships and (iii) capture the required causal discoveries to be applied to ontology evolution.

The proposed framework has several advantages over existing experimental design techniques. First, the idea of reusing ontological components can help to tackle real world learning problems. So, instead of repeating the efforts that have already been spent elsewhere to capture and create the same causal knowledge, one may reuse an existing domain ontology or some parts of it and make a considerable saving in term of time and cost. With such approach, we can also increase the reliability of the domain ontology by giving indication that it is continuously revised and evaluated through our ontology evolution process. Moreover, the serendipitous aspect when choosing the experimentations to perform is another advantage of the proposed strategy. This allows us to discover virgin areas and move away from what it is usually proposed by the research community.

While SemCaDo has the potential to intelligently choose experiments to perform, it is important that the reader be made aware of the limitation of the methodology. Indeed, until now, we have relied on a stringent set of assumptions when designing our active learning approach. Unfortunately, some of them are not always realistic and verifiable. So, at the next stage, we absolutely have to relax some of these SemCaDo's restrictions.

Among the numerous perspectives, an important issue concerns the best strategy to adopt in order to make better interactions with the ontology axioms during the causal discovery process. With such an extension, it will be possible to infer real causal structures directly from the ontology. We also propose to improve the third step of the SemCaDo algorithm by rejecting the ontology consistency and admitting axiom violations. In the third stage, the concept-node correspondence can be reformulated in a more tractable way. More precisely, there may be multiple correspondences between the node properties in the causal model and the concept attributes in the ontology. In such expanded settings, it seems most relevant to use probabilistic relational models [25] or object-oriented Bayesian networks [51] in which it is easy to define probability distribution over the attributes of the objects in the model. A last point concerning our perspectives is to proceed without assuming causal sufficiency. Recall that we still eliminate a number of latent variables that can be part of the model to pick up. Those hidden variables can be of particular relevance to establish useful models for achieving a correct causal inference and predicting the effects of some external criteria. In that case, we need a richer formalism than DAGs. The two major paradigms that can be used to model these systems are ancestral graphs [53] and semi-Markovian causal models [50]. This perspective gives us more points of comparison with MyCaDo++ algorithm [45].

Finally, SemCaDo can be custom-designed to explicitly pinpoint causality when disposing of a real system to intervene upon. Potentially fruitful areas that can adopt a similar serendipitous experimental design include chemistry, physics, ecology, marketing, health care and traffic.

References

[1] S. Albagli, R. Ben-Eliyahu-Zohary, S. Shimony, Markov network based ontology matching, J. Comput. Syst. Sci. 78 (1) (2012) 105–118.

[2] P. Anantharam, K. Thirunarayan, A. Sheth, Traffic analytics using probabilistic graphical models enhanced with knowledge bases, in: The 2nd International Workshop on Analytics for Cyber-Physical Systems (ACS-2013) at SIAM International Conference on Data Mining (SDM13), 2013, pp. 13–20.

[3] S.A. Andersson, D. Madigan, M.D. Perlman, A characterization of Markov equivalence classes for acyclic digraphs, Ann. Stat. 25 (2) (1997) 505–541.

[4] G.D. Bader, M.P. Cary, C. Sander, Pathguide: a pathway resource list, Nucleic Acids Res. 34 (Database-Issue) (2006) 504–506.

[5] M. Ben Ishak, P. Leray, N. Ben Amor, Ontology-based generation of object oriented bayesian networks, in: A. In Nicholson (Ed.), Proceedings of the Eighth UAI Bayesian Modeling Applications Workshop (UAI-AW 2011), Barcelona, Spain, vol. 818, 2011, pp. 9–17.

[6] M. Ben Ishak, P. Leray, N. Ben Amor, A two-way approach for probabilistic graphical models structure learning and ontology enrichment, in: Proceedings of the International Conference on Knowledge Engineering and Ontology Development (KEOD 2011) part of the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management IC3K, Paris, France, 2011, pp. 189–194.

[7] M. Ben Messaoud, Semcado: An Approach for Serendipitous Causal Discovery and Ontology Evolution, Ph.D. Thesis, University of Tunis & University of Nantes, 2012.

[8] M. Ben Messaoud, P. Leray, N. Ben Amor, Integrating ontological knowledge for iterative causal discovery and visualization, in: ECSQARU'09, Verona, Italia, 2009, pp. 168–179.

[9] M. Ben Messaoud, P. Leray, N. Ben Amor, Semcado: a serendipitous causal discovery algorithm for ontology evolution, in: The IJCAI-11 Workshop on Automated Reasoning about Context and Ontology Evolution (ARCOE-11), Barcelona, Spain, 2011, pp. 43–47.

[10] M. Ben Messaoud, P. Leray, N. Ben Amor, Semcado: a serendipitous strategy for learning causal bayesian networks using ontologies, in: ECSQARU'11, Belfast, Northern Ireland, 2011, pp. 182–193.

[11] M. Ben Messaoud, P. Leray, N. Ben Amor, Active learning of causal bayesian networks using ontologies: a case study, in: The International Joint Conference on Neural Networks (IJCNN-13), Dallas, US, 2013, pp. 1–8.

[12] E. Blanchard, M. Harzallah, H. Briand, P. Kuntz, A typology of ontology-based semantic measures., in: 2nd INTEROP-EMOI Open Workshop on Enterprise Models and Ontologies for Interoperability at the 17th Conference on Advanced Information Systems Engineering (CAISE'05), vol. 160, 2005, pp. 407–412.

[13] L.M.D. Campos, J.G. Castellano, Bayesian network learning algorithms using structural restrictions, Int. J. Approx. Reason. (2007) 233–254.

[14] D. Chickering, Learning Bayesian networks is NP-complete, Learning Data: Artif. Intell. Statist. V (1996) 121–130.

[15] D.M. Chickering, Learning equivalence classes of bayesian-network structures, J. Mach. Learn. Res. 2 (2002) 445–498.

[16] D.M. Chickering, Optimal structure identification with greedy search, J. Mach. Learn. Res. 3 (2002) 507–554.

[17] A.M. Cohen, W.R. Hersh, A survey of current work in biomedical text mining, Briefings Bioinformatics 6 (1) (2005) 57–71.

[18] G.F. Cooper, E. Herskovits, A bayesian method for the induction of probabilistic networks from data, Mach. Learn. 9 (1992) 309–347.

[19] A. Devitt, B. Danev, K. Matusikova, Constructing bayesian networks automatically using ontologies, in: Second Workshop on Formal Ontologies Meet Industry. FOMI '06, Trento, Italy, 2006.

[20] Z. Ding, Y. Peng, A probabilistic extension to ontology language OWL, in: Proceedings of the 37th Hawaii International Conference on System Sciences (HICSS '04), 2004.

[21] F. Eberhardt, C. Glymour, R. Scheines, N – 1 Experiments Suffice to Determine the Causal Relations Among N Variables, Department of Philosophy, Carnegie Mellon University, Technical Report CMU-PHIL-161.

[22] G. Flouris, D. Manakanatas, H. Kondylakis, D. Plexousakis, G. Antoniou, Ontology change: classification and survey, in: The Knowledge Engineering Review, vol. 23, 2008, pp. 117–152.

[23] N. Friedman, M. Goldszmidt, Learning bayesian networks with local structure, in: E. Horvitz, F.V. Jensen (Eds.), Morgan Kaufman, UAI, 1996, pp. 252–262.

[24] N. Friedman, M. Linial, I. Nachman, Using bayesian networks to analyze expression data, J. Comput. Biol. 7 (2000) 601–620.

[25] L. Getoor, B. Taskar, Introduction to Statistical Relational Learning, The MIT Press, 2007.

[26] C. Glymour, G. Cooper, Computation, Causation and Discovery, MIT Press, Cambridge, MA, 1999.

[27] T.R. Gruber, Towards principles for the design of ontologies used for knowledge sharing, Int. J. Human–Comput. Stud. 43 (5–6) (1995) 907–928.

[28] Y.B. He, Z. Geng, Active learning of causal networks with intervention experiments and optimal designs, JMLR 9 (2008) 2523–2547.

[29] F. Jensen, An Introduction to Bayesian Networks, Springer Verlag, New York, 1996.

[30] B. Jeon, I. Ko, Ontology-based semi-automatic construction of bayesian network models for diagnosing diseases in e-health applications, in: FBIT, IEEE Comput. Soc., 2007, pp. 595–602.

[31] A.M. Khattak, K. Latif, S. Lee, Y.K. Lee, Ontology evolution: a survey and future challenges, in: U- and E-Service, Science and Technology, vol. 62, 2009, pp. 68–75.

[32] J.L.Y. Koh, H. Ding, M. Costanzo, A. Baryshnikova, K. Toufighi, G.D. Bader, C.L. Myers, B.J. Andrews, C. Boone, Drygin: a database of quantitative genetic interaction networks in yeast, Nucleic Acids Res. 38 (Database-Issue) (2010) 502–507.

[33] D. Koller, N. Friedman, Probabilistic Graphical Models – Principles and Ttechniques, MIT Press, 2009.

[34] O. Lassila, R. Swick, Resource Description Framework (RDF) Model and Syntax Specification. WWW Consortium, 1998.

[35] S.L. Lauritzen, Graphical Models, Oxford University Press, Oxford, UK, 1996.

[36] S.L. Lauritzen, D.J. Spiegelhalter, Local computations with probabilities on graphical structures and their application to expert systems, J. R. Statist. Soc. Ser. B (Meth.) 50 (2) (1988) 157–224.

[37] A. Maedche, B. Motik, L. Stojanovic, N. Stojanovic, in: User-Driven Ontology Evolution Management, Springer-Verlag, 2002, pp. 285–300.

[38] S. Mani, G. Cooper, Causal discovery using a bayesian local causal discovery algorithm, Proc. MedInfo (2004) 731–735.
[39] K. McGarry, S. Garfield, N. Morris, S. Wermter, Integration of hybrid bio-ontologies using bayesian networks for knowledge discovery, in: A.S. d'Avila Garcez, P. Hitzler, G. Tamburrini (Eds.), NeSy, CEUR Workshop Proceedings, CEUR-WS.org, vol. 230, 2007.
[40] D. McGuinness, F.V. Harmelen, OWL Web Ontology Language overview, W3C Recommendation 10 (2004) 1–19.
[41] C. Meek, Causal inference and causal explanation with background knowledge, in: Proceedings of the Eleventh Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-95), Morgan Kaufmann, San Francisco, CA, 1995, pp. 403–410.
[42] C. Meek, Causal inference and causal explanation with background knowledge, in: P. Besnard, S. Hanks (Eds.), Morgan Kaufmann, UAI, 1995, pp. 403–410.
[43] S. Meganck, Towards an Integral Approach for Modelling Causality, Ph.D. Thesis, INSA Rouen and Vrije Universiteit Brussels, 2008.
[44] S. Meganck, P. Leray, B. Manderick, Learning causal bayesian networks from observations and experiments: a decision theoretic approach, in: V. Torra, Y. Narukawa, A. Valls, J. Domingo-Ferrer (Eds.), MDAI, Lecture Notes in Computer Science, vol. 3885, Springer, 2006, pp. 58–69.
[45] S. Meganck, S. Maes, P. Leray, B. Manderick, Learning semi-markovian causal models using experiments, in: M. Studen, J. Vomlel, (Eds.), Probabilistic Graphical Models, 2006, pp. 195–206.
[46] K.P. Murphy, Active Learning of Causal Bayes Net Structure, Tech. Rep., University of California, Berkeley, USA, 2001.
[47] J. Pearl, Fusion, propagation, and structuring in belief networks, Artif. Intell. 29 (3) (1986) 241–288.
[48] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.
[49] J. Pearl, Graphical models, causality and intervention, in: Statistical Science, vol. 8, 1993, pp. 266–269.
[50] J. Pearl, Causality: Models, Reasoning and Inference, Cambridge University Press, 2000.
[51] A. Pfeffer, Probabilistic Reasoning for Complex Systems, PhD Thesis, Stanford, 2000.
[52] R. Rada, H. Mili, E. Bicknell, M. Blettner, Development and application of a metric on semantic nets, IEEE Trans. Syst. Man Cybern. 19 (1) (1989) 17–30.
[53] T. Richardson, P. Spirtes, Ancestral graph Markov models, Ann. Stat. 30 (4) (2002) 962–1030.
[54] D. Settas, A. Cerone, S.Fenz, Generation of bayesian networks using the antipattern ontology, in: 9th ACIS International Conference on Software Engineering Research, Management and Applications (SERA'11).
[55] S. Shimizu, P.O. Hoyer, A. Hyvrinen, A.J. Kerminen, A linear non-gaussian acyclic model for causal discovery, J. Mach. Learn. Res. 7 (2006) 2003–2030.
[56] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, B. Futcher, Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization, Mol. Biol. Cell 9 (12) (1998) 3273–3297.
[57] P. Spirtes, C. Glymour, R. Schienes, Causation Prediction and Search, Springer-Verlag, New York, 1993.
[58] S. Spirtes, G. Glymour, R. Scheines, Causation, Prediction and Search, MIT Press, 2001.
[59] L. Stojanovic, N. Stojanovic, S. Handschuh, Evolution of the metadata in the ontology-based knowledge management systems, in: M. Minor, S. Staab (Eds.), German Workshop on Experience Management, vol. 10, 2002, pp. 65–77.
[60] M. Tagliasacchi, M. Masseroli, Anomaly-free prediction of gene ontology annotations using bayesian networks, in: J.J.P. Tsai, P.C.-Y. Sheu, H.C.W. Hsiao (Eds.), BIBE, IEEE Computer Society, 2009, pp. 107–114.
[61] T. Verma, J. Pearl, Equivalence and synthesis of causal models, in: P.P. Bonissone, M. Henrion, L.N. Kanal, J.F. Lemmer, (Eds.), UAI, 1990, pp. 255–270.
[62] F. Wu, D.S. Weld, Automatically refining the wikipedia infobox ontology, in: WWW '08: Proceeding of the 17th international conference on World Wide Web, ACM, New York, NY, USA, 2008, pp. 635–644.
[63] D. Xuan, L. Bellatreche, G. Pierra, A versioning management model for ontology-based data warehouses, in: A.M. Tjoa, J. Trujillo (Eds.), DaWaK, Lecture Notes in Computer Science, vol. 40-81, Springer, 2006, pp. 195–206.
[64] Y. Yang, J. Calmet, Ontobayes: an ontology-driven uncertainty model, International Conference on Computational Intelligence for Modelling, Control and Automation (CIMCA) (2005) 457–463.
[65] Y. Yang, J. Calmet, From the ontobayes model to a service oriented decision support system, in: Proceedings of International Conference on Computational Intelligence for Modelling, Control and Automation (CIMCA'06), 2006, pp. 127–127.
[66] S. Zhang, J. Cao, Y.M. Kong, R.H. Scheuermann, Go-bayes: gene ontology-based overrepresentation analysis using a bayesian approach, Bioinformatics 26 (7) (2010) 905–911.
[67] S. Zhang, Y. Peng, X. Wang, BayesOWL: a prototype system for uncertainty in semantic web, in: Proceedings of the International Conference on Artificial Intelligence, 2009, pp. 678–684.
[68] H. Zheng, B. Kang, H. Kim, An ontology-based bayesian network approach for representing uncertainty in clinical practice guidelines, in: F. Bobillo, P.C.G. da Costa, C. d'Amato, N. Fanizzi, F. Fung, T. Lukasiewicz, T. Martin, M. Nickles, Y. Peng, M. Pool, P. Smrz, P. Vojts, (Eds.), URSW, vol. 327, 2007.