Gene expression

**Literature-based priors for gene regulatory networks**

E. Steele¹*, A. Tucker¹, P.A.C. ‘t Hoen² and M.J. Schuemie²

¹Centre for Intelligent Data Analysis, School of Information Systems, Computing and Mathematics, Brunel University, Uxbridge UB8 3PH, UK, ²Biosemantics Association, a collaboration between the Department of Human Genetics from the Leiden University Medical Center, Leiden and Department of Medical Informatics from the Erasmus University Medical Center, Rotterdam, The Netherlands

Received on October 31, 2008; revised on March 23, 2009; accepted on April 20, 2009

Advance Access publication April 23, 2009

Associate Editor: Joaquin Dopazo

ABSTRACT

**Motivation:** The use of prior knowledge to improve gene regulatory network modelling has often been proposed. In this article we present the first research on the massive incorporation of prior knowledge from literature for Bayesian network learning of gene networks. As the publication rate of scientific papers grows, updating online databases, which have been proposed as potential prior knowledge in past research, becomes increasingly challenging. The novelty of our approach lies in the use of gene-pair association scores that describe the overlap in the contexts in which the genes are mentioned, generated from a large database of scientific literature, harnessing the information contained in a huge number of documents into a simple, clear format.

**Results:** We present a method to transform such literature-based gene association scores to network prior probabilities, and apply it to learn gene sub-networks for yeast, Escherichia coli and Human organisms. We also investigate the effect of weighting the influence of the prior knowledge. Our findings show that literature-based priors can improve both the number of true regulatory interactions present in the network and the accuracy of expression value prediction on genes, in comparison to a network learnt solely from expression data. Networks learnt with priors also show an improved biological interpretation, with identified sub-networks that coincide with known biological pathways.

**Contact:** emma.steele@brunel.ac.uk

**Supplementary information:** Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Gene regulatory networks (GRNs) represent how genes interact in various cellular processes by describing how the expression level, or activity, of genes in turn affect the expression of the other genes. Reverse-engineering GRN models can help biologists understand and gain insight into genetic conditions and diseases, and subsequently this has become a rapidly expanding area of Bioinformatics research (Friedman et al., 2000; Hartemink et al., 2002; Segal et al., 2004). However, whilst the microarray provides the most available genome-wide data source on gene expression, there are concerns over its reliability and the reproducibility of results across microarray platforms or laboratories (Tan et al., 2003; MAQC consortium, 2006). Microarray data are subject to both biological variations across samples and experimental noise, which may be introduced throughout the stages of the experiment. In addition, a single microarray dataset usually contains a large number of genes (commonly thousands), but the number of samples is much lower, making it difficult to extract reliable regulatory interactions. However, these drawbacks can be alleviated by incorporating other complementary data sources into the modelling process.

In this article we present the first research on the massive incorporation of prior knowledge from literature for the construction of Bayesian network (BN)-based GRN models. We use a comprehensive collection of literature and show that this content helps to improve the modelling process. The ability to use such a large body of prior knowledge lies in the use of advanced biological text-mining techniques such as literature-based gene concept profiling (Jelier et al., 2007; Schuemie et al., 2007a), which allow a measure of association between a pair of genes to be calculated based not only on the co-occurrence of entities in the same document, but also on indirect relations, where genes are linked via a number of documents. An association matrix for gene-pairs can be generated, where each entry represents the strength of the relationship between genes, based on a database of scientific literature.

Bayesian networks (BNs) (Pearl, 1991) have become a popular method for computational modelling of GRNs from expression data, since they are able to represent the network both graphically and quantitatively (probability distributions quantify the strength of dependencies between the nodes in the network) and thus are relatively easy to interpret by non-statisticians (e.g. biologists). They are also provided with a natural mechanism for incorporating prior knowledge relating to the network structure through informative structure priors (Castelo andSiebes, 2000), which we make use of in this research. We build on existing techniques by presenting a methodology for deriving a prior probability distribution over candidate network structures based on literature-based gene association matrices and investigate how weighting the influence of the prior knowledge can affect the learnt network structure.

In the most notable comparable work, Imoto et al. (2003) use energy functions to incorporate prior knowledge sources into Bayesian GRN models and propose the incorporation of many types of different prior knowledge, including literature-based knowledge extracted from regulatory interactions that are recorded in the Yeast Proteome Database (YPD). Later, Werhli and Husmeier (2007)
extended the approach of Imoto et al. to multiple sources of prior knowledge and applied their approach on combining protein–protein interactions and KEGG pathways with expression data. Our use of advanced text-mining techniques provides an advantage over this research, since databases such as YPD and KEGG rely on the addition of manual annotations and as the volume of scientific publications becomes prohibitively large, keeping up-to-date information within them becomes increasingly challenging.

Informative structure priors have previously been used to incorporate prior knowledge into BN-based GRN models. For example, Bernard and Hartemink (2005) use the technique to incorporate transcription factor (TF) binding site data location. However, it has not been applied with the type of literature-based information used in this research, and to our knowledge weighting the score from the prior has not been addressed.

The remainder of the article is organized as follows. In Section 2, we describe the informative priors methodology and literature-based concept profiling in more detail. Section 3 details our results on three real sub-networks. Finally, in Section 4 we discuss our findings and outline directions for future research.

2 METHODS

2.1 Bayesian networks

BNs are graph-based models of probability distributions that capture properties of conditional independence between variables. A BN consists of two components—a directed acyclic graph (DAG) consisting of links between nodes that represent variables in the domain, and a set of conditional probability distributions associated with each node. In GRN modelling, the expression level of genes are represented by nodes in the network graph, whilst dependencies between genes are represented by the directed edges.

We use a score-based search method to learn a BN that represents a GRN from microarray expression data. A search is conducted through the space of possible networks, scoring each structure, to identify the network with the maximum score. A variety of search strategies can be used, the simplest being a greedy hill-climb. We use a simulated annealing approach in order to limit local maxima. The search begins with an empty network. At each stage of the search, the networks in the current neighbourhood are found by applying operators such as add arc, remove arc and reverse arc to the current network.

To score candidate networks, we use the Bayesian information criterion (BIC), which is a combination of the model log-likelihood and a penalty term that favours less complex models:

\[
\text{BIC} = \log P(S) + \log P(S|D) - 0.5k \log(n)
\]

where \(S\) represents the model, \(D\) is the data, \(n\) is the number of observations (sample size) and \(k\) is the number of parameters. \(\log P(S)\) is the prior probability of the network model \(S\), \(\log P(S|D)\) is the log-likelihood, while the term \(0.5k \log(n)\) is a penalty term, which specifically penalizes more complex models with more parameters. The BIC is good for dealing with small samples of data as is common with microarray data, as the penalty term helps to prevent overfitting.

Several DAGs may represent the same set of dependencies amongst variables. A set of such DAGs belong to the same equivalence class. Equivalent graphs agree on the same underlying undirected structure, but the direction of some edges may vary (Pearl and Verma, 1991). Thus, an equivalence class of a set of DAGs can be represented using a partially DAG (PDAG), where only some edges are directed. Chickering (1995) derived an algorithm for constructing the PDAG representing the equivalence class for any DAG; we use this to convert the learnt BNs to their equivalence classes.

Since the search-and-score method can lead to a different learnt graph on each run, Friedman et al. (1999) devised a method for computing the statistical confidence of features within a BN, based on a well-known statistical method, Efron’s Bootstrap (Efron and Tibshirani, 1993). Given a dataset \(D\) containing \(N\) observations, a new dataset is created by resampling \(N\) times, with replacement from \(D\). This process is repeated \(m\) times, resulting in \(m\) learnt BNs. An estimate of confidence for each feature is computed by the proportion of networks that contain that feature. We make use of the bootstrapping method to generate more robust network structures, defining a feature as the existence of an edge between two nodes in the network. Thus, the resulting bootstrapped network has a confidence estimate assigned to each network edge. A PDAG from a bootstrapped network by thresholding. If an edge has a confidence above the threshold, it is included in the PDAG (and if edges are found in both directions—e.g. from node \(i \rightarrow j\) and \(j \leftarrow i\), then the edge is undirected).

2.2 Informative literature-based priors

Usually the prior probability \(\log P(S)\) in the BIC score is chosen to be uninformative—that is, it is a uniform prior, where every network structure is equally likely. In this research, we consider the use of an informative prior (i.e. which is not uniformly distributed over each possible network structure) based on knowledge contained in the scientific literature.

2.2.1 Calculating the prior probability of a network structure

The prior probability distribution for candidate network structures assigns each network a probability such that all probabilities sum to 1. However, enumerating all possible structures is infeasible in most cases. A more intuitive method to calculate the prior probability is by using an edge decomposition technique developed by Castelo and Siebes (2000). This method is based on using probabilities for the existence of each edge, which are provided by an expert or based on some prior knowledge. If \(a\) and \(b\) are two nodes in a network where \(B\) is the expert prior knowledge then:

\[
p(a \rightarrow b | B) + p(a \leftarrow b | B) + p(a \leftrightarrow b | B) = 1
\]

where \(a \rightarrow b, a \leftarrow b\) indicate directed edges and \(a \leftrightarrow b\) indicates that an edge exists but \(b\) does not exist. In other words, the probabilities of the edge existing in either direction or not existing at all sum to 1. Making the assumption that the prior information on the existence of each edge are independent of one another, we can multiply the probabilities together to obtain the probability of the whole network \(G\), such that:

\[
P(G|B) = \prod_{(v_i \rightarrow v_j) \in G} p(v_i \rightarrow v_j | B) \prod_{(v_i \leftarrow v_j) \in G} p(v_i \leftarrow v_j | B)
\]

where \(v_i\) and \(v_j\) are nodes in \(G\), \(v_i \rightarrow v_j\) represents an edge between \(v_i\) and \(v_j\) in either direction and \(v_i \leftarrow v_j\) represents an edge that does not exist.

In other words, the prior probability of the whole network \(G\) is formed by multiplying the probabilities for each edge in \(G\) to exist, and for each edge not to exist. We note that in our case this assumption may be violated if the different sources of the prior information are not independent, which can lead to edge probabilities becoming dependent. For example, if one edge probability is high, another related edge may have a higher probability. This could lead to the introduction of surplus false positive (FP) edges into the learnt networks. However, as noted by Castelo and Siebes, the assumption of independence ensures that the calculation of the prior probability distribution is tractable. Then, taking logs:

\[
\log P(G|B) = \sum_{(v_i \rightarrow v_j) \in G} \log(p(v_i \rightarrow v_j | B)) + \sum_{(v_i \leftarrow v_j) \in G} \log(p(v_i \leftarrow v_j | B))
\]

Thus, by the edge decomposition method, the log prior probability of a network \(G\) (as required for the BIC score of a network) can be calculated by summing the log prior probabilities that each edge the network \(G\) contains does exist and that each edge not present in the network \(G\) does not exist.

2.2.2 Weighting the prior

The influence of the prior can be varied by including a weight in the score calculation:

\[
\text{BIC} = w \log P(S) + \log P(S|D) - 0.5k \log(n)
\]
where \( w \) is between 0 and 1 and referred to as the prior weight. A weight of 0 corresponds to a uniform (uninformative) prior, whilst a weight of 1 includes the full log prior probability in the score calculation.

2.2.3 Literature-based gene concept profiling Information about biomedical concepts such as genes can be summarized using a technique known as concept profiling (Jelier et al., 2007; Schuemie et al., 2007a). The technique uses a thesaurus containing concepts; we use a combination of the UMLS Metathesaurus (McCray and Miller, 1998) and our own gene thesaurus, constructed by combining information from several databases, including Entrez Gene, Uniprot, and the Saccharomyces Genome Database.

The concept recognition software Peregrine, which also disambiguates homonyms (Schuemie et al., 2007b), can detect occurrences of thesaurus-concepts in Medline articles published after 1980, resulting in a list of occurrences of concepts. For each concept occurrence, concept profiles are constructed. A concept profile itself contains concepts: it is a vector of concepts with weights, where the weight describes the strength of the association between the concept and the concept to which the profile belongs. The weight is the uncertainty between the occurrence of one concept and the occurrence of the other concept (Jelier et al., 2008); it expresses the relative amount of information gained about whether concept \( X \) occurs in document \( D \) by knowing that concept \( Y \) occurs in document \( D \). Since each gene is a concept in our thesaurus, we can construct concept profiles for all genes.

Since concept profiles are weight vectors, we can calculate Pearson’s correlation coefficient for each pair of genes and subsequently a gene correlation matrix, which is based on knowledge contained in the literature. Interestingly, we can even calculate the correlation between genes that have never been mentioned together before, based on shared concepts in the respective concept profiles.

The resulting correlation matrix provides an indication of whether a set of genes are closely related. However, we should note that this correlation score only provides information about relationships between genes in the broadest sense. For example, it does not indicate whether a certain regulatory relationship only holds under certain conditions. To solve such ambiguities is beyond the capabilities of current natural language processing techniques. By using homonym disambiguation in our concept recognition process, we generate the most reliable correlation scores currently possible, but still these scores should be taken as indications of probabilities, not as proven biological facts.

2.2.4 Edge prior probabilities from literature-based knowledge We estimate the probabilities for the existence of each edge based on the gene-pair correlation matrix obtained from the literature. The correlation value provides a measure of whether two genes are related according to the literature. We translate this correlation value into a probability that an edge exists between that gene pair, based on the idea that a higher correlation translates to a higher probability. We adopt an approach based on the statistical confidence of the correlation values. This is because even correlation values that are small (in absolute terms) may represent a significant association, if they are relatively higher than the majority of values. We first generate the distribution of correlation values for a random selection of gene pairs and fit this to a normal distribution. For each gene-pair correlation value of interest, its associated \( P \)-value, \( p \), is calculated based on how far the correlation value deviates from the fitted normal distribution (based on a two-tailed confidence interval) so outlying correlations are more significant. Finally, we define the probability for an edge existing between the gene pair as \( P(\text{edge exists})=1-p \). Please note that this probability is in fact the probability that a particular correlation was not drawn from the distribution of random gene-pair correlations, and is therefore not equal to the probability that the edge exists. Although more advanced methods exist for deriving a more accurate probability estimation, such as those in Segal et al. (2002) and Efron (2007), we have found that this simple method already generates useful priors that reflect the information within the literature.

2.3 Experiments and evaluation To evaluate the performance of networks learnt with a literature prior compared with those networks learnt from expression data alone, we compare the regulatory relationships found in the networks: TFs and the target genes that they regulate (i.e. the child genes of TFs in the networks). We use microarray expression datasets from three different organisms: yeast, *Escherichia coli* and Human. For each organism, a gene correlation matrix was constructed based on all abstracts contained in Medline. The expression values were discretized into three states using an equal frequency based method (i.e. for each gene, one-third of values are categorized as ‘low’, one-third as ‘normal’ and one-third as ‘high’). Discretization is a technique commonly applied to real gene expression data for its use to model GRNs (Friedman et al., 2000; Hartemink et al., 2002). It is of particular benefit with real datasets that have a small number of samples and/or can be noisy. The method still allows complex regulatory relationships to be modelled, whilst avoiding the need to deal with parameterized continuous distributions.

For each organism, hierarchical clustering was used to identify groups of related genes in the literature correlation matrix, and a subgroup of genes was formed by combining the clusters that contained TFs. Each subgroup contained 200–300 genes, which allowed the inference of a larger scale network, whilst maintaining the efficiency of learning. Selecting genes from literature-based clusters that include known TFs increases the occurrence of ‘true’ (documented) regulatory relationships within the subgroup of genes, since these will only occur with known TFs as parent genes. Due to the large number of genes in the selected subgroup, we restricted the network structure during learning to only include edges where parent nodes are documented TFs. This meant that each edge in the learnt network had the potential to be documented as a regulatory relationship (thus removing the possibility of unavoidable (FPs) that would occur when a parent node is not a documented TF).

For each organism, seven bootstrapped networks were learnt: the expression data network (with prior weight 0) and six posterior networks, learnt with the prior weight at 0.2, 0.4, 0.5, 0.6, 0.8 and 1, respectively. We refer to the following terminology throughout the article: A network learnt from microarray expression data with a literature-based prior \( \text{prior weight } > 0 \) is the posterior network. A network learnt from expression data only \( \text{(i.e. a prior weight of 0) is referred to as the expression data network.} \)

Previous research on learning GRNs has often evaluated learnt networks by comparing them to documented gene interactions that form a ‘true network’ (usually compiled from online databases of confirmed regulatory interactions) in terms of true positives (TPs) and FP. A TP is an edge that is present in both the learnt and true networks. A FP is an edge that is present in the learnt network but not in the true network. However, since no negative relationships are actually defined in the literature, we cannot be sure that FP edges are actually not regulatory interactions. Additionally, we note that this type of comparison should be treated with some caution since information distilled into a database containing regulatory interactions essentially comes from the literature, which is also where our prior information comes from. However, a comparison to documented interactions can still assist us in measuring the effect of using a prior to learn the network.

Since we actually learn bootstrapped networks, where a confidence is attached to each edge, we threshold the network edges at values from 0 to 1 (at steps of 0.01) to form a collection of network graphs and compare each one with the true network. It is then possible to plot the TP and FP rates for each thresholded network, forming a receiver operator characteristic (ROC) curve, which allows one to view the performance of a classifier graphically. A global measure of the classifier performance, often used in classification problems, is the area under the ROC curve (AUC). AUC is a value between 0 and 1. We use the AUC to compare the networks generated with different prior weights. In general, the closer the AUC is to 1 the better the overall performance of the network.
Another method of network analysis is the prediction of gene expression values on an independent dataset. If a gene node in a posterior network has increased accuracy in prediction on an independent dataset than the same gene node in an expression data network, we can say that the prior does add value to the learnt network. To predict gene expression values, we estimate the conditional probability distributions using the same expression dataset from which the structure was learnt, using a maximum likelihood parameter estimation method (since we have complete data). Using the parameterized structure, we can then predict the (discretized) expression value of each gene, based on the expression values of its influencing genes in the network, over samples from unseen independent datasets. We can measure the success of gene expression prediction using prediction accuracy, which is the proportion of samples where the prediction is correct.

In order to measure the statistical significance of the differences between TF prediction accuracies across networks with different prior weights, we run the bootstrap learning process several times for each dataset and apply the Cochran–Mantel–Haenszel (CMH) test (McDonald, 2008), which can be used for repeated tests of independence. The objective is to establish whether two variables are independent, conditional on a third variable that identifies the repeat tests. The null hypothesis is that the two variables are independent of each other within each repetition—that having one value of one variable does not mean it is more likely to have one value of the second variable, or in other words that there is no significant difference between the two variables. In this case, we wish to establish whether the prediction performances of two different networks are independent where each TF identifies the repeat tests. Therefore, in order to obtain a significant result, the null hypothesis should be rejected in favor of the alternate hypothesis, that there is a significant difference between the prediction performances of the TFs in each network.

3 RESULTS

3.1 Yeast

For yeast, we based network learning on cell-cycle expression data (Spellman et al., 1998) for a group of 204 genes. Out of the 204 genes, 22 are identified as TFs in the YeastRact database (Teixeira et al., 2006), which lists documented regulatory interactions in yeast. Figure 1 shows the AUC for the networks learnt with each prior weight. The expression data network (a prior weight of 0) obtains an AUC of 0.56. As the prior weight increases, so does the AUC, up to the prior weight of 0.8, where the AUC peaks at 0.65. Supplementary Figure 1 compares the documented links (TPs) in the expression data-only network with those in the network learnt with prior weight 0.8, which has the highest AUC. At a prior weight of 1 the AUC dips slightly to just under 0.65. (Recall that the prior weight of 1 indicates that the prior knowledge is fully weighted in the score during learning.) This indicates that including the prior does add knowledge to the learnt network, but a balance between the literature and expression data is required.

Expression values were predicted for all TFs in each network (different prior weights) on an unseen cell-cycle expression dataset (Pramila et al., 2006). In general, we found that TFs in the posterior networks obtained higher predictive accuracies than the same TFs in the expression data network (Table 1). Using the CMH test across all TFs, the posterior networks attain significantly higher accuracies with $P < 0.002$. TFs in the posterior network learnt with a prior weight of 0.6 exhibit the most significant difference to the expression data network, with $P = 0.0001$. The prediction accuracies for each TF in the networks learnt with prior weights of 0 and 0.6, respectively are shown in Figure 2. The expression value prediction and comparison to documented interactions (Fig. 1) show the same networks performing well—the network learnt with prior weight 0.6 also gains a high AUC value, although it is not the absolute maximum.

3.2 Escherichia coli

For our second set of experiments, we considered network learning on E. coli expression data (Sangurdekar et al., 2006). This dataset records transcriptional responses to more than 30 chemical and physiological perturbations. The selected group of 262 genes contains 17 TFs, according to the RegulonDB online database of E. coli regulatory interactions (Salgado et al., 2006).

Figure 3 shows the AUC for the networks learnt with each prior weight. The pattern exhibited is very similar to the yeast results. The expression data network (a prior weight of 0) obtains an AUC of 0.67. As the prior weight increases, so does the AUC, up to the prior weight of 0.8, where the AUC peaks at 0.81. At a prior weight of 1 the AUC dips slightly to just under 0.8. This shows that the use of the prior adds to the network many edges that represent confirmed interactions, but are not represented in the expression data (Supplementary Fig. 2 compares the confirmed interactions in the expression data-only network with those in the network learnt with prior weight 0.8, which has the highest AUC). This may be because the expression data focuses on a particular type of experiment—chemical and physiological perturbations. It is worth noting that E. coli is a particularly well-studied organism and this may contribute towards such improvements in network performance in terms of AUC—there is more literature available, which is also more reliable and accurate, than for a less-studied organism.

Expression values were predicted for all TFs in each network (different prior weights) on an unseen cell-cycle expression dataset (Faith et al., 2007), which contains a wide range of different experiments with over 250 samples. This dataset does not contain data for all genes in the subset, so for each TF we only make predictions using the bootstrap samples where there are no target genes with missing data. Where there are no bootstrap samples without missing target gene data the TF is not included. Thus, prediction accuracies are compared over 11 TFs. Table 1 details the average TF prediction accuracies for each posterior network in comparison to the expression data-only network, and the corresponding $P$-value calculated using the
Table 1. Yeast and E. coli expression value prediction results

<table>
<thead>
<tr>
<th>Posterior network (prior weight)</th>
<th>Yeast</th>
<th>E. coli</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average TF prediction accuracy</td>
<td>P-value from CMH-test</td>
</tr>
<tr>
<td></td>
<td>Expr. only network</td>
<td>Posterior network</td>
</tr>
<tr>
<td>0.2</td>
<td>0.409</td>
<td>0.441</td>
</tr>
<tr>
<td>0.4</td>
<td>0.409</td>
<td>0.463</td>
</tr>
<tr>
<td>0.5</td>
<td>0.409</td>
<td>0.442</td>
</tr>
<tr>
<td>0.6</td>
<td>0.409</td>
<td>0.459</td>
</tr>
<tr>
<td>0.8</td>
<td>0.409</td>
<td>0.446</td>
</tr>
<tr>
<td>1</td>
<td>0.409</td>
<td>0.456</td>
</tr>
</tbody>
</table>

This table compares the expression data-only (prior weight 0) network with each posterior network (prior weights 0.2–1), through the average TF prediction accuracies and the significance of the CMH test.

Fig. 2. Comparison of expression value prediction between the yeast networks learnt with prior weight 0 and 0.6 for the TF genes.

Fig. 3. Comparison of AUC values for each E. coli network generated with a different prior weight from 0 (no prior) to 1 (fully weighted prior).

Fig. 4. Comparison of expression value prediction over TF genes, between the E. coli networks learnt with prior weights of 0 and 1.

CMH test. There is an increase in TF predictive accuracies in the posterior networks, and this is a significant increase \((P \leq 0.05)\) for four of the posterior networks. In particular, the posterior network generated with prior weight 1 (full prior weighting) shows the most significant difference \((P = 0.00001)\). Figure 4 plots the prediction accuracies for each TF in the networks learnt with prior weights of 0 and 1.

3.3 Human

To demonstrate that our approach is useful for the modelling of GRNs in higher eukaryotes, we test our methodology on a third dataset concerning muscle differentiation studied in cultures of primary human muscle precursor cells. This dataset consists of expression time series (7 time points between 0 and 14 days of differentiation) of in vitro muscle differentiation has been generated for six human individuals: three healthy and three patients with Duchenne muscular dystrophy (DMD), which are known to display certain differentiation defects (Sterrenburg et al., 2006). For this biological system, there is no comprehensive database of regulatory relationships available, and no other suitable microarray expression datasets for evaluating prediction accuracy. Instead, we comment on the learnt networks and their biological interpretation.
In particular, we considered the differences between the posterior networks generated using different prior weights. As expected, the effect of increasing the prior weight was the inclusion of more edges supported by the literature. Our hypothesis was that spurious edges, i.e. edges between genes that exhibit a pattern of co-regulation (e.g. correlated expression patterns) but have no regulatory relationship, would disappear from the network as the prior knowledge gained more weight in the learning process. This indeed is what we observe. Supplementary Figure 3 shows the networks learnt with each prior weight. In particular, in the networks with a prior weight of 0.4 or greater, we can identify subnetworks that are related to certain biological functions such as cell-cycle control. The cell-cycle network is relevant to the biological system under study since cell cycle arrest is one of the first steps in myoblast differentiation. In the network with prior weight of 0.4, CCNH, which belongs to the family of cyclins that are involved in cell-cycle control, forms a central node with several daughter genes (CDKN1B, CDK4, CDK7, CDK9, CCND3, FRAP1, MDM2). Fewer of these edges are present in the expression data-only network (prior weight 0). There is also literature support for the regulatory relationship between CCNH and MDM2 and CDKN1B (Datta, 2002; Mandalb et al., 1998). Supplementary Figure 4 highlights the target genes of CCNH in a sub-network for each of the prior weights.

NCO2, a nuclear co-receptor that inhibits muscle differentiation, is another central node in the network with prior weight 0.4. Consistent with literature (Bailey et al., 1999), NCO2 demonstrates reduced expression during differentiation, in particular in DMD myotubes. In the network with prior weight 0.4, there are links visible between NCO2, SKIP and the histone deacetylases (HDAC3, PCAF) that are known to work together in the acetylation of the important muscle TFs MYOD1 and MEF2 (Gregoire et al., 2007). As before, this presumably central role of NCO2 and the histone deacetylation pathway is only evident upon incorporation of the literature prior.

Other possible gene relationships for which there is literature support are not present in the learnt networks. However, this is expected since the networks are learnt from expression data from a highly specific biological system in which not all possible relationships will be existent and also since not all literature-derived relationships are of regulatory nature.

From the manual inspection of networks, we believe that a literature prior weight of between 0.4 and 0.6 is the best choice to identify relevant regulatory edges. Higher prior weights appear to lead to the inclusion of too many edges due to literature associations that are not of regulatory nature. This is lower than the optimum prior weights found for the yeast and E.coli networks, which were both above 0.6. This may be because there is less literature on the human organism, whereas yeast and E.coli are both well-studied. Where there is less literature, it may be less reliable, so more weight needs to be assigned to the microarray expression data.

4 DISCUSSION

This article has presented some of the first research on the incorporation of prior knowledge from a large body of relevant literature for BN learning of gene networks. Using advanced text-mining techniques, information from a collection of documents can be represented using a correlation matrix. We make use of an informative prior probability distribution over BN structures, the natural mechanism for incorporating prior knowledge into BN learning, together with an existing method for computing the probability of a network structure using edge-wide decomposition. Building on these existing techniques, this article has contributed a method for translating literature-based gene-pair correlations to network edge prior probabilities and investigated the effects of weighting the influence of the prior knowledge during learning.

Based on our experimental results, we found that the posterior networks (those learnt with prior knowledge) were closer to a ‘true’ network of documented regulatory interactions than the expression data network. In yeast and E.coli, the posterior network with a prior weight of 0.8 gave AUCs of 0.65 and 0.81, improvements over the expression data networks of 0.56 and 0.67, respectively. This shows that the prior can add many edges that represent confirmed interactions, but are not exhibited in the expression data. Microarray experiments are often focused on a particular subsystem of genes, so prior knowledge can assist in ‘filling the gaps’ for genes that are not the particular experimental focus of the expression dataset. Experiments on the human dataset also provided evidence that incorporation of literature information results in removal of spurious correlations and generates networks containing modules of functionally related genes.

In yeast and E.coli the comparison of TF predictive accuracies was used as another method of network validation. In general, the predictive accuracy improved in the posterior networks with a prior weight of 0.8 gave AUCs of 0.65 and 0.81, improvements over the expression data networks of 0.56 and 0.67, respectively. This indicates that the posterior network structure is more robust for making predictions outside of the original dataset.

Proper weighting of the literature information appears to be needed—whilst the inclusion of the prior information helps to reduce spurious regulatory relationships, higher prior weights can lead to the inclusion of too many edges due to literature associations that are not of regulatory nature (e.g. proteins in the same multi-protein complex). In addition, the optimal prior weight may be related to the amount of literature available. The optimal prior weight for the Human network was lower than for yeast and E.coli, which are well-studied organisms with a lot of related literature. Where there is less literature, more weight may need to be assigned to the microarray expression data.

Comparable related work (Bernard and Hartemink, 2005; Inoto et al., 2003; Werhli and Husmeier, 2007) in incorporating prior knowledge into BN-based GRN models has concentrated on developing the methodology and less on the biological content of prior knowledge. Where literature-based knowledge sources were used, these were based on databases such as KEGG. Our use of advanced text-mining techniques provides a powerful advantage over this research, as it allows up-to-date information from a huge amount of literature to be used. In addition, in previous research where real datasets have been used for evaluation, typically this has been on a small scale using <40 genes in total. In this work, we have performed evaluation on networks of 200–250 genes, for three different organisms. One potential method to improve upon our abstract literature search would be to search the full text in Medline as has been done in Natarajan et al. (2006).

Further work will involve an investigation on optimising the prior weight and extending the modelling techniques. For example, temporal information can be incorporated through the use of dynamic BNs, which will allow cyclic behaviour to be introduced. This research is also part of a wider project on combining multiple
sources of heterogeneous data for GRN modelling. As part of this, other future research in this area will include combining prior knowledge with multiple sources of expression data.

**Funding:** Royal Society (grant 2006/R3); VENI-grant from the Dutch Organization for Scientific Research (NWO grant 2005/03808/ALW to P.A.C.’IH).

**Conflict of Interest:** none declared.

**REFERENCES**


