Using Qualitative Probability in Reverse-Engineering Gene Regulatory Networks

Zina M. Ibrahim, Alioune Ngom, and Ahmed Y. Tawfik

Abstract—This paper demonstrates the use of qualitative probabilistic networks (QPNs) to aid Dynamic Bayesian Networks (DBNs) in the process of learning the structure of *gene regulatory networks* from microarray gene expression data. We present a study which shows that QPNs define monotonic relations that are capable of identifying regulatory interactions in a manner that is less susceptible to the many sources of uncertainty that surround gene expression data. Moreover, we construct a model that maps the regulatory interactions of genetic networks to QPN constructs and show its capability in providing a set of candidate regulators for target genes, which is subsequently used to establish a prior structure that the DBN learning algorithm can use and which 1) distinguishes spurious correlations from true regulations, 2) enables the discovery of sets of coregulators of target genes, and 3) results in a more efficient construction of gene regulatory networks. The model is compared to the existing literature using the known gene regulatory interactions of *Drosophila Melanogaster*.

Index Terms—Gene regulatory networks, reverse-engineering genetic networks, dynamic Bayesian networks, qualitative probabilistic networks, qualitative reasoning.

1 Introduction

The past decade has witnessed the inception of computational problems concerned with making sense of the massive influx of biological data generated by microarray technology and extracting useful information and biological insight from the data.

As a result, a variety of mathematical techniques and computational models have been devised to solve the problems at hand. One such model which has generated a lot of research interest is concerned with finding means to uncover the complex true gene-to-gene interactions governing the gene expression data obtained from microarrays [23] and modeling the discovered connectivity through a network, called a *gene regulatory network* (GRN) [23]. The resulting computational problem, termed *reverse-engineering gene regulatory networks* [4] from gene expression profiles, is currently one of the central problems in Systems Biology as it can provide great insight to the internal working of the cell [27].

However, the nature of microarray gene expression data makes reconstructing such a network from the available data a difficult task. This is because microarray data are highly dimensional, describing the expression levels of as many as tens of thousands of genes for a small number of samples or at relatively few experimental conditions or time points [4]. Microarray data are also noisy, sparse and governed by imprecision [8], making it difficult to make informed judgement about the interactions of the genes and how they affect each other. Needless to say, this does not only make the convergence to a single network describing the data at hand a more challenging task, but questions arise on whether or not the learned model is representative of the true genetic interactions governing the observed data and if it can, in fact, be used to reach a functional understanding of the mechanisms underlying the synergies among the cellular genetic components [8].

The literature contains a variety of approaches to tackle the task at different levels of detail [27] such as differential equations [2], [25], Boolean networks [15], state-space models [11], clustering methods [6], neural networks [30], fuzzy systems [12], and Bayesian networks [18], [8]. While each of these models has advantages and pitfalls, the Bayesian approach has attracted special attention because of its inherent capability of capturing the stochastic nature and noise of microarray data [4], [8]. More specifically, *Dynamic Bayesian Networks* (DBNs), which extend Bayesian Networks to capture temporal information and cyclic relations, have been successfully applied to extract regulatory information from time-series microarray data [18], [31], [22] and learn large-scale networks. The body of work focusing on the DBN approach has also tackled issues such as the incorporation of time lags [29], [32], improving the convergence rates [18] as well as combining perturbation experiments with time-series experiments in order to improve the quality of the inferred network [5].

Along with these major efforts comes a great room for improvement as there are two fundamental issues that are being continuously studied with respect to DBNs. The first pertains to the accuracy of the learned model to maximize

- *Z.M. Ibrahim and A. Ngom are with the School of Computer Science, University of Windsor, 401, Sunset Avenue, Windsor, ON N9CB3P4, Canada. E-mail: {ibrahim, angom}@uwindsor.ca.*
- *A.Y. Tawfik is with the French University in Egypt, Shurooq city, Cairo Ismailia Desert Road, Egypt. E-mail: ahmed.tawfik@ufe.edu.eg.*

Manuscript received 3 May 2010; revised 7 July 2010; accepted 11 July 2010; published online 22 Sept. 2010.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBBSI-2010-05-0112.
Digital Object Identifier no. 10.1109/TCBB.2010.98.

the relations predicted that in fact correspond to true regulations, while the second is due to the fact that DBNs remain far from being efficient, specially given the data's large size [3].

This work is concerned with improving the results of learning gene regulatory networks with Dynamic Bayesian Networks using the aid of commonsense information extracted from the gene expression data. More specifically, we examine the use of *Qualitative Probabilistic Networks* (QPNs) [26], [21], which are qualitative abstractions of Bayesian Networks, in providing a better characterization of gene regulation relations among genetic components. QPNs are directed acyclic graphs (DAGs) that preserve the independence and structure properties of Bayesian Networks but instead of keeping track of the local conditional probability distribution of each node in the network, QPNs only observe how the probabilities of the various nodes are affected by changes in the probabilities of their immediate parents. These effects are described in nonnumerical terms such as increasing, decreasing, constant or unknown. In this paper, we use these relationships to define the meaning of gene regulation and use the resulting formalism to guide the DBN learning algorithm for discovering the topology of the regulatory network underlying the genetic data.

The paper is centered around the idea that having a better-defined notion of regulation, which exploits higher level commonsense information extracted from the gene expression profiles, complements the analysis of the data. Our argument is motivated by the many sources of uncertainty surrounding the data, and the availability of useful qualitative information awaiting to be mined. To begin with, the numbers given in microarray experiments represent outcomes of a single and nonrepeated experiment. This is especially important in determining how dependable the data are given the sparse nature of the resulting measurements [7]. In addition, the dynamic nature of the expression process, and the fact that it depends on factors that may not be known [4], makes the numbers more untrustworthy because it is currently not known whether the variables affecting the expression at different intervals are constant through the experiment [23]. Despite the above, the uncertainty surrounding microarray data does not prevent the extraction of useful qualitative information that can be used to uncover the underlying genetic interactions and effectively reason about them to obtain biological insight. In fact, microarray data contain information pertaining the conditional dependence among the genes in question, variable time delays, and the combined effects of complexes of end products over genes. Although this information can be modeled correctly using the Bayesian approach [16], there is other information of a strictly qualitative nature that can be extracted due to the monotonicity of genetic interactions. More specifically, instead of using conditional probabilities to uncover the type of regulatory relation present between two (or more) genes (being of a stimulatory or an inhibiting nature), defining the conditions under which seemingly conditionally dependent genes do, in fact, exhibit regulatory ties is not directly derivable using the probabilistic model. Instead, the qualitative relations defined by QPNs can provide better clues to regulation as they have an explicitly defined notion of *influence*, making one perfectly capable of formally defining the behavior of regulation relations, and if used properly, they can be used to either uncover the network model or produce a candidate set of possible regulators that can reduce the search space for a DBN, which is the approach that we will follow here.

Another important motivation is the intricacy of biological pathways and the ongoing challenge of their discovery. It is now accepted that in order to obtain biological insight, it is viable to examine data from different sources in the aim of forming an integral examination of cellular interactions, e.g., gene expression and protein-protein interactions [8]. Integrating data from the various sources brings about issues such as compatibility and standardization of the numbers obtained from the different technologies. As a result, being able to extract higher level information providing clues to the meaning of the numbers may serve as a good vehicle for integration, given that the focus of such qualitative information is on how the numbers change instead of what they have changed to [14].

The contribution of the paper lies in presenting an improved model for learning the structure of gene regulatory networks from microarray data by incorporating qualitative knowledge in the DBN learning algorithm. The model presented here improves the quality of the learned network and is computationally less costly.

The rest of the paper is organized as follows: we begin by introducing Qualitative Probabilistic Networks as abstractions of Bayesian Networks, in Section 2. In Section 3, we identify properties that QPNs lack and which must be present if QPNs are to be useful for our purpose and devise a new model which incorporates these properties. The result is a new qualitative formalism we call *Dynamic Qualitative Probabilistic Networks* (DQPNs) which we use for the rest of the paper. Section 4 details the steps we followed to construct DQPNs from time-series data, and how they are used to enhance the process of learning the structure of gene regulatory networks using Dynamic Bayesian Networks. The experimental evaluation of our approach is given in Section 5, followed by a conclusion which provides a summary and some future directions in Section 6.

2 Qualitative Probabilistic Networks

Qualitative reasoning is now a well-established area in Artificial Intelligence [28]. The field is concerned with explaining and predicting the behavior of physical phenomena without (or with the minimal use of) numerical information. It is motivated by the observation that people are capable of drawing subtle conclusions about many aspects of the physical world using less data than numerical and quantitative methods require. A subfield of qualitative reasoning is concerned with modeling probabilistic systems qualitatively and is based on the idea of building a reasoning system that makes full use of the principles underlying probabilistic reasoning but instead of using exact probabilities, it captures how probabilities change using categorial knowledge [26]. This is done by replacing conditional probabilities by relations describing how a variable's likelihood changes given the probability of the variables upon which it is conditionally dependent. The change is modeled by qualitative terms such as increase,

decrease, no change, or an unknown change [26]. The idea has been extended to formulate qualitative equivalents of Bayesian networks, termed as QPNs.

QPNs are DAGs that represent a qualitative abstraction of Bayesian Networks [21], [26]. Formally, a QPN is given by a pair $G = (V(G), Q(G))$, where $V(G)$ is the set of nodes capturing random variables and $Q(G)$ is the set of arcs capturing the conditional dependence among the variables as in Bayesian Networks. Instead of a known conditional probability distribution however, the arcs of a QPN capture qualitative relations by finding monotonic characteristics in the local conditional probability distribution of each node based on the idea of first-order stochastic dominance [21]. The resulting relations are used to establish properties over the probabilities of events and are of two types, binary qualitative influences and tertiary qualitative synergies [26].

Influences describe how the change of the value of a single variable affects that of another, with the effect being categorized as positive, negative, constant, or unknown.

A positive influence exists between a parent node X and its child Y (X is said to positively influence Y, written as $I^+(X, Y)$) if observing higher values for X makes higher values of Y more probable, regardless of the value of any other node which may directly influence Y (i.e., any other parent of Y, denoted by W) as given in Definition 1. The definition assumes that the variables X and Y are binary and places a partial order on their values such that for a variable X with two values x and $\neg x$, $x > \neg x$. Negative, constant, and unknown influences are analogously defined by replacing the $>$ sign by $<$, $=$, and ?, respectively. While we use binary variables here to define influences for simplicity, the definition can be easily extended for multivalued variables by placing the values in their appropriate locations in the inequality.

Definition 1 (Positive Influence).

$$I^+(X, Y) \text{ iff } Pr(y|x, W) > Pr(y|\neg x, W).$$

An example of a QPN illustrating influences is given in Fig. 1. In the figure, $V(G) = \{\text{Gene A}, \text{Gene B}, \text{Gene C}, \text{Gene D}, \text{Gene E}\}$ and $Q(G) = \{(\text{Gene A}, \text{Gene D}), (\text{Gene A}, \text{Gene C}), (\text{Gene B}, \text{Gene C}), (\text{Gene C}, \text{Gene E})\}$. The only information encoded in the arcs is the signs of the influences from one node to another extracted from the conditional probability tables of each node. For instance, the negative influence exerted by Gene A on Gene D comes naturally from Gene D's conditional probability table given its parent Gene A. A similar picture can be drawn to conclude I^+(A,C) and I^+(B,C). In the case of I^+(A,C), W of Definition 1 is the set {B} and the sign of the influence is obtained by comparing the probabilities $Pr(c|a, B)$ (which is 1.05) and $Pr(c|\neg a, B)$ (which is 1.0). With I^+(B,C), W of Definition 1 is the set {A} and the sign of the influence is the result of the comparison of probabilities $Pr(c|b, A)$ (which is 1.7) and $Pr(c|\neg b, A)$ (which is 0.35).

Although qualitative influences define the basic interactions among variables, they are not always sufficient to capture all the interactions that exist in the network. This is the case when it is necessary to identify the combined effect of a pair of parent nodes in union on another variable. For

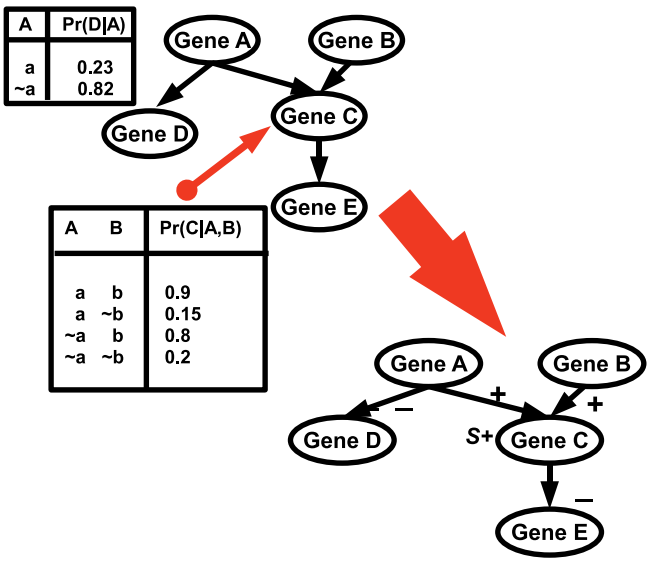

Fig. 1. A Bayesian network with the corresponding QPN.

this, the concept of qualitative synergies is created in order to model the interaction among the influences between three nodes in a network's diagraph. Qualitative synergies are essentially of two classes depending on the type of interaction, mainly *additive* and *product* synergies, and can be positive, negative, constant, or unknown as in the case with influences. Since product synergies are not of direct relevance to this work, we will do away with a discussion about them.

Additive synergies describe the situations in which the combined influence of the parents on their common child is greater than the individual influence of each parent on the child [26]. For example, a positive additive synergy of two nodes X and Y on their common child Z, written as $S^+(\{X, Y\}, Z)$, exists if the sum of their joint influence on Z is greater than the sum of their separate influences regardless of the value of any direct ancestor W of Z other than X and Y as given in Definition 2. As in the case of influences, the definition is stated for binary variables but can be similarly extended to multivalued ones.

Definition 2 (Positive Additive Synergy) [26]. *$S^+(\{X, Y\}, Z)$ iff for any values x, y, z of X, Y, Z, respectively, and for any variable W such that $W \in pa(Z)/\{X, Y\}$, we have*

$$Pr(z|x, y, W) + Pr(z|\neg x, \neg y, W) > Pr(z|x, \neg y, W) + Pr(z|\neg x, y, W).$$

where $pa(Z)$ denotes the set of Z's parents; therefore, $pa(Z)/\{X, Y\}$ is the set of all Z's parents except for X and Y. In Fig. 1, Gene A and Gene B exhibit a positive additive synergy on their common child Gene C as the label $S+$ placed over the node C shows. This relation can be verified from Gene C's conditional probability table given its parents; in this case, $W = \{\}$. Negative and constant additive synergies are analogously defined.

Observed evidence is propagated through the network via qualitative operators that combine influences and produce their net effects. There are two such operators serving different topologies of arcs. When evaluating the

TABLE 1
Sign Multiplication ($\otimes$) and Sign Addition ($\oplus$) Operators

$\otimes$	+	−	0	?
+	+	−	0	?
−	−	+	0	?
0	0	0	0	0
?	?	?	0	?

$\oplus$	+	−	0	?
+	+	?	+	?
−	?	−	−	?
0	+	−	0	?
?	?	?	?	?

net effect of influences in a chain (such as the combined influence of Gene A on Gene E), the sign multiplication operator given in the left portion of Table 1 is used (resulting in a negative net influence). On the other hand, parallel connections (such as the individual influences of Gene A on Gene C and that of Gene B on Gene C) are evaluated using the sign addition operator given on the right portion of the table (resulting in a net positive influence). The signs propagate through the network until the net effect of the evidence is observed by the polynomial-time sign-propagation algorithm [24].

It is worth noting that QPNs suffer from coarseness, which can result in many ambiguous signs as Table 1 shows. However, because our aim is to use QPNs to only discover the topology of genetic networks, we will not discuss means for resolving the conflicts that can arise. The interested reader can refer to [21] for a general discussion and to [13] for a more biologically relevant application of conflict resolution.

3 QPNs for Gene Regulation

Using the intuition that if some gene g_1 is said to regulate another gene g_2, then observing higher expression values for g_1 renders higher expression levels of g_2 more likely in the case of upregulation or less likely in the case of downregulation, one can map regulatory relations to qualitative QPN influences and use QPNs to model the topology of gene regulatory networks. Hence, the key to our approach is formally establishing a mapping between QPN constructs and gene regulation relations.

However, there are two crucial aspects in which QPNs and gene regulatory networks differ. First, because QPNs preserve the DAG structure of Bayesian Networks, they are incapable of handling cyclic relations which are abundant in gene regulatory networks. Second, in contrast to binary influences and tertiary synergies, gene regulation relations may hold between an arbitrary number of parents and their children. To deal with these two limitations of QPNs, this section defines additional properties and constructs for QPNs to make them more usable for our purpose.

3.1 Handling Cyclic Relations: Dynamic Qualitative Probabilistic Networks

Here, we present Dynamic QPNs (DQPNs) as a temporal extension of QPNs to enable them to handle time-series data and enable cyclic interactions. The model is an improvemnt on the model prsented in [18] and an extension to our work in [13].

3.1.1 Terminology

Let U be a set of n variables drawn from Pr, an unknown probability distribution on U, and let T be a totally ordered set of m temporal slices such that $T_1 \ldots T_m \in T$. We denote the set of variables in each temporal slice by U^t $(1 \leq t \leq m)$ and the set of n variables in U^t by X_i^t $(1 \leq i \leq n)$.

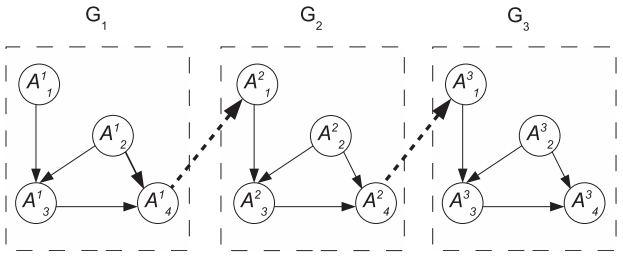

Fig. 2. An example of G.

Definition 3 (Temporal Snapshot). *Let $G = (V(G), Q(G))$ be a DAG such that* **G** *is the qualitative probabilistic network representing U. An instance G_t of G represents a temporal snapshot of G in time slice T_t such that G_t retains the DAG structure of G.*

Example 1. Consider Fig. 2 representing a fictitious graph G capturing the I-map for Pr, the joint probability distribution on $U = \{A_1, A_2, A_3, A_4\}$. Each instance G_t of G $(1 \leq t \leq 3$ in the figure) represents a snapshot of G, where the variables in each temporal slice are given by $U_t = \{A_1^t, A_2^t, A_3^t, A_4^t\}$.

Definition 4 (Dynamic Instance). *Let G_t be as given in Definition 3. G_t defines a dynamic instance of the QPN whose structure is defined by G and is given by $G_t = (V(G_t), \{Q(G_t) \bigcup T(G_t)\})$,[1] where $V(G_t)$ and $Q(G_t)$ are instances of $V(G)$ and $Q(G)$, respectively, at time slot t, and $T(G_t)$ describes the interslot conditional dependence between variables in $V(G_t)$ and its immediate neighbor $V(G_{t+1})$.*

Example 2. In the graph given in Fig. 2, for each G_t,

$$V(G_t) = U_t,$$
$$Q(G_t) = \{(A_1^t, A_3^t), (A_2^t, A_3^t), (A_3^t, A_4^t), (A_2^t, A_4^t)\},$$

and $T(G_t) = \{(A_4^t, A_1^{t+1})\}$.

Both $Q(G)$ and $T(G)$ encode a set of arcs for G to capture the set of qualitative relations representing how variables influence each other. For this, we redefine the concept of a qualitative influence to capture not only within-slot relations, but also interslot ones. Before doing so however, we first present the definition of a DQPN below.

Definition 5 (Dynamic QPN). *Let $(G_1 = (V(G_1), \mathcal{Q}(G_1)), \ldots, G_m = (V(G_m), \mathcal{Q}(G_m)))$ be a total ordering of the m instances of G such that $T(G_t) \neq \emptyset$, $\forall 1 \leq t \leq m-1$. Then the compound graph of $G_1, \ldots, G_m$ defines a Dynamic Qualitative Probabilistic Network over G and is given by*

$$\bigcup_{t=1}^{m} G_t = \left(\bigcup_{t=1}^{m} V(G_t), \bigcup_{t=1}^{m} \mathcal{Q}(G_t) \right).$$

1. For readability purposes, we will refer to $\{Q(G_t) \bigcup T(G_t)\}$ as $\mathcal{Q}(G_t)$ in this work.

3.1.2 Qualitative Influences in a DQPN

Definition 6 (Positive DQPN Influence). *Let G_t and G_{t+1} be two adjacent subgraphs of the DQPN defined over G. Further, let X and Y be such that $X, Y \in V(G)$. A direct positive influence is exerted by node X over node Y, written as $I^+(X,Y)$ iff for all values x^i of X and y^j, $\neg y^j$ of Y, and for all integer values i and j such that $1 \leq i, j \leq m$ and $i - j \in \{0, 1\}$ we have*

$$Pr(x^i|y^j, W) > Pr(x^i|\neg y^j, W).$$

The superscripts i and j denote the temporal slot to which the instances x, y, and $\neg y$ belong. Moreover, the definition enforces a temporal order over its components by requiring that variables can only directly influence other variables that belong to the same temporal slot ($i = j$) or those that belong to the next immediate slot ($i = j + 1$). As in QPNs, W represents all other direct influences on Y other than X. Negative, zero, and unknown influences are analogously defined.

As the influences defined for DQPNs preserve the underlying principles of those defined for QPNs, they respect the combinatorial properties defined in Table 1 and can therefore be propagated according to their rules as in QPNs.

3.2 Generalized Joint Influences

As stated earlier, because regulation is a many-to-many relationship, single influences and binary synergies are not sufficient for their description. There must be a way to establish the combined influence on many parent nodes over their common child in order to be able to define those relations. For this, we define the notion of a *generalized joint influence* of a set of k variables $X_1, \ldots X_k$ over a target variable Y which describes the monotonic relationship between the values of the variables $X_1, \ldots X_k$ jointly and that of Y. Definition 7 below illustrates a positive generalized joint influence $J^+(\{X_1, \ldots X_k\}, Y)$. In the definition, the superscript i denotes the time slots at which the value of the child node y is observed while the superscripts $j_1, \ldots, j_k$ denote the time slots at which the influencing parents $X_1, \ldots, X_k$ are observed.

Definition 7 (Positive Generalized Joint Influence). *$J^+(\{X_1, \ldots X_k\}, Y)$ iff for value y of Y observed at time slot i and for any combination of values for variables $X_1, \ldots X_k$ observed at time slots $j_1, \ldots, j_k$ such that $j_1, \ldots, j_k \leq i$:*

$$\begin{aligned} &Pr\big(y^i|{x_1}^j, W\big) > Pr\big(y^i|\neg {x_1}^j, W\big), \\ &\qquad\qquad\qquad\qquad when\ k = 1, \\ &Pr\big(y^i|x_1^{j_1}, .., x_k^{j_k}, W\big) + Pr\big(y^i|\neg x_1^{j_1}, .., \neg x_k^{j_k}, W\big), > \wp \\ &\qquad\qquad\qquad\qquad when\ k > 1. \end{aligned}$$

Where $\wp$ is the sum of the conditional probability of Y given any combination of values for $X_1, \ldots X_k$ other than $x_1^{j_1}, \ldots, x_k^{j_k}$ and $\neg x_1^{j_1}, \ldots \neg x_k^{j_k}$.

It can be seen that the case of binary synergies can be directly extracted from the definition by setting $k = 2$ and that negative and zero joint generalized influences can be analogously defined by replacing $>$ by $<$ and $=$, respectively.

In our next steps, we will use generalized joint influences of DQPNs to guide the process of identifying regulator genes for a given target. When referencing the influences defined above, we will use the notation $J^{\varrho}(\{X_1, \ldots X_k\}, Y)$, where $\varrho \in \{+, -, 0\}$.

4 Our Approach

In this section, we describe the use of DQPNs and the generalized joint influences defined over them to aid the construction of a DBN from microarray data. The approach is based on 1) using DQPN generalized joint influences to identify the set of regulators for each gene; 2) estimating time lags of regulations from the expression data; and 3) infusing the qualitative knowledge in a DBN learning algorithm by using the candidate set of regulators to reduce the search space of possible models. The steps of our approach are detailed below.

4.1 Constructing the Qualitative Model

This step makes use of the monotonic relations corresponding to regulator-target interactions for the identification of the set of potential regulators of a specific gene. The step is twofold: first a quantitative analysis is performed based on comparing the times of significant initial change in expression levels of the genes to construct an initial set of candidate regulators for each gene, and then another step follows to discover those candidates that exhibit a monotonic behavior with respect to the target gene and discard spurious interactions by building the subsets of regulators that jointly exhibit a generalized influence over the target gene as described in Section 3.2.

4.1.1 Gathering Potential Regulators

The quantitative step is not unique to our work and is based on the hypothesis that more often than not, regulators exhibit an earlier up- or down-change in their expression levels than that of the regulated genes [32]. This is of course not always the case as will be clear in Section 5.2 but we think that it is a good estimator of regulation relations that can be improved and built upon later on.

A gene is said to be up- or downregulated if its expression level goes up (in the case of up-regulation) or goes down (in the case of downregulation) by a certain fold change α. Since the aim of this step is to identify all potential regulators, we decided to use modest cutoffs of $\alpha = 1.1$ for up-regulation or $\alpha = 0.9$ for downregulation in order not to overlook potential regulators.

Once the genes with significant fold changes have been identified, then for each gene g, the genes with simultaneous or preceding fold changes are placed in the set of potential regulators of g. Establishing this set for every gene marks the completion of the quantitative part of this step.

A simple fictitious example illustrating this step is given in Fig. 3. In the figure, gene g_3 has both g_1 and g_2 as potential regulators because its expression level had a significant increase at time step 4, which follows the time steps for which g_1 and g_2 had a significant increase in their expression levels.

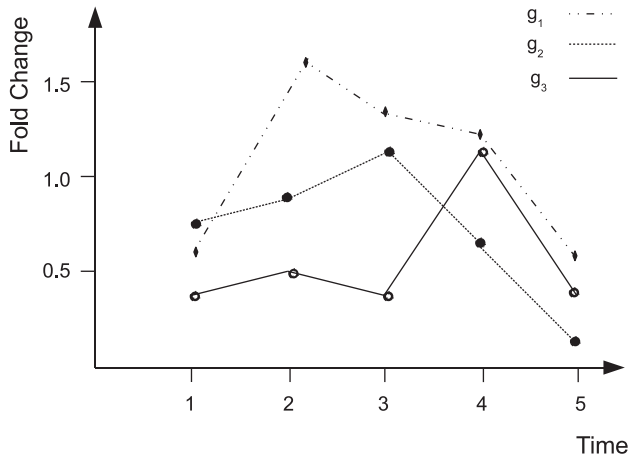

Fig. 3. A hypothetical example illustrating possible regulators.

4.1.2 Extracting Most-Likely Regulators

A quick critical examination of the procedure described above reveals that 1) it can potentially incorporate many spurious relations because it assumes that the change in expression levels entails a regulatory relation and does not consider a more well-defined notion of regulation (such as the one we provided in Section 3.2), which is the case for all stochastic approaches currently existing in the literature; and 2) it does not distinguish between i) coregulation, where several genes collectively activate or inhibit the expression of a target gene and ii) simple regulation, where a target gene has a set of regulators, each individually regulating the gene without the need for the other regulators to be present. This is where the qualitative relations defined over our model come into the picture in a procedure described in Algorithm 1. The idea is to find the maximum number of potential regulators that exhibit monotonic effects on the expression of the regulated gene, and call the resulting set the most-likely regulators of the gene. The algorithm receives as an input a gene g along with the set R of its potential regulators identified using the quantitative method described above. The output is a collection O of subsets of R where each individual subset contains the genes that together coregulate g.

Algorithm 1.
Require: Gene g and set R of its potential regulators.
Ensure: Set O contains the most-likely regulating sets of g.
1: **for** $k = |R|$ to 1 **do**
2: $\forall R_{sub} \subseteq R : |R_{sub}| = k$
3: **if** $\forall r_1^{i_1}, \ldots r_k^{i_k} \in R_{sub}, J^{\varrho}(\{r_1^{i_1}, \ldots, r_k^{i_k}\}, g)$ **then**
4: $O \leftarrow R_{sub}$
5: **end if**
6: **end for**
7: **for** $O_{sub} \subset O$ **do**
8: **if** $\exists\, O_{sub2} \subseteq O : O_{sub2} \subset O_{sub}$ **then**
9: $O \leftarrow O - O_{sub2}$
10: **end if**
11: **end for**

Lines 1-6 of the algorithm construct the subsets of R of decreasing size whose elements jointly exhibit a generalized influence over g. For each subset R_{sub} of R of size k (line 2), if the elements of R_{sub} satisfy some generalized influence J^{ϱ} over g (condition in line 3), then R_{sub} is added to O, the set of most-likely regulating sets of g (line 4).

The second phase of the algorithm (lines 7-11) removes redundant subsets by making sure that any proper subsets of O_{sub} (denoted by O_{sub2}) are not included in the set of all potential regulators given that its superset is included (line 8). This phase also establishes the distinction between joint and individual regulators by ensuring that for every subset of potential regulators O_{sub} of O, one-element subsets made of its individual members are not included in the final output O as this corresponds to stating that each element of O_{sub} individually regulates g.

It is important to note that the time delays of the elements of the collection O for every regulated gene g are directly encoded in the construction of the set as the condition checks for the generalized joint influence given in line 3.

Moreover, there are several points worth noting with respect to the use of generalized joint influences of Definition 7 and Algorithm 1 for discovering regulatory relationships. They are as follows:

1. The temporal precedence properties of generalized joint influences are more relaxed than in Definition 6 of DQPN influences. This is to allow the discovery of regulation relations between genes that may not belong to two consecutive time slots as fold changes of regulating genes may occur much earlier than those of target genes.
2. Generalized joint influences describe the combined influence of multiple parents such that all the influences yield the same sign, be it positive, negative, or constant. As a result, a target gene node may have two or more sets of generalized joint influences exerted on it by different subsets of its regulators according to how the elements of each subset satisfy the definition of the corresponding generalized joint influence.
3. Unknown influences generated by Definition 7 correspond to no regulation in the resulting gene regulatory network.

4.2 Time-Lag Estimation

One issue with respect to the use of DBNs to model gene regulatory networks is that DBNs construct conditional distributions over fixed time intervals measured according to the time series. This has been found to be problematic [29] as it can miss potential regulation relations. However, because our approach incorporates the time lag between each gene's expression and that of its potential regulators by marking the difference between their significant fold changes, the resulting model will not suffer from this problem.

Hence, for each gene g, we collected 1) the sets of joint regulator genes O and 2) their corresponding time lags. The resulting adjacency list L of length N contains this information for all N genes such that for each gene in the list L[j], $1 \leq j \leq N$, a linked list containing L[j]'s set of most-likely regulators O is added where each node of the list represents one subset of joint regulators (O_{sub} in Algorithm 1) along with their times of significant fold changes as the algorithm has shown.

4.3 Aided Learning with Qualitative Joint Influences

A score-maximizing learning algorithm that utilizes the additional aid of the list L was used to construct the target GRN. The corresponding criterion for maximizing the score is composed of two quantities. The first is, the prior structure of the network established by L which contains the most-likely regulators of every target gene constitutes a model that is used in this step as the base for a model search using DBN learning. The second is the marginal likelihood of the data, which measures how the model fits the microarray data. For each target gene, its conditional probability given its regulators is constructed from the expression data and were used to compute marginal likelihood scores.

5 Experimental Results

5.1 Data and Preprocessing

We used the gene expression time-course data set of the *Drosophila Melanogaster* genetic network obtained from the Drosophila interaction database to compare our approach with that of [32]. The data set contains 4,028 genes whose expression levels were sampled at 74 time points covering the four life cycle stages of embryonic, larval, pupal, and adulthood [1].

The original data set is quantized into fold-change series by computing the ratio of expression of each gene g at two consecutive time points x_t and $x_{t+\Delta}$. The resulting set contains the fold changes enabling the establishment of times of significant change in expression for all the genes. Missing values are computed using a simple linear interpolation by obtaining the mean of the preceding and following neighbors in the expression time series for the specific gene. When the missing expression is a start or an end point in the time series, it is replaced by the nearest observed neighbor's value (resulting in no significant fold change).

5.2 Accuracy Evaluation

In order to obtain an initial visual image of the performance of our method, we first used it to construct the GRN of a selected set of 12 genes from our data set. The selected genes have been reported to describe the larval somatic muscle development stage of Drosophila Melanogaster and contain a total of 18 known interactions.

A comparison between our approach and that of [32] for this set of genes is shown in Fig. 4. While our approach successfully identified 15 interactions, using [32]'s approach only identified 11 of the total 18 interactions. Upon a close examination, we found that the missing interactions from our network are due to the assumption of regulators having an earlier or simultaneous expression time than regulated genes. For instance, examining CG9843, which regulates CG7447, it turns out that CG9843 has a much shorter half-life than CG7447. As a result, the regulator's mRNA will take much longer in reaching a steady-state level of up- or downregulation compared to the regulated gene, resulting in an apparent later change of expression. Since this assumption is also made by [32], the method did not identify these interactions either. Moreover, our approach identified coregulating genes using the synergetic definition of the initial model. These coregulations along with

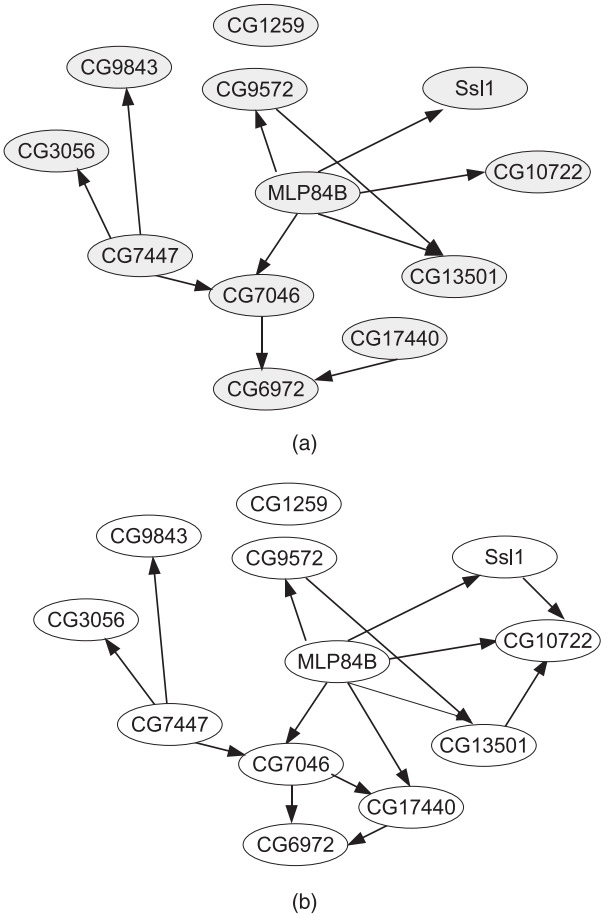

Fig. 4. Muscle development network in Drosophila's larval stage (a) using [32]'s DBN, and (b) using our approach.

feedforward loops were largely missed by [32]'s method, resulting in a smaller number of identified relations. Apart from (G9843, G7447), our network is missing the interactions (CG6972,CG2046) and (CG13501,CG17440).

Comparison of larger subsets of the networks is given in Table 2. In the table, the experiments are labeled by $\mathrm{DBN}_{qual}(N, I)$ or $\mathrm{DBN}_{ZN}(N, I)$ denoting the approach used (DBN_{qual} refers to our method and DBN_{ZN} refers to the method used in [32]) with N denoting the number of genes involved in the network while I refers to number of known interactions. The sizes of the networks were selected randomly and the subset of genes involved being based on the current interaction diagram of the Drosophila genetic network.

For each network size, we conducted 10 runs and reported the average performance measures of the number of correctly identified edges or true positive edges (C), the number of misidentified edges or false positive edges (F) which have been identified by the learning algorithm but do not exist in the real network, and the number of missed edges (M) which are edges that exist in the real network but were either unidentified or given the wrong regulator-regulated gene direction in the inferred network. We calculated precision as the ratio $\mathrm{C}/(\mathrm{C}+\mathrm{M})$ and recall as the ratio $\mathrm{C}/(\mathrm{C}+\mathrm{M})$ and listed them accordingly in Table 2.

The results given in the table show the clear improvement our approach presents in terms of both precision and

TABLE 2
Results of Comparing DBN_{qual} and DBN_{ZN} Using Differently-Sized Sample Networks

	Correct edges (C)			False Edges (F)			Missed Edges (M)			Precision	Recall	Average Time
	min	max	avg	min	max	avg	min	max	avg			avg
$\mathrm{DBN}_{\mathrm{qual}}$ (12, 18)	13	17	15	0	4	2	1	3	2	0.88	0.88	12.9
$\mathrm{DBN}_{\mathrm{ZN}}$ (12, 18)	9	12	10.5	2	9	5.5	5	5	5	0.66	0.66	19.5
$\mathrm{DBN}_{\mathrm{qual}}$ (40, 80)	51	62	54.5	15	19	17	21	28	24.5	0.689	0.76	31.9
$\mathrm{DBN}_{\mathrm{ZN}}$ (40, 80)	38	46	42	35	41	38	40	49	44.5	0.486	0.525	38.4
$\mathrm{DBN}_{\mathrm{qual}}$ (50, 95)	61	69	65	29	42	35.5	28	34	31	0.677	0.647	23.1
$\mathrm{DBN}_{\mathrm{ZN}}$ (50, 95)	45	51	48	48	57	52.5	51	68	54.5	0.468	0.478	34.3
$\mathrm{DBN}_{\mathrm{qual}}$ (80, 100)	68	79	73.5	41	49	45	38	43	40.5	0.645	0.62	57.3
$\mathrm{DBN}_{\mathrm{ZN}}$ (80, 100)	53	64	58.5	45	51	48	41	49	45	0.56	0.46	86.6

recall. Our improved precision is due to the discovery of joint regulations and feedforward loop identifications as discussed earlier. Our increased recall is due to the better definition of regulation provided by the monotonic relations of the synergies and influences that QPNs provide. The numbers clearly show that this definition helps in eliminating many spurious correlations that do not correspond to regulatory relations.

5.3 Efficiency Evaluation

Table 2 shows the average running time of the algorithms at each experimental setup. The time taken by our algorithm does not include the step of generating the most-likely regulators in our algorithm, so that the actual learning time of our algorithm can be compared with that of [32].

The improvement presented by our approach is contributed to the fact that Algorithm 1 provides a candidate set which minimizes the number of potential regulators so that the only possible regulators are those that exhibit the monotonicity of qualitative influences and synergies and exclude those exhibiting a correlation that does not correspond to a regulatory relation. This optimal candidate set is the main contributor to the better performance exhibited by our algorithm.

6 Conclusions

We have presented a model that uses qualitative probability to discover monotonic relations among genes by comparing their expression profiles and using the discovered qualitative relations to aid the DBN learning algorithm in constructing better and more efficient models of the corresponding GRN. We presented an experimental study that compares our results in terms of accuracy and efficiency with the approach found in [32], which is an accepted benchmark for DBN learning and found that the added qualitative knowledge highly improves the type of model inferred and the efficiency of the learning procedure. The results were compared using the Drosophila Melanogaster gene regulation data set.

Future directions include the comparison of the algorithm with non-Bayesian approaches for GRN construction as well as the exploration of using qualitative knowledge with other forms of high-throughput data. Another longer term aim is to examine the use of QPN constructs in the process of integrating data from multiple sources to form a global view of the various cellular interactions.

References

[1] M. Arbeitman et al., "Gene Expression During the Life Cycle of Drosophila Melanogaster," *Science,* vol. 297, pp. 2270-2275, 2002.

[2] T. Chen, H. He, and G. Church, "Modeling Gene Expression with Differential Equations," *Proc. Pacific Symp. Biocomputing,* pp. 29-40, 1999.

[3] D.M. Chickering, D. Heckerman, and C. Meek, "Large-Sample Learning of Bayesian Networks Is NP-Hard," *The J. Machine Learning Research,* vol. 5, pp. 1287-1330, 2004.

[4] P. D'Haeseleer, "Reconstructing Gene Regulatory Networks from Large Scale Gene Expression Data," PhD dissertation, Univ. of New Mexico, 2000.

[5] N. Dojer et al., "Applying Dynamic Baeysian Networks to Perturbed Gene Expression Data," *BMC Bioinformatics,* vol. 7, pp. 249-260, 2006.

[6] Z. Fang et al., "Comparisons of Graph-Structure Clustering Methods for Gene Expression Data," *Acta Biochimica et Biophysica Sinica,* vol. 38, no. 6, pp. 379-384, 2006.

[7] V. Filkov, S. Skiena, and J. Zhi, "Analysis Techniques for Microarray Time-Series Data," *J. Computational Biology,* vol. 9, pp. 317-330, 2002.

[8] N. Friedman, "Inferring Cellular Networks Using Probabilistic Graphical Models," *Science,* vol. 303, pp. 799-805, 2004.

[9] Y. Guo et al., "How is mRNA Expression Predictive for Protein Expression: A Correlation Study on Human Circulating Monocytes," *Acta Biochimica et Biophysica Sinica,* vol. 40, no. 5, pp. 426-436, 2008.

[10] D. Heckerman, D. Geiger, and D. Chickering, "Learning Baeysian Networks: The Combination of Knowledge and Statistical Data," *Machine Learning,* vol. 20, pp. 197-243, 1995.

[11] X. Hu and F. Wu, "Mining and State-Space Modeling and Verification of Sub-Networks from Large Biomolecular Networks," *BMC Bioinformatics,* vol. 8, pp. 324-342, 2007.

[12] X. Hu, M. Ng, F. Wu, and B. Sokhansanj, "Mining, Modeling and Evaluation of Sub-Networks from Large Biomolecular Networks and Its Comparison Study," *IEEE Trans. Information Technology in Biomedicine,* vol. 13, no. 2, pp. 184-194, Mar. 2009.

[13] Z. Ibrahim, A. Tawfik, and A. Ngom, "Qualitative Motif Detection in Gene Regulatory Networks," *Proc. IEEE Int'l Conf. Bioinformatics and Biomedicine (BIBM),* pp. 124-129, 2009.

[14] S. Iyenga and M. McGuire, "Imprecise and Qualitative Probability in Systems Biology," *Proc. Int'l Conf. Systems Biology,* 2007.

[15] H. Lähdesmäki, I. Shmulevich, and O. Yli-Harja, "On Learning Gene Regulatory Networks under the Boolean Network Mode," *Machine Learning,* vol. 52, nos. 1/2, pp. 147-167, 2003.

[16] T. Liu and W. Sung, "Learning Gene Network Using Conditional Dependence," *Proc. IEEE Int'l Conf. Tools with Artificial Intelligence,* pp. 800-804, 2006.

[17] W. Liu, K. Yue, S. Liu, and Y. Sun, "Qualitative-Probabilistic-Network-Based Modeling of Temporal Causalities and Its Application to Feedback Loop Identification," *Int'l J. Information Sciences,* vol. 178, no. 7, pp. 1803-1824, 2008.

[18] K. Murphy and S. Mian, "Modeling Gene Expression Data Using Dynamic Bayesian Networks," technical report, Computer Science Division, Univ. of California, 1999.

[19] L. Nie, G. Wu, W. Zhang, L. Nie, G. Wu, and W. Zhang, "Correlation of mRNA Expression and Protein Abundance Affected by Multiple Sequence Features Related to Translational Efficiency in Desulfovibrio Vulgaris: A Quantitative Analysis," *Genetics,* vol. 147, pp. 2229-2243, 2006.

[20] L. Pascal et al., "Correlation of mRNA and Protein Levels: Cell Type-Specific Gene Expression of Cluster Designation Antigens in the Prostate," *BMC Genomics*, vol. 23, no. 9, pp. 246-258, 2008.
[21] S. Renooij and L. Van der Gaag, "From Qualitative to Quantitative Probabilistic Networks," *Proc. Int'l Conf. Uncertainty in Artificial Intelligence*, pp. 422-429, 2002.
[22] B. Perrin et al., "Gene Networks Inference Using Dynamic Bayesian Networks," *Bioinformatics*, vol. 9, suppl. 2, pp. II138-II148, 2003.
[23] A. Pisabarro et al., "Genetic Networks for the Functional Study of the Genomes," *Briefings in Functional Genomics and Proteomics*, vol. 7, no. 4, pp. 249-263, 2008.
[24] F. Van Kouwen, S. Renooij, and P. Schot, "Inference in Qualitative Probabilistic Networks Revisited," *Int'l J. Approximate Reasoning*, vol. 50, no. 5, pp. 708-720, 2009.
[25] D. Weaver, C. Workman, and G. Stormo, "Modeling Regulatory Networks with Weight Matrices," *Proc. Pacific Symp. Biocomputing*, vol. 4, pp. 112-123, 1999.
[26] M. Wellman, "Fundamental Concepts of Qualitative Probabilistic Networks," *Artificial Intelligence*, vol. 44, pp. 257-303, 1990.
[27] L. Wessels, E. Someren, and M. Reinders, "A Comparison of Genetic Network Models," *Proc. Pacific Symp. Biocomputing (PSB)*, vol. 6, pp. 508-519, 2001.
[28] B. Williams and J. De Kleer, "Qualitative Reasoning about Physical Systems: A Return to Roots," *Artificial Intelligence*, vol. 51, nos. 1-3, pp. 1-9, 1991.
[29] Z. Xiing and D. Wu, "Modeling Multiple Time Units Delayed Gene Regulatory Networks Using Dynamic Baeysian Networks," *Proc. IEEE Int'l Conf. Data Mining*, pp. 190-195, 2005.
[30] R. Xu, D. Wunsch, II, and R. Frank, "Inference of Genetic Regulatory Networks with Recurrent Neural Network Models Using Particle Swarm Optimization," *IEEE/ACM Trans. Computational Biology and Bioinformatics*, vol. 4, no. 4, pp. 681-692, Oct. 2007.
[31] Y. Zhang et al., "Inferring Gene Regulatory Networks from Multiple Data Sources via a Dynamic Bayesian Network with Structural EM," *Proc. Int'l Conf. Data Integration in the Life Sciences*, pp. 204-214, 2007.
[32] M. Zou and S. Conzen, "A New Dynamic Bayesian Network (DBN) Approach for Identifying Gene Regulatory Networks from Time Course Microarray Data," *Bioinformatics*, vol. 2, no. 1, pp. 71-70, 2005.

Zina M. Ibrahim obtained the PhD degree from the University of Windsor, Ontario, Canada, in 2000, under the joint supervision of Dr. Ahmed Y. Tawfik and Dr. Alioune Ngom. She is a post-doctoral research associate at the University of Windsor. In her research, she investigated the formulation of qualitative uncertainty approaches and their use in reverse-engineering gene regulatory networks.

Alioune Ngom is an associate professor at the University of Windsor, Ontario, Canada. Prior to joining the University of Windsor, he held the position of an assistant professor at the Department of Mathematics and Computer Science at Lakehead University, Thunder Bay, Ontario, Canada, from 1998 to 2000. During his short stay at Lakehead University, he cofounded Genesis Genomics Inc., in 1999, a biotechnology company that specializes in the analysis of mitochondrial genome and the design of biomarkers for the early detection of cancer. His main research interests include but are not limited to computational intelligence and machine learning methods and their applications in computational biology and bioinformatics problems such as microarray analysis, protein analysis, oligonucleotide selection, bioimage analysis, and gene regulatory network analysis. He is a member of IEEE.

Ahmed Y. Tawfik is an associate professor of informatics and computing at the French University in Egypt. Prior to joining the UFE, he has held the position of an associate professor at the University of Windsor, Ontario, Canada. His research interests include intelligent systems, multiagent systems, knowledge representation and reasoning, and knowledge discovery. He is a member of IEEE.