Citation for published version (APA):
Ibrahim, Z. M., Tawfik, A. Y., & Ngom, A. (2009). Qualitative Motif Detection in Gene Regulatory Networks. In *The International IEEE Conference on Bioinformatics and Biomedicine* (pp. 124-129)

Qualitative Motif Detection in Gene Regulatory Networks

Zina M. Ibrahim and Ahmed Y. Tawfik and Alioune Ngom
School of Computer Science
University of Windsor
Windsor, Ontario, Canada
Email: {ibrahim,atawfik,angom}[at]uwindsor.ca

***Abstract*—This paper motivates the use of Qualitative Probabilistic Networks (QPNs) in conjunction with or in lieu of Bayesian Networks (BNs) for reconstructing gene regulatory networks from microarray expression data. QPNs are qualitative abstractions of Bayesian Networks that replace the conditional probability tables associated with BNs by qualitative influences, which use signs to encode how the values of variables change. We demonstrate that the qualitative influences defined by QPNs exhibit a natural mapping to naturally-occurring patterns of connections, termed network motifs, embedded in Gene Regulatory Networks and present a model that maps QPN constructs to such motifs.**

The contribution of this paper is that of discovering motifs by mapping their time-series experimental data to QPN influences and using the discovered motifs to aid the process of reconstructing the corresponding gene regulatory network via Dynamic Bayesian Networks (DBNs). The general aim is to compile a model that uses qualitative equivalents of Dynamic Bayesian Networks to explore gene expression networks and their regulatory mechanisms. Although this aim remains under development, the results we have obtained shows success for the discovery of regulatory motifs in *Saccharomyces Cerevisiae* and their effectiveness in improving the results obtained in terms of reconstruction using DBNs.

I. Introduction

Networks modeling the dynamics of the interactions of genetic information in the cell have become increasingly better studied in recent years. Complex networks modeling the behavior of genetic components and end-products (e.g. genes, DNA, RNA and proteins) serve as graphical models to better study the dynamics of the internal state of the cell. There exist many types of such networks, each differing by the macromolecular components modeled (e.g. DNA, protein) and the type of interactions captured. As a result, these networks have been broadly termed *biological regulatory networks* [16]. Identifying and understanding these regulatory mechanisms appear nowadays as one of the key challenges in systems biology with potential applications in therapeutical targeting, drug design, diagnosis and disease management [1], [12], [20].

A *gene expression network* is one type of biological networks in which every node represents a genetic component or end-product and every edge represents a regulation relationship. It is a directed graph that models how genes influence (through activation or inhibition) other genes in a complex web of interactions during the gene expression process. Uncovering the topology of the network from microarray expression data is currently one of the focuses of systems biology. It is mainly a reverse-engineering task to identify the true regulatory system from the observed gene expression profiles [23]. The complexity of the task stems from the fact that not only the kind of data available is of high dimensionality and suffers from great noise [8], but also because the data provides the expression levels of a large number of genes (usually tens of thousands) at different but relatively few (usually a few dozens) temporal intervals or conditions. Hence, it is usually sparse, which makes uncovering causal relations more difficult.

There currently exist Bayesian approaches for learning the structure of genetic networks [14], [8], [25], [24]. They have been successfully used to learn large scale networks but remain far from being efficient, specially given the data's large size [4]. This relative success of Bayesian approaches motivates this work. On one hand, (Dynamic) Bayesian Networks have been successfully used to detect the conditional (in)dependence and time-delay relations governing the structure of gene expression networks [10], [14]. On the other hand, it is the qualitative nature of the information extracted from the data that brought about the benefits of the model [8]. Hence, formulating a model that is specifically tailored to represent this information (in addition to other qualitative information Bayesian approaches may not be able to capture), then efficient ways to obtain insight regarding the functional interactions governing the data maybe uncovered.

For this, we set out to investigate the various ways in which qualitative abstractions of Bayesian Networks [19] can be useful with respect to the problem at hand. We present a model termed *Dynamic Qualitative Probabilistic Networks* (DQPNs), which extends the existing qualitative probabilistic networks to deal with temporal data. DQPNs are presented as an alternative to Dynamic Bayesian Networks that 1) focuses on the qualitative relations the time-series data presents for the discovery of the interactions in the regulatory networks 2) is more efficient than Dynamic Bayesian networks.

In this paper, we formally presented DQPNs and sketch the bases for using it for the reconstruction of gene regulatory networks. Experimentally, we use the model to explicitly define regulatory relations and discover patterns commonly occurring in regulatory networks, termed regulatory network motifs. We successively use the patterns defined, in conjunction with DBNS, to reconstruct gene regulatory networks. The qualitative relations discovered show improved accuracy of DBNs via experiments conducted on the time series data of *Saccharomyces Cerevisiae*.

After an introduction to Qualitative Probabilistic Networks in section II, we present the formal model used in defining DQPNs, the temporal equivalents of QPNs in section III and construct dynamic framework that can capture gene regulatory motifs and model them accordingly. The experiments of section V verify the mapping from DQPN constructs to network motifs and establish the usefulness of such mapping by showing that DBNs used to reconstruct gene regulatory networks can have an increased accuracy if making use of the aid of the qualitative motifs we defined. Some conclusions and future directions are presented in section VI.

II. Qualitative Probabilistic Networks

Qualitative Probabilistic Networks (QPNs) are directed acyclic graphs that represent a qualitative abstraction of Bayesian Networks [19], [22] . Formally, a QPN is given by a pair G = ($V(G)$,$Q(G)$), where $V(G)$ is the set of nodes capturing the variables of the domain being represented 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 among the variables by finding monotonic characteristics in the conditional probability distribution based on the idea of first-order stochastic dominance. The resulting relations are used to establish properties over the probabilities of events, and are of two types, qualitative influences and synergies [22].

Influences describe how the change of the value for one variable affects that of another and are of four types, positive, negative, constant and unknown.

A positive influence exists between two variable X and 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 variable which may directly influence Y (denoted by W) as given in Definition 1 below. The inequality 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 defined analogously.

Definition 1: **Positive Influence:**

$$I^+(X,Y)\ iff\ Pr(y|x,W) \geq Pr(y|\neg x,W)$$

An example of a QPN is given in Figure 1. In the figure, $V(G)$ = {A,B,C,D,E} and $Q(G)$ = {(B,C),(A,D),(C,D),(D,E),(B,E)}. The only information encoded in the arcs are the signs of the influences from one node to another. For instance, the figure shows that node A positively influences node D, while it has a negative influence on B.

Observed evidence is propagated through the network via qualitative operators that combine influences and produces their net effect. There are two such operators serving different topologies of arcs. When evaluating the net effect of influences in a chain, the sign multiplication operator given in the left portion of Table I, is used. For example, in order to obtain the effect of A on E via the path A-B-E, we have two examine a chain of two influences, that of A on B and of B on E. On the other hand, parallel connections are evaluated using the sign addition operator given in the right portion of the table. For example, two influences in parallel are required to establish the net effect of nodes A and C on node D, that of A on D and of C on D. The signs propagate through the network until the net effect of the evidence is observed on the required node or all the nodes are known to have been visited twice by the polynomial-time sign-propagation algorithm [6].

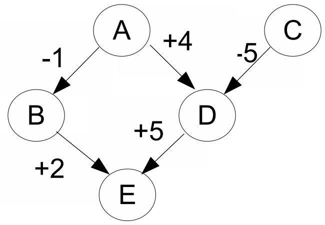

Fig. 1. An Example QPN

TABLE I
Sign multiplication ($\bigotimes$) and sign addition ($\bigoplus$) Operators [22]

$\bigotimes$	+	−	0	?
+	+	−	0	?
−	−	+	0	?
0	0	0	0	0
?	?	?	0	?

$\bigoplus$	+	−	0	?
+	+	?	+	?
−	?	−	−	?
0	+	−	0	?
?	?	?	?	?

It is worth noting that the original representation of QPNs [22] suffers from coarseness that has been dealt with in later work [19]. The resulting ambiguity is resolved by refining the model by incorporating more detail in the representation.

As can be seen in Table I, the coarseness of the representation results in many ambiguous signs. This has been dealt with in [19] by refining the model to incorporate more details, and subsequently reducing the chance of obtaining an ambiguous sign. For this [19] distinguish between strong and weak influences (where a strong positive influence of X on Y, termed $I^{++}(X,Y)$, carries more weight than a weak one, termed $I^+(X,Y)$ (with the same nomenclature used for negative, zero and unknown influences). [19] also provide a method for comparing indirect qualitative influences along different paths with respect to their strengths for trade-off resolution by retaining the length of the paths over which influences have been multiplied. For this, every influence's sign is augmented by a superscript, called the signs multiplication index, and is used as an indicator of its strength. Higher values of multiplication indices indicate a longer path and as a result, a weaker influence. This enables generalizing the sign-propagation algorithm of [6] by adapting the $\bigoplus$ and $\bigotimes$ operators to the different types of influences as given in Tables II and III.

TABLE II
ENHANCED SIGN MULTIPLICATION ($\bigotimes$) OPERATOR [19]

$\bigotimes$	$++^j$	$+^j$	0	$-^j$	$--^j$	$?$
$++^i$	$++^{i+j}$	$+^j$	0	$-j$	$--^{i+j}$	$?$
$+^i$	$+^i$	$+^{i+j}$	0	$-^{i+j}$	$-^i$	$?$
0	0	0	0	0	0	0
$-^i$	$-^i$	$-^{i+j}$	0	$+^{i+j}$	$+^i$	$?$
$--^i$	$--^{i+j}$	$-^j$	0	$+^j$	$++^{i+j}$	$?$
$?$	$?$	$?$	0	$?$	$?$	$?$

TABLE III
ENHANCED SIGN ADDITION ($\bigoplus$) OPERATOR [19]

$\bigoplus$	$++^j$	$+^j$	0	$-^j$	$--^j$	$?$
$++^i$	$++^{ij}$	$++^i$	$++^i$	a)	$?$	$?$
$+^i$	$++^j$	$+^{i,j}$	$+^i$	$?$	d)	$?$
0	$++^j$	$+^j$	0	$-^j$	$--^j$	$?$
$-^i$	b)	$?$	$-^i$	$-^{i,j}$	$--^{i,j}$	$?$
$--^i$	$?$	c)	$--^i$	$--^i$	$--^{i,j}$	$?$
$?$	$?$	$?$	$?$	$?$	$?$	$?$

a)	$++^{i,-j}, if \;\; i \leq j;$	?, otherwise
b)	$++^{-i,j}, if \;\; j \leq i;$	?, otherwise
c)	$--^{i,-j}, if \;\; i \leq j;$	?, otherwise
c)	$--^{i,-j}, if \;\; j \leq i;$	?, otherwise

III. TOWARDS QUALITATIVE REGULATION OF GENETIC NETWORKS: DYNAMIC QPNS

In this section, we present Dynamic QPNs (DQPNs) as a temporal extension of QPNs to qualitatively model a genetic network and show how it is used to model the commonly occurring motifs of gene expression networks.

A. 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 $T_1 ... T_m \in T$. 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 A_i^t $(1 \leq i \leq n)$.

Definition 2: **Temporal Snapshot:**

Let $G = (V(G), A(G))$ be a directed acyclic graph (DAG) such that G is an I-map for Pr, the joint probability distribution defined on U^1. 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 Figure 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 3: **Dynamic Instance:**

Let G_t be as given in definition 2. G_t defines a dynamic instance of the QPN whose structure is defined by G and is given by $G_t = (V(G_t), \{A(G_t) \bigcup T(G_t)\})$ [2], where $V(G_t)$ and $A(G_t)$ are instants of $V(G)$ and $A(G)$ respectively at time slot t, and $T(G_t)$ describes the inter-slot conditional dependence between variables in $V(G_t)$ and its immediate neighbor $V(G_{t+1})$.

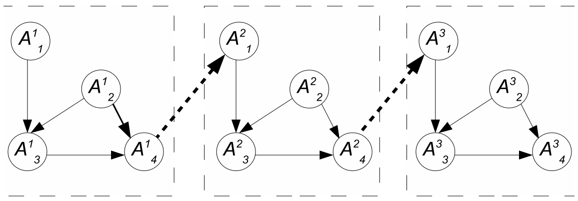

Fig. 2. An Example of G

Example 2: In the graph given in Figure 2, for each G_t, $V(G_t) = U_t$, $A(G_t) = \{(A_1^t, A_3^t), (A_2^t, A_3^t), (A_3^t, A_4^t), (A_2^t, A_4^t)\}$ and $T(G_t) = \{(A_4^t, A_1^{t+1})\}$.

Both of $A(G)$ and $T(G)$ encode a set of hyperarcs for G to capture a set of qualitative relations representing how variables influence each other. For this, we re-define the concept of a qualitative influence to capture not only within-slot relations, but also inter-slot ones. Before doing so however, we first present the definition of a Dynamic Qualitative Probabilistic Network (DQPN) below.

Definition 4: **Dynamic QPN:**

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

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

B. Qualitative Influences in a DQPN

Definition 5: **Positive DQPN Influence:**

Let G_t and G_{t+1} be two adjacent subgraphs of the DQPN defined over G. Further, let B and C be such that $B, C \in V(G)$. A direct positive influence is exerted by node B over node C, written as $S^+(B, C)$ iff for all values c_i^x of C and b_j^y, b_k^y of B with $b_i^y > b_k^y$, and for all integer values x and y such that $1 \leq x, y \leq m$ and $x - y \in \{0, 1\}$ we have:

$$Pr(C \geq c_i^x | b_j^y, w) \geq Pr(C \geq c_i^x | b_k^y, w)$$

Where w represents any combination of values for the set of nodes W which represent all other direct influences on C other than B. The superscripts x and y denote the temporal slot to which the instances c_i, b_j and b_k belong. The definition necessitates that variables can only directly influence other variables that belong to the same temporal slot $(x = y)$ or those that belong to the next immediate slot $(x - y = 1)$. Negative, zero and unknown influences are analogously defined.

In order to resolve the likely-to-occur ambiguities, we mimic the mechanisms given in [19] and define *indirect influences* that are augmented with two levels of strength and a multiplication index as given in Definition 6.

Definition 6: **Strongly Positive DQPN Influence:**

[1] G is the qualitative probabilistic network representing U

[2] For readability purposes, we will refer to $\{A(G_t) \bigcup T(G_t)\}$ as $Q(G_t)$ in this work.

Let B and C be two nodes in the DQPN defined over G. Furthermore, let tr be a trail from B to C. Let W be all the other nodes that can influence C and that do not belong to the trail from B to C. Then the qualitative influence from node B to node C along trail tr is strongly positive with multiplication index μ, $\mu \in \mathbb{N}$, written as $S^{++^{\mu}}(B, C, tr)$ iff for all values c_i^x of C and b_j^y, b_k^y of B with $b_i^y > b_k^y$

$$Pr(C \geq c_i^x | b_j^y, w) - Pr(C \geq c_i^x | b_k^y, w) \geq \alpha^{\mu}$$

Moreover, the qualitative influence of B on C along trail tr is weakly positive with multiplicative index μ, $\mu \in \mathbb{N}$, written as $S^{+^{+}}(B, C, tr)$ iff

$$0 \leq Pr(C \geq c_i^x | b_j^y, w) - Pr(C \geq c_i^x | b_k^y, w) \leq \alpha^{\mu}$$

Where w represents any combination of values from the set W and $x - y \in \{0, 1\}$. The value μ is given by the length of the trail tr and $\alpha = [0 - 1]$ is the cut-off value used for distinguishing between strong and weak influences and which can be chosen by an expert [3]. In addition to the cut-off value α which distinguish strong from weak influences, influences of the same strength can be compared using their μ value, where higher values indicate a longer trail tr, and as a result, a weaker influence [19].

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

IV. Defining Genetic Network Motifs Using DQPNs

Gene expression networks tend to be very complex with a large number of nodes and arcs connecting them. This has motivated studies that define simple patterns of interconnections between small groups of nodes. These patterns appear at high frequencies in naturally-occurring networks (including biological networks) and tend to increase in number monotonically as the size of the network increases. This is in contrast to synthetic, randomly-generated networks in which such patterns tend to sharply decrease in number as the size of the network grows [21]. Hence, these patterns define subgraphs that occurs at high frequencies in the network and which can serve as building blocks of the network. Such patterns have been termed *regulatory network motifs* [21] [13] and have been shown to carry significant information about the network's overall organization and functionality [11]. The motifs present a way of uncovering the structural design principles of gene expression networks is by breaking down their complex wiring into basic components.

[21] identifies three motifs that occur frequently in gene expression networks that have been shown to appear at frequencies greater 10 standard deviations greater than their mean number of appearances in randomized networks [21]. These motifs are the feed-forward in which a node X regulates another node Y such that they both regulate a third node Z, bi-fan motifs in which two nodes concurrently regulate two other nodes, and single-input module motifs which defines a set of nodes under the control of the same type of regulation (positive or negative) of one node, and are shown in figure 3.

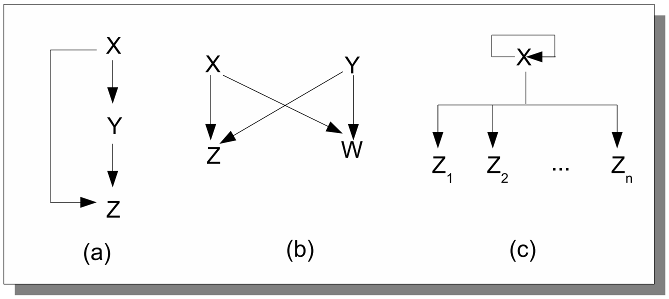

Fig. 3. (a) The Feed-forward loop motifs (b) The Bi-fan motif (c) The Single-input Module Motif.

If one to represent the gene-to-gene interactions in an expression experiment using a DQPN, where each subgraph G_t, $1 \leq t \leq m$ represents a snapshot of the genetic interactions of the cell during time slot T_t modeled by a QPN, then $A_1^t, ..., A_n^t \in U^t$ represents the expression levels of all the genes involved at slot T_t. In this context, a qualitative influence naturally corresponds to a regulatory relation between two nodes (genes). As a result, defining the motifs given in figure 3 is directly obtained from the construct of the DQPN as given in definitions 7 - 9 below.

Definition 7: **Feed-forward loops**

A feed-forward loop exists in a genetic network modeled by a DPQN defined over G iff for two subgraphs G_t and G_{t+1}[4]:

$$S^{\delta_1}(A_i^t, A_i^{t+1}, tr_1) \wedge S^{\delta_2}(A_i^t, A_i^{t+1}, tr_2), \text{ where } tr_1 \neq tr_2$$

Where $\delta_1, \delta_2 \in \{++, --, +, -, ?, 0\}$. The above definition states that a feed-forward loop exists on a variable (gene) A_i if it influences its own expression through two different trails (by stimulating different genes that will subsequently stimulate its expression). Bi-fans are similarly defined below.

Definition 8: **Bi-fans**

A bi-fan among four genes $A_a^t, A_b^t, A_c^{t+1} and A_d^{t+1}$ exists in a genetic network modeled by a DPQN defined over G iff for two subgraphs G_t and G_{t+1}

$$S^{\delta_1}(A_a^t, A_c^{t+1}, 1) \wedge S^{\delta_2}(A_b^t, A_c^{t+1}, 1) \wedge$$
$$S^{\delta_3}(A_a^t, A_d^{t+1}, 1) \wedge S^{\delta_4}(A_b^t, A_d^{t+1}, 1).$$

Where $\delta_1, \delta_2, \delta_3$ and $\delta_4 \in \{++, --, +, -, ?, 0\}$.

Definition 9: **Single Input Module (SIM)**

A SIM motif of a gene X_t on n other genes $A_1^{t+1}, ..., A_n^{t+1}$ exists in a genetic network modeled by a DPQN defined over G iff for two subgraphs G_t and G_{t+1}

$$S^{\delta}(X_t, A_1^{t+1}, 1) \wedge .... \wedge S^{\delta}(X_t, A_n^{t+1}, 1)$$

Where $\delta \in \{++, --, +, -, ?, 0\}$.

V. Experimental Results

A. Uncovering the Network Motifs Using QPNs

We conducted a set of experiments to verify the mapping between qualitative influences and the motifs formalized in

[3] The choice of α is part of our current experimental work.

[4] Note that only two time slots are sufficient for the definition of the loop as DQPNs naturally preserve the Markov property.

TABLE IV
NODES = NUMBER OF NODES (GENES) IN THE RUN, EDGE$_{avg}$ = AVERAGE NUMBER OF EDGES FOR 10 RUNS OF NETWORKS OF SIZE N FEED-FORWARD = AVERAGE NUMBER OF FEED-FORWARD MOTIFS FOR THE 10 RUNS, BI-FAN = AVERAGE NUMBER OF BI-FAN MOTIFS FOR THE 10 RUNS

Nodes	Edge$_{avg}$	Feed-forward	Bi-fan
85	154	16	209
185	372	18	430
285	518	21	825
385	698	29	1092
485	912	46	1437
585	997	52	1745

TABLE V
THE RESULT OF COMPARING THE ANALYSIS PROVIDED BY DBN$_{ZC}$ WITH THE SAME METHOD GUIDED WITH OUR QUALITATIVE NETWORK MOTIFS FOR THE YEAST TRANSCRIPTION DATASET COMPRISING OF 2875 GENES. THE ONLY PRIOR KNOWLEDGE INCLUDED IS THE KNOWLEDGE OF OUR QUALITATIVE MOTIFS AND NOTHING ABOUT THE YEAST CELL CYCLE IS GIVEN TO TEST THE HYPOTHESIS OF AN IMPROVED DETECTION OF REGULATOR-TARGET RELATIONS AND A BETTER CONSTRUCTION OF THE TARGET NETWORK. I = IDENTIFIED RELATIONSHIPS, M = MISIDENTIFIED RELATIONSHIP, S = SPECIFICITY

Method	I	M	S
DBN$_{ZC}$	17	3	9.8%
DBN$_{ZC}$ + Qualitative Motifs	26	2	10.7%

definitions 7, 8 and 9. The data set used for the purpose is based on the YPD (Yeast protein database) (S2) and was obtained from the data set used in [13] and contains 1079 interactions of 688 genes describing the regulation relationships of the transcriptional regulatory network of *Saccharomyces Cerevisiae*. The data comprises of three columns representing regulating genes, regulated genes and the mode of regulation. Not only that the number of motifs detected by our influences matches those of [13], but also upon retesting the hypothesis with differently-sized subsets of the data set, the number of motifs discovered by our influences was found to monotonically increase with the size of the data (as expected in real biological networks) as table IV shows.

The latter finding was achieved by constructing six additional experiments each testing the hypothesis for a subset of the full data set having a specific size. Each experiment consisted of 10 runs, all of the same size (number of nodes) but differ in connectivity (number of arcs). The algorithm describing the mapping of section III-B was tested on each of the 60 resulting runs and used to output the number of feed-forward loops and bi-fan motifs in each run. The results given in table IV visibly show the monotonic increase of the number of motifs with the number of nodes in the interaction data set.

B. The Second Experiment

The second set of experiments were conducted to build qualitative influences between genes by examining their expression levels, map the relevant influences to network motifs and use them to guide the construction of a DBN. The aim of the experiment is to assess the accuracy of the approach in recovering the structure of the DBN from the expression data with the aid of the discovered motifs by comparing it to the unguided DBN approach of [25].

For this experiment, we used the *Saccharomyces Cerevisiae* time series data from Choo et al [3], which contains data for ten time points. The first step was to examine the microarray data to investigate the strength of the various regulatory interactions by assigning each pair of genes a correlation coefficient γ capturing the degree to which two genes are co-expressed. We used cut-off values of $\gamma_+ \geq 1.2$ for a positive regulation and $\gamma_- \leq 0.7$ for a negative regulation to separate possible direct regulation from spurious interactions and used an approach similar to that of [25] to identify potential regulators and regulees. The cut-off values were chosen to match those of [25] for a controlled experiment.

We then designed an algorithm that reads through the collected pairs and their normalized expression levels and builds a database of qualitative influences that are detected by examining the genes pair-wise. We constructed an $M \times M$ matrix of influences exhibited among the genes. Each cell in the matrix is given a sign of either $+, -, 0, ?$. In our experiment, an unknown or a zero sign given in cell $m[i][j]$ designates a no correlation between the respective genes (at locations i and j). The mapping presented in section III-B is used to construct the set of feed-forward loop motifs discovered in the data.

The set of motifs constructed is then used as prior knowledge to guide the construction of the yeast gene regulatory network using [25]'s method, referred to in this work as DBN$_{ZC}$. We evaluated the method in terms of accuracy of the reconstructed network. More specifically, the guidance provided by the motifs discovered increased the specificity [5] as table V shows.

VI. CONCLUSIONS, CURRENT AND FUTURE WORK

This paper introduced, DQPNs, a formal model for capturing qualitative causal knowledge in time-series data. The model serves as a qualitative equivalent of Dynamic Bayesian Networks which uses signs to capture the direction of change of probabilities corresponding to the conditional probabilities of the original DBNs. The model makes use of the arc-based relations to introduce an efficient equivalent of DBN which captures the conditional independence relations the same way. We used the qualitative relations, namely *influences*, of DQPNs to model commonly-occurring motifs of gene regulatory network and showed a natural mapping between such motifs and DQPN influences. We evaluated the mapping via experiments which show that the regulatory networks motifs identified using the mapping we defined are equivalent to those identified in [13] for the same data set and that the motifs discovered via our formalism exhibit the expected property of increasing in number as the size of the regulatory network increases.

[5] Specificity is the percentage of correctly predicted known gene relationships out of the total number of predicted gene relationships.

Also, as an initial step to using DQPNs to recover the structure of gene regulatory networks, we adopted motifs captured from expression data to serve as representatives of the conditional independence relations in DBN graphs and used them to direct the reconstruction process of gene regulatory networks using DBNs. The result is an increased specificity and a decrease in the number of misidentified regulations.

We are currently working on the realization of a model for completely reconstructing gene regulatory networks using DQNs. We are at the stage of of incorporating time lags into the model and testing the hypothesis of 'the full specification of conditional probabilities is not necessary to reconstruct the regulatory relations in a gene regulatory network and only a subset of the quantitative data available is required. Because DQPNs deploy arc-based reasoning, they are expected to be much more efficient than their quantitative equivalents.

References

[1] Alizadeh, A. et al: *Distinct Types of Diffuse Large B-cell Lymphoma Identified by Gene Expression Profiling*. Nature 403:503-511 (2000)

[2] Bulashevka, S. et al: *Pathways of Urothelial Cancer Progression Suggested by Bayesian Network Analysis of Allelotyping Data*. International Journal of Cancer 110:850-856 (2004)

[3] Choo, R.J. et al: *A Genome-wide Transcirptional Analysis of the Mitotic Cell Cycle*. Molecular Cells 2(1): 65-73 (1998)

[4] Chickering, D. and Heckerman, D. and Meek,C.: *Large-Sample Learning of Bayesian Networks is NP-Hard*. The Journal of Machine Learning Research 5:1287-1330 (2004)

[5] D'haeseleer, P.: *Reconstructing Gene Networks from Large Scale Gene Expression Data*. Ph.D. dissertation, University of New Mexico (2000)

[6] Druzdzel, M. and Henrion, M.: *Efficient Reasoning in Qualitative Probabilistic Networks*. Proceedings of the AAAI National Conference on Artificial Intelligence: 548-553 (1993)

[7] Filkov, V. and Skiena, S. and Zhi, J.: *Analysis Techniques for Microarray Time-Series Data*. Journal of Computational Biology 9:317-330 (2002).

[8] Friedman, N.: *Inferring Cellular Networks Using Probabilistic Graphical Models*. Science 303:799-805 (2004)

[9] Guo, Y. et al: *How is mRNA Expression Predictive for Protein Expression? A Correlation Study on Human Circulating Monocytes*. Acta Bochimica et Biophysica Sinica 40(5):426-436 (2008)

[10] Liu,T. and Sung, W.: *Learning Gene Network Using Conditional Dependence*. Proceedings of the IEEE International Conference on Tools in Artificial Intelligence: 800-804 (2006)

[11] Hinman, V. and Nguyen, A. and Cameron, A. and Davidson, E.: *Developmental Gene Regulatory Network Architecture Across 500 Million Years of Echinoderm Evolution*. Proceedings of the National Academy of Sciences of the United States of America 100(23):13356-61 (2003)

[12] Hood, L. et al: *Systems Biology and New Technologies Enable Predictive and Preventive Meidcine*. Science 306:640-643 (2004)

[13] Milo, R. et al: *Network Motifs: Simple Building Blocks of Complex Networks*. Science 298: 824-827 (2002)

[14] Murphy, K. and Mian, S.: *Modelling Gene Expression Data using Dynamic Bayesian Networks.* Technical Report, University of California (1999)

[15] Noveen, A. and Hartsentein, V. and Chuong, C.M.: *Gene Nentworks and Supernetworks: Evolutionary Conserved Gene Interactions.* Molecular Basis of Epithelial Apppendage Morphogenesis: 371-391 (1998)

[16] Pisabarro, A. et al: *Genetic Networks for the Functional Study of Genomes*. Briefings in Functional Genomics and Proteomics 7(4):249-263 (2008)

[17] Renooij, S. and Parsons, S. and Pardieck, P.: *Using Kappas as Indicators of Strength in Qualitative Probabilistic Networks*. European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty: 87-99 (2003)

[18] Renooij, S. et al: *Pivotal Pruning of Tradeoffs in Qualitative Probabilistic Networks*. International Conference on Uncertainty in Artificial Intelligence: 515-522 (2000)

[19] Renooij, S. and Van der Gaag, L.: *Enhanced Qualitative Probabilistic Networks for Resolving Trade-offs*. Artificial Intelligence 172(12-13): 1470-1494 (2008)

[20] Segal, E. and Friedman, N. and Koller, D. and Regev, A.: *A Module Map Showing Conditional Activity of Expression Modules in Cancer*. Nature Genetics 36:1090-1098 (2004)

[21] Shen-Orr, S. and Milo, R. and Mangan, S. and Alon,U.: *Network motifs in the Transcriptional Regulation Network of Escherichia Coli*. Nature Genetics 31: 64-68 (2002)

[22] Wellman, M.: *Fundamental Concepts of Qualitative Probabilistic Networks*. Artificial Intelligence 44:357-303 (1990)

[23] Wessels, L. and Someren, E. and Reinders, M.: *A Comparison of Genetic Network Models*. Pacific Symposium on Biocomputing (PSB) 6:508-519 (2001)

[24] Zhang, Y. et al: *Inferring Gene Regulatory Networks from Multiple Data Sources Via a Dyanamic Bayesian Network with Structural EM*. Data Integration in the Life Sciences: 204-214 (2007)

[25] Zou, M. and Conzen, S.: *A New Dynamic Bayesian netowrk (DBN) Approach for Identifying Gene Regulatory networks from Time Course Microarray Data*. Bioinformatics 2(1):71-70 (2005)