Serendipitous strategy for learning causal bayesian networks using ontologies: a case study

MONTASSAR BEN MESSAOUD
LARODEC, Institut Supérieur de Gestion Tunis
LINA, Laboratoire d'Informatique de Nantes Atlantique
benmessaoud.montassar@hotmail.fr

PHILIPPE LERAY
LINA, UMR 6241, France
Ecole Polytechnique de l'Univ. de Nantes
philippe.leray@univ-nantes.fr

NAHLA BEN AMOR
LARODEC, Institut Supérieur de Gestion Tunis
41, Avenue de la liberté, 2000 Le Bardo, Tunisie
nahla.benamor@gmx.fr

***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 spread our previous works which foster greater collaboration between causal discovery and ontology evolution so as to evaluate them on real case study.**

I. INTRODUCTION

A directed acyclic graph (DAG) (also called a Bayesian network (BN)) is a powerful tool for representing domains with inherent uncertainty. Due to the Markov equivalence property, when learning the Completed Partially Directed Acyclic Graph (CPDAG) from observational data and randomly choosing one possible complete instantiation in the equivalence space, we are left with an unresolved causal structure. In order to provide a causal interpretation for BNs, an extension, called Causal Bayesian networks (CBNs), is introduced with the goal to provide a convenient framework for causal modeling and reasoning.

Contrary to the non-Gaussian learning methods (also called LiNGAM) which use pure observational data (D_{obs}), the causal discovery in CBNs often requires interventional data (D_{int}). In this work, we don't 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 approach to only CBNs.

This paper provides a continuation as well as an extension of our previous works [1][2][3][4] in which we introduce an algorithm that actively chooses the experiments to perform based on the semantic distance calculus. Further developments along this direction have been made in order to design the experimental work appropriately.

The remainder of this paper is arranged as follows: Section 2 gives the necessary background for both CBNs and ontologies. In Section 3, we discuss some related works that have addressed the problem of active causal discovery and in Section 4 we present a case study to demonstrate the practical application of our active learning design in real biological system. Concluding remarks and future works are given in Section 5.

II. BASIC CONCEPTS & BACKGROUND

A. Causal Bayesian Networks

A Causal Bayesian network (CBN), also known as a Markovian model, is represented by a tuple (*G,P*), where: (i) *G* is a DAG, called a causal graph, over a set X=$\{X_1, X_2,.., X_n\}$ of vertices, and (ii) a probability distribution *P*(v), over the set X of discrete variables that correspond to the vertices in *G*.

The probabilistic interpretation views G as representing conditional independence restrictions on P: Each variable is independent of all its non-descendants given its direct parents in the graph. This leads to express a global factorization of the joint probability distribution (JPD) over the set of random variables in the graph.

In addition to the usual conditional independence interpretation, the CBN is also given a causal interpretation since each arc is identified as a direct causal influence between a parent and a child node. For this reason, CBNs are considered as proper bayesian networks (BNs) but the reverse is not necessarily true.

The main difference between the two formalisms lies in the nature of the data needed to learn the structure. Due to nonidentifiability, when building our model from pure observational data, we may not have enough information to discover the true structure of the graph and the causal model will be restricted to the Completed Partially Directed Acyclic Graph (CPDAG).

However, if we are able to intervene on the system, that is, to set some of its variables to user-specified values, we can infer the directions of arcs which are not specified in the particular Markov equivalence class. At this point, we should note that intervening on a system may be very expensive, time-consuming or even impossible to perform. This implies that the choice of variables to experiment on can be vital when the number of interventions is restricted.

B. Ontologies

There are different definitions in the literature of what should be an ontology. The most notorious was given by Tom Gruber [21], stipulating that an ontology is an *explicit* specification of a *conceptualization*. The "conceptualization", here, refers to an abstract model of some phenomenon having real by identifying its relevant concepts. The word "explicit" means that all concepts used and the constraints on their use are explicitly defined.

In the simplest case, an ontology describes a hierarchy of concepts (i.e. classes) related by taxonomic relationships (is-a, part-of). In more sophisticated cases, an ontology describes domain classes, properties (or attributes) for each class, class instances (or individuals) and also the relationships that hold between class instances. It is also possible to add some logical axioms to constrain concept interpretation and express complex relationships between concepts.

Hence, more formally, an ontology can be defined as a set of labeled classes C=$\{C_1, ..., C_n\}$, hierarchically ordered by the subclass relations (i.e. is-a, part-of relations). For each concept C_i we identify k meaningful properties p_j, where j $\in$ [1, k]. We use H_i to denote the finite domain of instance (i.e. concretizing the ontology concepts by setting their properties values) candidates with each concept C_i and c_i to denote any instance of C_i. We also use R to represent the set of semantical (i.e non-hierarchical) relations between concepts and R_c to represent the subset of causal ones. Finally, formal axioms or structural assertions $<c_i, c_j, s>$ can be included, where $s \in S$ is a constraint-relationship like "must, must not, should, should not, etc".

Practically speaking, the ontologies are often a very large and complex structure, requiring a great deal of effort and expertise to maintain and upgrade the existing knowledge. Such proposals can take several different forms such as a change in the domain, the diffusion of new discoveries or just an information received by some external source [26].

There are many ways to change the 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 [27].

The ontology evolution can be of two types [27]:

- 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 a given ontology.

In order to establish the context in which the ontology evolution takes place, the principle of ontology continuity should be fulfilled [28]. It supposes that the ontology evolution should not make false an axiom that was previously true. When changes do not fulfill the requirement of ontological continuity, it is not any more an evolution, it is rather an ontology revolution.

III. Active causal discovery: State of the art

The main purpose of causal discovery with active learning is to derive meaningful patterns from a limited data. Fundamentally, active learning methods are designed to guide the learning process through the most informative interventions and maximally reduce the variance in the model. Several approaches for actively learning CBNs have been proposed during the last decade.

[17] studied a Bayesian scoring metric that can incorporate both observational and experimental data. Using a similar metric [13] designed an algorithm to select experiments that minimizes an expected loss. A similar but more general algorithm has been proposed in [22] to learn BN structures using posterior distributions of structures based on decision theory. [24] investigates the performance of another active learning approach that is more suitable for modeling continuous data. The method employs an expected loss function that should be expressed in terms of the size of transition sequence equivalence classes [25].

In some recent works, [23] presents a theoretic approach in which the causal discovery is considered as a two person game between Nature and Scientist. The scientist attempts to discover the true causal structure and Nature tries to make discovery as difficult as possible. [5][7] have also proposed the MyCaDo (My Causal Discovery) algorithm for learning CBNs from perfect observational data and experiments. Using traditional structure learning techniques, they learn CPDAG from observational data and then try to discover the directions of the remaining edges by means of experiments. To choose the best ones, they proposed a utility function reflecting the gain (i.e. the number of undirected edges and those susceptible to be inferred) and costs of the experiments. MyCaDo forms a key component in our actual contribution, named Semantic Causal Discovery (SemCaDo), in which we incorporate available knowledge from domain ontologies. The original character of SemCaDo algorithm is essentially its ability to discover and reuse the capitalized knowledge in CBNs. As inputs, SemCaDo needs an observational dataset and a corresponding domain ontology. Then it will proceed through three consecutive phases:

- During the first, we will try to fully exploit the semantic causal relations encoded in the ontology by injecting them in the structure learning process. 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.
- The second phase will then tackle the problem of consistently orienting the rest of undirected edges in the CPDAG. It seems obvious that the gained information of the utility function employed in MyCaDo is essentially the node connectivity which serves to orient the maximal number of edges but not necessary the most informative ones. To cope with this limitation, the strategy we propose in our approach makes use of a semantic distance calculus 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. By this way, we will accentuate the serendipitous aspect of the proposed strategy and investigate new and unexpected causal relations on the graph.

- In the third step, we follow the same edge orientation strategy as in MyCaDo. 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. The causal discoveries will be then extracted and interpreted for an eventual ontology evolution. In this way, the causal relations will be traduced as semantic causal relations between the corresponding ontology concepts.

In previous works [2][3], we resort to simulations to give accurate results on the performance of SemCaDo compared to MyCaDo in various situations (i.e. MyCaDo serves as comparison reference to SemCaDo since both of them share the same assumptions and use the same input data). In the next section, we present an experimental evaluation using real cellular network to confirm the effectiveness of the proposed technique.

IV. Validation on S. 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 [6] and the lack of facilities for conducting genetic [1] (resp. environmental [2]) 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 Sacharomyces cerevisiae cell cycle microarray data and the corresponding Gene Ontology annotations.

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 Saccharomyces cerevisiae cell cycle microarray data 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 data set contains 73 sampling points for all genes. Each of them is measured in different phases of the yeast cell cycle. According to [9], about 800 of these genes have been reported with varying transcripts over the cell cycle stages.

[1]Gene knockout (deletion of the gene), or overexpression (setting the expression level higher than its usual level).

[2]change in one or more non-genetic factors, such as a change in environment, nutrition, pressure or temperature.

- ◇ *Gene ontology:* Most of the Saccharomyces cerevisiae genes are annotated with specific biological functions from the Gene Ontology (GO) [29], 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.

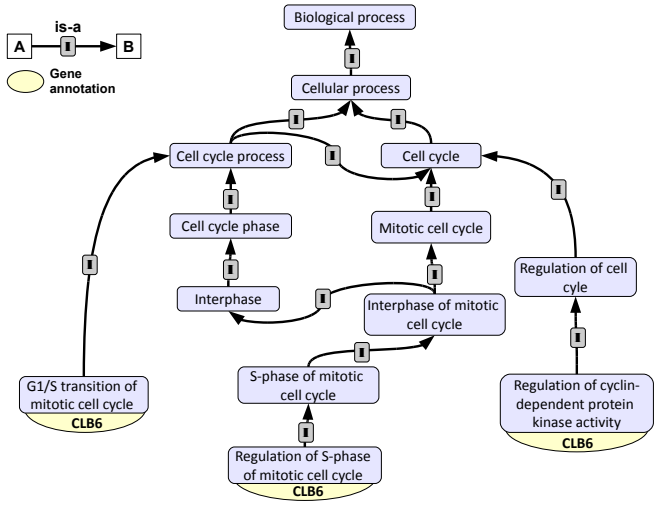

Fig. 1. CLB6 multiple localizations in GO

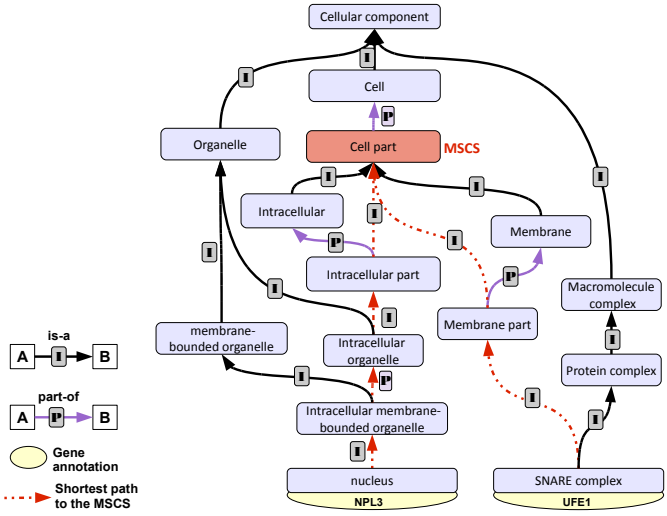

Fig. 2. Semantic distance between two annotated genes in GO.

The directed edges between concept nodes represent either subsumption links ("is-a") or composition relationships ("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, at the most specific levels possible. For instance, as shown in Figure 1, the gene CLB6 is involved in:

- the regulation of cyclin-dependent protein kinase activity (GO:0000079),
- the regulation of S phase of mitotic cell cycle (GO:0007090),
- the 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 Figure 2 the multiple paths that can be found between them. Using our simple path based method, we set the cell part term (colored in red) as the most specific common subsumer (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 way to identify causal relations is to use the so-called 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) [3], enable a convenient access to genes in terms of the biological pathways in which they intervene (See the DRYGIN screen capture for the top regulatory pathways involving the gene CLB6 in Figure 3).

2) Experimental design: When applying our approach in the context of biological field, we were forced to change some of the initial CBN-ontology correspondences that we provide in [2][3]. According to Table I, the GRN nodes which correspond to genes will be assigned to the most specific level of the Gene Ontology using term annotations

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

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

(i.e. instances). Then there would no longer be any need to use the observational and experimental data since we dispose of an appropriate causal model based on we simulate experimental treatments. The rest of the correspondences remain unchanged.

TABLE I

THE SET OF ALL POSSIBLE CORRESPONDENCES BETWEEN THE GRN AND THE GENE ONTOLOGY.

Gene Regulatory Network	Gene Ontology
Nodes	Concept instances (i.e. GO annotations)
Causal dependencies	Semantic causal relations
Causal inference	Logic rule reasoning

To make a meaningful performance comparison between MyCaDo and SemCaDo algorithms, we will detail the three main blocs of our experimental strategy:

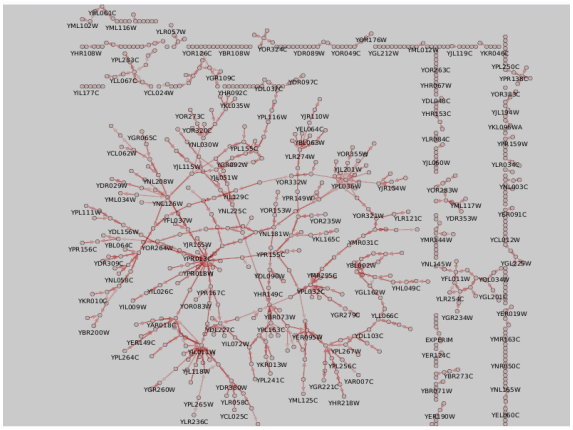

Fig. 4. Graphical representation of the entire GRN employed for the experimentations

- *Structure learning:* Our alternative way for implementing the MyCaDo (resp. SemCaDo) approaches is to use the Gene Regulatory Network (GRN) of [8] as a starting causal model and the GO structure as a source for calculating semantic distances between genes. From a modelling standpoint, a GRN can be thought 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).

Using the Yeast Sacharomyces cerevisiae cell cycle microarray data, [8] proved that they were able to extract a finer structure of regulatory interactions between genes. Their heuristic approach was 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 which assumes that a gene interaction exists between two genes if no variable in the model mediates the dependence between them. The second feature is synonymous of 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. Using the Tulip Software [30], a screen capture of the considered causal graph is shown in Figure 4.

The main reason for choosing the GRN of [8] as starting model is that it is free from assumptions and don't reuse any prior knowledge. We also note that interactions between genes other than causal relationships (i.e. directed edges with sufficiently high confidence in the order between genes) are not considered.

We initially need this causal graph to sample normally distributed data that will be fed to the Greedy search (GS) algorithm [18] 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 full posterior (BDe) score and is faster to compute with large amounts of data. 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 constructed, we need to execute the DAG-to-CPDAG algorithm [10].

- *Causal discovery process:*
  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 [5][7] and choose equal measures of importance when calculating the expected utilities (i.e. α=β=1).
  To perform the experimentation on the best node, we have 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 [19]. We force X_{best} to take on random values and then sample the post-intervention distribution to get our 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 experimented on is the cause or the effect of its neighboring variables.
  Another point to consider in our experimental study concerns the calculation of the semantic gain:

$$Semantic\_Gain(X_i) = \sum_{X_j \in O(X_i)} dist_{Sem}(mscs(O^*(X_i)), X_j^*) \tag{1}$$

  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 minimal distance, in separating edges, between C_i and C_j.
  Throughout our simulation phase, we measured the sum of Rada's distances [21] relative to the new directed edges in the graph and update a semantic cumulus after each SemCaDo (resp. MyCaDo) iteration.
  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.
- *Pathway repository 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.

3) Results & interpretation: The present experimental study employed a two key steps to evaluate our strategy, in both a 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 analy-

sis, 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.

- ⋆ *The beneficial effects of prior knowledge:* Four scenarios have been analyzed, as shown in the two plots in figures 5 and 6. 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 graph). After each test run, we counted the number of correctly (resp. inversely) directed edges and non-existent edges obtained with SemCaDo (Refer to Figure 5) and measure the time execution needed to reconstruct the full structure using GS (resp. MyCaDo and SemCaDo) algorithm (Refer to Figure 6). In order to properly compare and contrast the empirical results, the above-mentioned algorithms must be simulated on 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.

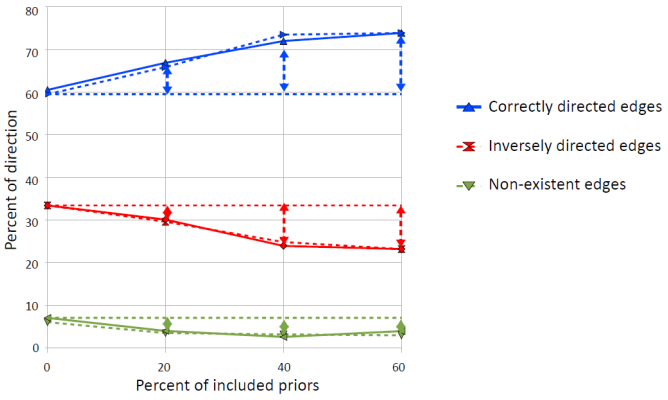

Fig. 5. Enhancing the structure learning performance of SemCaDo (resp. MyCaDo) by exploiting causal prior knowledge

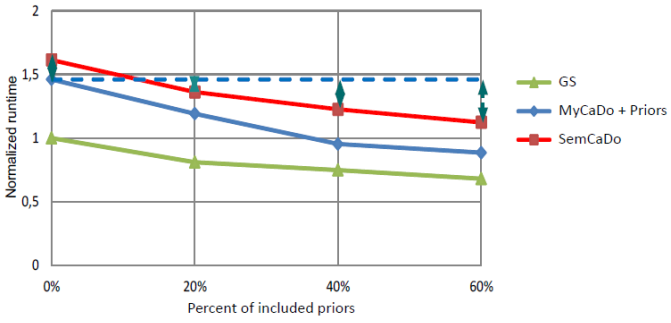

Fig. 6. Computational time for SemCaDo (resp. MyCaDo and GS) algorithms

We first set out to investigate the efficacy of studying the impact of causal prior knowledge on the quality of the learning reconstruction. Figure 5 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.

Let's now reexamine the same four scenarios from the perspective of time execution. As we stated earlier, 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 Figure 6, the green line in the plot corresponds to the time execution of GS. The other curves in red and blue displayed the runtimes relative to both SemCaDo and MyCaDo. For expository ease, we normalize the time execution of each algorithm by the GS (without priors) runtime, so the normalized runtimes will be values between 0 and 2. Here, without priors, we can depict that MyCaDo+priors needed three quarters of the time that the SemCaDo used. Similarly as in Figure 5, we plotted the execution time relative to applying MyCaDo without integrating priors in dotted horizontal line. Under the same parameter setting, 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 significantly outperforms MyCaDo for all the tested graphs but it appears to be 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.

⋆ *Qualitative comparison between SemCaDo and MyCaDo:*
Given the above parameter settings, our analysis here will focus on a qualitative comparison between MyCaDo and SemCaDo. The corresponding results are reported in Figure 7 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. Here we would like to propose a different approach whose aim is to promote the experimentation on the more distant genes according to the GO. We also continue to assume the use of Rada distance [21]. Table II can be used in conjunction with Figure 7 to obtain additional statistical 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.

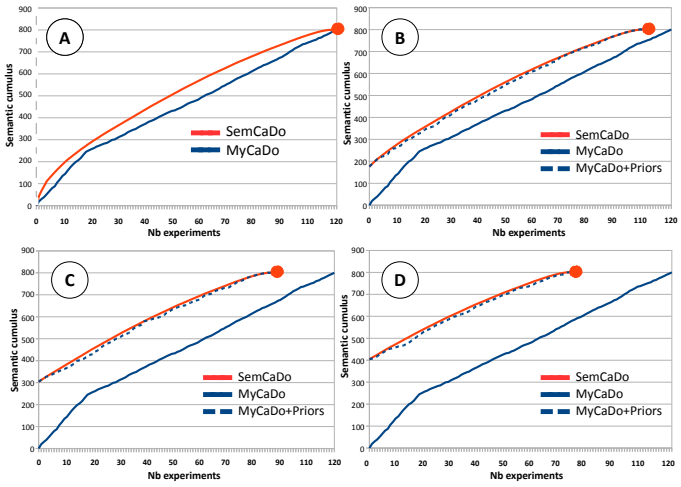

Fig. 7. Comparison between MyCaDo and SemCaDo without any prior knowledge (a) and after integrating 20 %, resp. 40% and 60% (b, c, d).

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 Figure 7.a). The difference in areas between the two curves was about 10% 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 (Figure 7.b) , we earned a cumulus margin of about 22% from the beginning. The difference in areas pass to 29% and we save nearly ten experiments. This increasing trend continues when incorporating 40% of the initial causal relations (Figure 7.c) to obtain 37% as a cumulus margin, 38% as total difference in areas between the two curves and 19 less experiments. We finish with the integration of 60% of the initial causal relations (Figure 7.d) to reach a cumulus margin of about 51%, a total difference in areas between the two curves exceeding the 44% and save more than 30 unnecessary experiments.

TABLE II
STATISTICAL ANALYSIS OF SEMCADO AND MYCADO CURVES IN FIGURE 7.

	Causal integration	Cumulus margin gain	Difference between SemCaDo and MyCaDo curves areas	Saved experiments
Fig 7.A	0%	0%	10%	0
Fig 7.B	20%	22%	29%	10
Fig 7.C	40%	38%	37%	19
Fig 7.D	60%	51%	44%	31

Nevertheless, there is no obvious disparity among the curve of SemCaDo and that of MyCaDO + priors since the two curves were highly promoted from the beginning. Additionally, it is noteworthy that regulatory relations between "semantically" adjacent genes are much more widespread in GRNs. It is also possible that some semantically highly weighted edges were introduced with the initial causal priors. But be that as it may, it is clear that the shape of the SemCaDo curve still gives the best performance. In fact, a lot of experiments and efforts have been saved compared to MyCaDo and the most informative interventions have been reported earlier in the experimental process. This allows a significant gain in term of relevant experimentations especially when there is not enough budget to cover all the required interventions. Our belief is that SemCaDo top-ranked genes can be targets for medical treatment of genetic diseases and opportunities to obtain further knowledge about the biological mechanisms that underly their gene expression. Potentially, this gives us scope to explore virgin areas when developing our knowledge-base on pathway modeling.

V. CONCLUSIONS

In this paper, we emphasize the potential application of the SemCaDo approach on real biological system design using S. Cerevisiae cell cycle microarray data and Gene Ontology. Based on the experimental results, we provided solid evidence that SemCaDo achieves better performance than MyCaDo, its competing algorithm. Our quantitative analysis outlined the SemCaDo ability to completely recover the true causal structure and reduce the learning task complexity. On the other hand, the qualitative analysis also showed the role of our active design to identify the most serendipitous experiments and move away from what it is usually proposed

by the research community. Nevertheless, the main problem is that there is no commonly accepted benchmark to help us to go further towards developing experimental tests that can lead to more rigorous results.

REFERENCES

[1] M. Ben Messaoud, P. Leray and N. Ben Amor, "Integrating Ontological Knowledge for Iterative Causal Discovery and Visualization", *ECSQARU'09*, pp. 168-179, Verona, Italia.

[2] M. Ben Messaoud, P. Leray and N. Ben Amor, "SemCaDo: a serendipitous strategy for learning causal bayesian networks using ontologies", *ECSQARU'11* , pp. 182-193, Belfast, Northern Ireland.

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

[4] M. Ben Messaoud, "SemCaDo: an approach for serendipitous causal discovery and ontology evolution", *Ph. D. thesis, University of Tunis & University of Nantes*, 2012.

[5] S. Meganck, "Towards an integral approach for modelling causality", *Ph. D. thesis, INSA & Rouen University*, 2008.

[6] C. Lovell, "Active learning for discovery in the laboratory: characterising biomolecular systems", *ALRA: Active Learning in Real-world Applications*, Bristol, GB, 2012.

[7] S. Meganck, P. Leray and B. Manderick, "Learning Causal Bayesian Networks from Observations and Experiments: A Decision Theoretic Approach", *In MDAI, Edited by V. Torra, Y. Narukawa, A. Valls and J. Domingo-Ferrer*, vol. 3885 of Lecture Notes in Computer Science, Springer, pp. 58-69.

[8] N. Friedman, M. Linial and I. Nachman, "Using Bayesian networks to analyze expression data", *Journal of Computational Biology*, vol. 7, pp. 601-620, 2000.

[9] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein and B. Futcher, "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization", *Molecular Biology of the Cell*, vol. 9, no. 12, pp. 3273-3297, 1998.

[10] D. M. Chickering, "Learning Equivalence Classes of Bayesian-Network Structures", *Journal of Machine Learning Research*, vol. 2, pp. 445-498, 2002.

[11] J. Pearl, "Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference", Morgan Kaufmann Publishers Inc, USA, 1988.

[12] J. Pearl, "Causality: models, reasoning and inference", Cambridge University Press, 2000.

[13] S. Tong and D. Koller, "Active Learning for Structure in Bayesian Networks", *International Joint Conference on Artificial Intelligence*, pp. 863-869, 2001.

[14] Y. He and Z. Geng, "Active learning of causal networks with intervention experiments and optimal designs", *JMLR*, vol. 9, pp. 2523-2547, 2008.

[15] G. Li and T. Leong, "Active Learning for Causal Bayesian Network Structure with Non-symmetrical Entropy", *PAKDD*, vol. 5476, pp. 290-301, Springer, 2009.

[16] A. Larjo, H. Lähdesmäki, M. Facciotti, N. Baliga, O. Yli-Harja and I. Shmulevich, "Active learning of Bayesian network structure in a realistic setting", *In Fifth International Workshop on Computational Systems Biology (WCSB)*, Leipzig, Germany, 2008.

[17] G. F. Cooper and C. Yoo, "Causal Discovery from a Mixture of Experimental and Observational Data", *In Kathryn B. Laskey and Henri Prade, editors, UAI*, pp 116-125, Morgan Kaufmann, 1999.

[18] N. Friedman and M. Goldszmidt, "Learning Bayesian Networks With Local Structure", *In Eric Horvitz and Finn Verner Jensen, editors, UAI*, pp 252-262, Morgan Kaufmann, 1996.

[19] J. Pearl, "Graphical Models, Causality And Intervention", *Statistical Science*, vol. 8, pp 266-269, Comment to Spiegelhalter et al., 1993.

[20] R. Rada, H. Mili, E. Bicknell and M. Blettner, "Development and application of a metric on semantic nets", *IEEE Transactions on Systems, Man and Cybernetics*, vol. 19, no. 1, pp. 17-30, 1989.

[21] T. R. Gruber, "Towards Principles for the Design of Ontologies Used for Knowledge Sharing", *International Journal Human-Computer Studies*, vol. 43, Issues 5-6, pp. 907-928, 1995.

[22] K. P. Murphy, "Active Learning of Causal Bayes Net Structure", *Technical report*, Department of Computer Science, UC Berkeley, 2001.

[23] F. Eberhardt, "Causal discovery as a game", *Journal of Machine Learning Research-Proceedings Track*, vol. 6, pp. 87-96, 2010.

[24] I. Pournara and L. Wernisch, "Reconstruction of gene networks using Bayesian learning and manipulation experiments", *Bioinformatics*, vol. 20, no. 17, pp. 29342942, 2004.

[25] J. Tian and J. Pearl, "Causal discovery from changes: A bayesian approach", *UCLA Cognitive Systems Laboratory, Technical Report (R-285)*, 2001.

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

[27] A. M. Khattak, K. Latif, S. Lee and Y. K. Lee, "Ontology Evolution: A Survey and Future Challenges", *In U- and E-Service, Science and Technology*, vol. 62, pp. 68-75, 2009.

[28] D. Xuan, L. Bellatreche and G. Pierra, "A Versioning Management Model for Ontology-Based Data Warehouses", *In A. Min Tjoa and Juan Trujillo, editors, DaWaK*, vol. 4081 of Lecture Notes in Computer Science, pp. 195-206, Springer, 2006.

[29] The Gene Ontology Consortium, "Gene Ontology: tool for the unification of biology", *Nature Genetics*, vol. 25, pp. 25-29, 2000.

[30] D. Auber, "Tulip: A huge graph visualization framework", *In Graph Drawing Softwares. Edited by P. Mutzel and M. Jünger, Mathematics and Visualization*, Springer-Verlag, pp. 105-126.