Cover Page

Universiteit Leiden

The handle http://hdl.handle.net/1887/19044 holds various files of this Leiden University dissertation.

Author: Anvar, Seyed Yahya
Title: Converging models for transcriptome studies of human diseases : the case of oculopharyngeal muscular dystrophy
Issue Date: 2012-06-06

NETWORKS IN BIOLOGY
beyond differential expression

PART TWO

The identification of informative genes from multiple datasets with increasing complexity

Seyed Yahya Anvar[1,2,*], Peter A.C. 't Hoen[2] and Allan Tucker[1]

In microarray data analysis, factors such as data quality, biological variation, and the increasingly multi-layered nature of more complex biological systems complicates the modelling of regulatory networks that can represent and capture the interactions among genes. We believe that the use of multiple datasets derived from related biological systems leads to more robust models. Therefore, we developed a novel framework for modelling regulatory networks that involves training and evaluation on independent datasets. Our approach includes the following steps: (1) ordering the datasets based on their level of noise and informativeness; (2) selection of a Bayesian classifier with an appropriate level of complexity by evaluation of predictive performance on independent data sets; (3) comparing the different gene selections and the influence of increasing the model complexity; (4) functional analysis of the informative genes. In this paper, we identify the most appropriate model complexity using cross-validation and independent test set validation for predicting gene expression in three published datasets related to myogenesis and muscle differentiation. Furthermore, we demonstrate that models trained on simpler datasets can be used to identify interactions among genes and select the most informative. We also show that these models can explain the myogenesis-related genes (genes of interest) significantly better than others ($P < 0.004$) since the improvement in their rankings is much more pronounced. Finally, after further evaluating our results on synthetic datasets, we show that our approach outperforms a concordance method by Lai *et al.* in identifying informative genes from multiple datasets with increasing complexity whilst additionally modelling the interaction between genes. We show that Bayesian networks derived from simpler controlled systems have better performance than those trained on datasets from more complex biological systems. Further, we present that highly predictive and consistent genes, from the pool of differentially expressed genes, across independent datasets are more likely to be fundamentally involved in the biological process under study. We conclude that networks trained on simpler controlled systems, such as *in vitro* experiments, can be used to model and capture interactions among genes in more complex datasets, such as *in vivo* experiments, where these interactions would otherwise be concealed by a multitude of other ongoing events.

1 Center for Intelligent Data Analysis, School of Information Systems, Computing and Mathematics, Brunel University, Uxbridge, Middlesex, UB8 3PH, UK **2** Center for Human and Clinical Genetics, Leiden University Medical Center, the Netherlands.

* To whom correspondence should be addressed at: s.y.anvar@lumc.nl

BACKGROUND

High-throughput gene expression profiling experiments have increased our understanding of the regulation of biological processes at the transcriptional level. In bacteria (Bockhorst et al., 2003) and lower eukaryotes, such as yeast (Segal et al., 2003), modeling of regulatory interactions between large numbers of proteins in the form of regulatory networks has been successful. A regulatory network represents relationships between genes and describes how the expression level, or activity, of genes can affect the expression of other genes. The network includes causal relationships where the protein product of a gene (e.g. transcription factor) directly regulates the expression of a gene but also more indirect relationships. Modeling has been less successful for more complex biological systems such as mammalian tissues, where models of regulatory networks usually contain many spurious correlations. This is partly attributable to the increasingly multi-layered nature of transcriptional control in higher eukaryotes, e.g. involving epigenetic mechanisms and non-coding RNAs. However, a potential major reason for the decreased performance is due to *biological complexity* of datasets which can be defined as the increase of biological variation and the presence of different cell types, which is not compensated by an increase in the number of replicate data points available for modeling. There is an urgent need to identify regulatory mechanisms with more confidence to avoid wasting laborious and expensive wet-lab follow-up experiments on false positive predictions.

The main paradigms of this paper are that regulatory interactions that are consistently found across multiple datasets are more likely to be fundamentally involved and that these regulatory interactions are easier to find in datasets with less biological variation. In the end, regulatory networks trained on less complex biological systems could thus be used for the modeling of the more complex biological systems. We do this using a novel computational technique that combines Bayesian network learning with independent test set validation (using error and variance measures) and a ranking statistic. Whilst Bayesian networks and Bayesian classifiers have been used with great success in bioinformatics (Friedman et al., 2000; Xu et al., 2004), an important weakness has been that, when trying to build models that reveal genuine underlying biological processes, a highly accurate *predictive* model is not always enough (Grossman and Domingos, 2004). The ability to *generalize* to other datasets is of greater importance (Peña et al., 2005). Simple cross-validation approaches on a single dataset will not necessarily result in a model that reflects the underlying biology and therefore will not generalize well. Our approach is to exploit multiple datasets of increasingly complex systems in order to identify more informative genes reflecting the underlying biology.

Bayesian networks have been an important concept for modeling uncertain systems (Pearl, 1986; Buntine, 1996; Heckerman, 1998; Friedman and Koller, 2003). In the last decade several researchers have examined methods for modeling gene expression datasets based on Bayesian network methodology (Segal et al., 2003; Friedman et al., 2000; Xu et al., 2004). These networks are directed acyclic graphs (DAG) that represent the joint probability distribution of variables efficiently and effectively (Friedman et al., 1997). Each node in the graph represents a gene, and the edges represent conditional independencies between genes. Bayesian networks are popular tools for modeling gene expression data as their structure and parameters can easily be interpreted by biologists.

Bayesian classifiers are a family of Bayesian networks that are specifically aimed to classify cases within a data set through the use of a class node. The simplest is known as the naïve Bayes classifier (NBC) where the distribution for every variable is conditioned upon the class and assumes

independence between the variables. Despite this oversimplification, NBCs have been shown to perform very competitively on gene expression data in classification and feature selection problems (Grossman and Domingos, 2004; Fielding, 2007; Tobler et al., 2002). Other Bayesian classifiers, which often have higher *model complexity* as they contain more parameters, involve learning different networks such as trees between the variables and therefore relax the independence assumption (Friedman et al., 1997). The logical conclusion is the general Bayesian Network Classifier (BNC) which simply learns a structure over the variables including the class node. In this paper, we explore the use of the NBC, and the BNC for predicting expression on independent datasets in order to identify informative genes using classifiers of differing complexity.

Accordingly, in order to optimize the classifier and choose the best method, we need to consider the classifiers' bias and variance. Since bias and variance have an inverse relationship (Fielding, 2007), which means decreasing in one increases the other, cross-validation methods can be adopted in order to minimize such an effect. The k-fold cross-validation (Fielding, 2007; Stone, 1974) randomly splits data into k folds of the same size. A process is repeated k times where k-1 folds are used for training and the remaining fold is used for testing the classifier. This process leads to a better classification with lower bias and variance (Kohavi, 1995) than other training and testing methods when using a single dataset. In this paper, we exploit bias and variance using both cross-validation on a single dataset and also independent test data in order to learn models that better represent the true underlying biology. In the next section we provide a description of the gene identification algorithm for identifying gene subsets that are specific to a single simple dataset as well as subsets that exist across datasets of all biological complexity. We used van den Bulcke *et al.* (2006) proposed model for generating synthetic datasets to validate our findings on real microarray data. Moreover, we evaluate the performance of our algorithm by comparing the ability of this model in identifying the informative genes and underlying interactions among genes with the concordance model. Finally, we present the conclusion and summary of our findings in the last section.

METHODS
Multi-Data Gene Identification Algorithm
The algorithm involves taking multiple datasets of increasing biological complexity as input and a repeated training and testing regime. Firstly, this involves a k-fold cross-validation approach on the single simple dataset (from now on we refer to this as the cross-validation data) where Bayesian networks are learnt from the training set and tested on the test set for all k folds. These folding arrangements have been used again for assessing a final model. The Bayesian Network learning algorithm is outlined in the next section.

The Sum Squared Error (SSE) and variance is calculated for all genes over these folds by predicting the measured expression levels of a gene given the measurements taken from others. Next, the same models from each k fold are tested on the other (more complex) datasets (the independent test data) and SSE and variance are again calculated. These SSE and variances are used to rank the genes according to their informativeness (which represents the most predictive and influential genes). Those that are ranked highly in the single-dataset cross-validation experiments will be informative, specific to the single datasets experiment, whereas those that are ranked highly on the independent datasets should be informative in a more general sense in that they are predictive (low SSE) and consistent (low variance) across datasets of all complexity. We evaluate the statistical significance of these rankings using a method proposed by Zhang *et al.* (2006). The full details are outlined in Algorithm 1 where *TrainD* represents the training data (cross-validation

Algorithm 1 - Multi Data Gene Identification Algorithm.

Input: *{TrainD, TestD$_1$, …TestD$_M$, folds}*
            for *k* = 1:*folds*
                        Learn *BN* **using Algorithm 2 on training folds of** *TrainD*
                        Score *SSE* **on test fold** *k* **of** *TrainD*
                        Score *SSE* **on all independent test datasets** {*TestD$_1$…TestD$_M$*}
            end for
            Calculate variance of *SSE* **over all** *k* **folds on** *TrainD* **and** {*TestD$_1$…TestD$_M$*}
            Create gene rankings: *trainR_SSE, train_var,*
            {*testR_SSE$_1$…testR_SSE$_M$*} **and**
            {*testR_var$_1$…testR_var$_M$*} **by ordering the genes**
            on the respective *SSE* **and** *variance* **scores**

Output: *trainR_SSE, train_var,*
            {testR_SSE$_1$…testR_SSE$_M$}
            {testR_var$_1$…testR_var$_M$}

---

data, here the relatively simple datasets), and *TestD$_1$* … *TestD$_M$* represent the more complex test datasets, independent test data.

Bayesian Network Structure Learning

The goal of learning gene regulatory networks using Bayesian network approaches is to establish the structure of the network and then to parameterize the conditional probability tables (Su and Zhang, 2006). As the number of possible network structures is huge, learning the structure of a network has a high computational cost. Since the effective learning of network structure engages a trade-off of bias vs. variance, the necessity of designing an algorithm in which it can generate an ideal structure for a given dataset, with a degree of biological complexity, is crucial (Chickering et al., 2004). In this study, instead of using well studied but unrealistic and sometimes not effective classifiers such as NBC and Tree Augmented Networks (TAN), we use an optimization approach that uses a simulated annealing search and the Bayes Information Criterion (BIC) as a scoring metric (Schwarz, 1978). The advantage of simulated annealing over other methods (like greedy searches or hill climbing) is that it aims to avoid local maxima (Friedman et al., 1997). We have chosen the BIC as a fitness function as it is less prone to overfitting through the use of a penalizing term for overly complex models.

Bayesian networks with more connections between their nodes require a higher number of parameters and as a result increase the complexity of the models exponentially (Lam and Bacchus, 1994). Therefore, we explore three different classes of model learning: the Selective Naïve Bayes (SNB) where only links between a class node representing differentiation status and a gene are explored, a search that explores structures with links between genes but limiting each gene to having only one parent (1PB). Limiting the number of parents in a Bayesian network is common practise but can be considered a crude approach to reducing parameters. As a result we also explore a full unlimited structure learning (NPB) and learn these structures using the simulated annealing with the BIC scoring metric (which naturally penalises overly complex networks). In this study, the initial state of the structure is an empty DAG with no link. In order to alter the network structures, three operators have been used within the simulated annealing. These operators are adding, removing, or swapping links to generate a new network for validation. These alterations

can be either accepted or rejected. The outline of this procedure can be found in Algorithm 2.

Prediction and Ranking

Zhang *et al.* (2006) proposed a method to convert a set of gene rankings into position p-values to evaluate the significance of a given gene. However, this involved working with resampling techniques upon a single dataset. Here, we use the ranking lists according to the model's average SSE and variance for both the original simple dataset and the independent test sets in order to generate position p-values. This requires us to include, a number of random genes which can be counted as uninformative genes. By comparing the actual ranking of the gene with the null distribution we can calculate the position p-values. In this paper we are using three independent datasets so we do not need to use resampling in order to generate more gene rankings as Zhang *et al.* (2006) did in their experiments. In addition, the different rankings will have different interpretations as some are based purely on the simple dataset whilst others are influenced by error and variance on the more biologically complex independent data.

Algorithm 2 - Simulated Annealing Structure Learning.

```
Input:    t_0, maxfc, D
          fc=0, t=t_0, t_n=0.001
          c=(t_n/t_0)^(1/maxfc)
          Initial bn to a Bayesian classifier with no inter-gene links
          results = bn
          oldscore=score(bn)
          while fc<maxfc do
                    for each operator do
                              apply operator to bn
                              newscore=score(bn)
                              fc=fc+1
                              dscore=newscore-oldscore
                              if newscore>oldscore then
                                        result=nbc
                              else if r(0,1)<e^(dscore/t) then
                                        Undo the operator
                              end if
                    end for
                    t=t x c
          end while
Output:   result
```

Datasets

With the aim of investigating the influence of the complexity of a gene expression dataset on the performance of classifiers in identifying the gene regulatory network, three gene expression datasets (with increasing biological variation) have been chosen for this study [GSE3858 (Cao et al., 2006), GSE1984 (Iezzi et al., 2004), and GSE989 (Tomczak et al., 2004)]. These three datasets are all concerned with the differentiation of cells into the muscle (Myogenic) lineage. During this process, mononucleated precursor cells stop to proliferate, differentiate and fuse with each other to become elongated multinucleated myotubes or myofibres. This in-vitro system mimics the formation of new muscle fibres in-vivo. The cell types differ between the different datasets:

- GSE3858: Embryonic fibroblasts (EF)

- GSE989 and GSE1984: C2C12 tumor cell line that has the potential for differentiation into different mesodermic lineages (mainly muscle and bone)

Also methods to drive cells into myogenic differentiation differ:

- GSE3858: Exogenous expression of the myogenic transcription factors are Myod and Myog.

- GSE989 and GSE1984: Serum Starvation

In addition, the study by Sartorelli included different treatments that affect the timing and efficiency of the myogenic differentiation process. The time points for sampling differ between the studies (**Table 1**). The class node reflecting the differentiation status had two possible states: undifferentiated (for all time points until myogenic differentiation was

Table 1 - Specification of three muscle differentiation datasets.

Dataset	Cell Type	Platform	Samples	Time Points
Tomczak	C2C12	Affy U74A	24	8
Cao	EF	Affy 430.2	36	4
Sartorelli	C2C12	Affy U74A	32	6

induced) and differentiated (for time points where myogenic differentiation had been induced). In the rest of this paper we call these datasets by the name of the first author (e.g. Cao instead of GSE3858).

Data Processing and Analysis

The raw microarray data were normalized and summarized with the RMA method (Irizarry et al., 2003), using the affy package in R. Only the 8904 probesets common to the Affymetrix U74A and 430.2 used in mentioned studies were considered in the analysis. All datasets were standardized to mean 0 and the standard deviation 1 across the genes. For the scope of this paper, first, we selected for each dataset a subset of 100 genes most affected by the induction of differentiation. These genes were identified with Student's t-test which compared samples from undifferentiated and differentiated cell cultures, disregarding the time of differentiation. An additional 50 genes were randomly selected to be able to calculate ranking p-scores described above and using the Kolmogorov-Smirnov test. For cross-validation we divided Cao dataset into 9 folds, Sartorelli into 8 folds, and Tomczak into 6 folds based upon the number of samples in each dataset. Simulated annealing has three attributes which should be set before starting the learning phase. It is crucial to set an appropriate initial temperature, sufficient number of iterations, and a convenient fitness function. In this study, the initial temperature has been set to 10 and it terminates at 0.001. The number of iterations has been set to 1000 for the first set of experiments only using most informative genes (top 100) and then we set the number of iterations to 1500 since we added 50 uninformative genes to the network. The code is implemented in Matlab 2007a using the Bayes Net toolbox (Murphy, 2001) to generate gene regulatory networks.

Analysis of myogenesis-Related genes

Myogenesis-related genes are defined as genes associated with the Gene Ontology term "Muscle Development" supplemented with all genes strongly associated with Myogenesis in the biomedical literature, as determined with the literature analysis tool Anni v2.0 (Jelier et al., 2008) with the association score greater than 0.02.

Analysis of Synthetic datasets

The use of datasets in which the underlying network is known enables us to validate the new algorithms that have been developed to identify gene regulatory networks and capture the most informative genes. van den Bulcke *et al.* (2006) proposed a new methodology to generate synthetic datasets where the network structure is known and biological, experimental, and model complexity can be manipulated. However, a disadvantage of this approach is that the generated networks can contain some overlapping pieces of the known network which may weaken the models being probabilistically independent (Haynes and Brent, 2009). Whilst SynTReN uses resampling from potentially overlapping networks, the generated data undergoes a robust statistical cross-validation regime ensuring that any prediction is applied to unseen data. The focus of this paper is upon the prediction of increasingly complex datasets, sampled from some underlying

biological process. Consequently, these synthetic datasets can be used for validating the performance of our methodology in identifying the informative genes and the interactions among them in real microarray data. SynTReN (Van den Bulcke et al., 2006) generates networks with more realistic topological characteristics and since we use this application to investigate the impacts of biological, experimental, and model complexity on identifying informative genes using the same sub-network is an advantage. Three datasets have been generated on the well-described network structure of *E. coli* (Ma et al., 2004) which contains 1330 number of nodes and 2724 interactions. These datasets have been generated in a manner that they can match the key characteristics of real microarray datasets we used in this study (for instance, limiting the number of genes that were selected for modelling to 150). This enables us to investigate the possibility of reproducing similar results on synthetic data which can be easily corrected for differences such as number of samples and time points per dataset (see **Additional file 1**) and avoid weakening the probabilistically independent assumption of the generated datasets.

Analysis of Concordance between datasets

The study of the concordance between microarray datasets has increased considerably in the past few years (Miron et al., 2006). However, a robust statistical method for examining the concordance or discordance among microarray experiments carried out in different laboratories is yet to develop. Methods such as multiplication of gene p-values in order to generate a list of rankings for concordance genes showed bias towards datasets with higher significance level (Rhodes et al., 2002). Lai *et al.* (2009) proposed a promising methodology (which we call concordance model) to investigate the concordance or discordance between two large-scale datasets with two responses. This method uses a list of z-scores, generated using a statistical test of differential expression, as an input to evaluate the concordance or discordance of two datasets by calculating the mixture model based likelihoods and testing the partial discordance against concordance or discordance. Additionally, the statistical significance of a test is being evaluated by the parametric bootstrap procedure and a list of gene rankings is being generated which can be used for integrating two datasets efficiently. In this paper we are using a set of gene rankings generated by this method to evaluate the performance of our model in identifying informative genes from multiple datasets with increasing complexity.

RESULTS

The aim of this study is to demonstrate firstly, the influence of model complexity in discovering accurate gene regulatory networks on multiple datasets with increasing biological complexity. Secondly, to investigate if cleaner and more informative datasets can be used for modelling more complex ones. Therefore, three public datasets that are concerned with the differentiation of cells into muscle lineage were chosen for this study. From a biological point of view, Sartorelli is the most complex dataset since it involves different treatments influencing myogenesis. Tomczak and Cao are less complex datasets. It is difficult to say how their complexity relates since Tomczak uses more heterogeneous stimuli to induce differentiation but has more time points, while Cao uses more defined stimuli (Myod or Myog transduction) and less time points. In order to meet the scope of this study, we evaluated the quality and informativeness of these datasets based on two criteria. Firstly, we calculated the average correlations between replicates as a measurement of noisiness of each dataset. Secondly, using Student's t-test method, we counted the number of differentially expressed genes with the significance levels of 0.05 and 0.01 as a measurement of informativeness (**Table 2**). Although the average correlations between replicates in all three datasets are very close, datasets differ in number of significant genes they hold. Tomczak is the most informative dataset as it includes the most number of significant genes and has a higher

average correlation value for the replicate samples in the dataset which represent the lowest level of noise. In contrast, Sartorelli contains the least differentially expressed genes with almost 12% of what Tomczak contains. Moreover, it has the lowest average correlation value and can be marked as the most complex dataset to model in this study as it has the highest noise level and the least number of informative genes. Therefore, we ordered these datasets by increasing biological complexity in the following way: Tomczak, Cao, and Sartorelli.

Table 2 - The average correlations between replicates and number of differentially expressed genes (based on BH corrected p-values) in each dataset.

Dataset	Correlation	Genes with a P-value (BH) less than	
		0.05	0.01
Tomczak	0.975	4602	3604
Cao	0.971	3668	2623
Sartorelli	0.964	1199	458

Comparison of classifiers and network analysis

We now explore how the different classifiers performed on these three datasets. Figure 1 shows the average error rate of the different classifiers trained on each given dataset. It can be seen that of the three classifiers, 1PB and NPB generated the same pattern and have very close error rates on cross-validation (training) sets. However, it is evident that NPB (particularly on Tomczak) performs poorer than 1PB on the independent test set, possibly due to overfitting as these models contain more parameters. Even though SNB performed poorly on both the cross-validation test and the independent data test, in some cases it could compete with NPB which appears to be too complex to predict some of the independent datasets accurately. Hence, 1PB has performed favorably, both in terms of average error rate and the difference between the cross-validation test and the independent data test (see **Additional file 1** for complete set of results).

According to Mac Nally (2000) simple models should be sought for various reasons. Firstly, simple models are more stable and capable of not overfitting to noise in the data which will influence the performance of classifier with future data. Secondly, they tend to provide a better insight into causality and interactions among genes. Finally, reducing the number of parameters will decrease the cost of validating a model for current and future data. However, we need a model that matches the complexity of data sets. Considering this argument along with our first set of results, we chose 1PB as a model that can capture the interactions among genes and does not overfit to noise. In order to understand the impacts of using different datasets for gene selection and training 1PB classifier (which will be discussed in the next section), we need to analyse the performance of the 1PB classifier on the top 100 (most informative) genes in more detail.

Additional file 1, Figure S7 represents the comparison of the error rate of the 1PB classifier on cross-validation versus the independent test. It is shown that the 1PB classifier trained on Tomc-

Figure 1 - The comparison of classifiers with increasing model complexity. Three Bayesian network models (SNB, 1PB, and NPB) have been trained using cross-validation set and validated on independent datasets. An average error rate of the classifiers' prediction has been calculated for each gene and an overall SSE on cross-validation set and independent test set are illustrated in this figure.
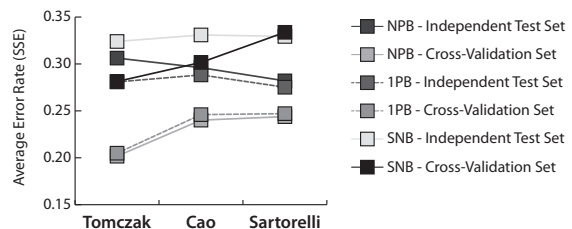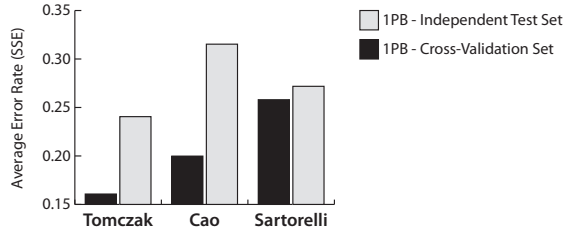
Figure 2 - Evaluating the accuracy of 1PB using different datasets for gene selection. We selected genes using only one dataset (black) at a time and compared the average error rate of 1PB classifier learnt and trained on a same dataset and validated on the other two datasets independently (grey).

zak performed significantly better on cross-validation and Sartorelli shows the lowest differentiation between cross-validation and the independent test with almost the same average error rate on the cross-validation set compared to Cao. Although the differentiation of average error rate on the cross-validation set and independent test set is high in Tomczak, this model produced the best models in terms of the lowest overall error rate. This figure raises the idea that Tomczak is the most informative dataset since it can model any dataset, regardless of the gene selection method, significantly better than the other alternatives. This will be discussed in more detail in the *Extraction of infotmative genes* section.

Comparison of gene selections with differing informativeness
We now look into how the different gene selections impact on the average error rate of the 1PB classifier for both cross-validation and the independent test. Figure 2 demonstrates the performance of the 1PB classifier in modeling datasets generated using different gene selections. Clearly, unlike Sartorelli, genes selected from Tomczak and Cao show very good performances on cross-validation. However, by looking at the average error rate of 1PB on independent test sets, we can see that the models learnt on Cao over-fitted the data and performed poorly on the independent test set (with the SSE of 0.32) whereas Sartorelli shows the lowest differentiation between the two sets. Overall the Tomczak selection performed the best both on cross-validation and the independent test.

It is important to adopt a methodology that can generate an accurate gene regulatory network, moreover, it is crucial to generate a model that can capture the significant genes and distinguish informative genes from uninformative ones. For this purpose, we added 50 randomly selected genes with high p-values (which imply less relatedness to Myogenesis) from the distribution. This also has the effect that it will increase the complexity of the datasets.

Figure 3 shows that there is a similar pattern on the average error rate of cross-validation. The additional random genes do not seem to affect Cao. It does, however, have an interesting impact on Sartorelli. The models learnt on Sartorelli (see **Additional file 1**) performed even poorer than SNB on the independent data sets and showed no significant changes when using different datasets for training. It is interesting because we know that the Sartorelli dataset is noisy and biologically complex and adding the random genes, which increases the complexity of the models in terms of more nodes and increases the risk of spurious links, produces a classifier which appears to be unable to capture the real gene interactions. The error rate and variance of models learnt on the Sartorelli selection is significantly high in comparison with Tomczak. By comparing figures 2 and 3, we can conclude that simpler and cleaner datasets tend to perform more reliably and have more stability while increasing the complexity. Since it is important to validate these models according to their variances, we demonstrated the average variance of each model on cross-validation and the independent test set in Additional file 1, Figure S8. Interestingly, we can see a similar pattern in the classifiers' variance in comparison with the average error rate (**Figure**

Figure 3 - The investigation of inference of adding more complexity to the model.
We investigated the inference of adding more complexity to the model by adding 50 randomly selected genes as uninformative on 1PB classifier performance. In this figure we compare the average error rate of 1PB classifier after adding 50 uninformative genes to the model.
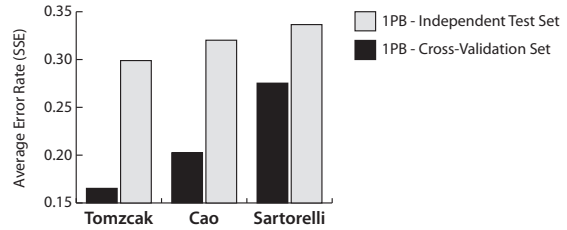
3). It is clear that we can raise the same conclusion as the simpler and cleaner datasets perform better than more noisy and complex ones. In this study, Tomczak performed favorably both in terms of bias and variance.

It is crucial to investigate if these findings are reproducible and are not prone to the number of samples and time points per dataset. Therefore, we applied our model on three synthetic datasets that have been generated by manipulating the biological, experimental, and model complexity of their known network structure using SynTReN application (Van den Bulcke et al., 2006). Additional file 1, Figure S9 illustrates that we can see a very similar pattern as we have seen on a real data where there is an increase on the average error rate of models learnt on multiple synthetic datasets with increasing biological variability. In the next section, before examining if these models can help us to capture the interactions in more complex datasets, we will investigate how well these models separate the informative genes from uninformative ones.

Extraction of informative genes
In order to test the ability of classifiers to separate informative genes from uninformative ones, we have looked at the result of the Kolmogorov-Smirnov test (KS test) on the ranking of genes according to their average error rate using a given model. Using this algorithm, we calculated the p-value, KS test, and the result of investigating the differentiation hypothesis along with the models' bias or variance. The results of this investigation are displayed in Additional file 1, Table S1 where Cao and Tomczak performed very well on cross-validation both in terms of bias and variance. However, models learnt on Sartorelli fail to separate between informative genes and uninformative genes as the scores are generally very low.

Generally, Tomczak outperformed Sartorelli and Cao and can be chosen as the most informative dataset in this study. Models learnt on Tomczak generated the lowest bias and variance and produced the best separation. In contrast, Sartorelli is the noisiest and less informative dataset while it failed to handle any increases in complexity (both biological and model wise) and generates models with highest bias and variance which also cause disability to separate informative genes from the others. Now the question is whether we can use a simpler and cleaner dataset to model more complex ones. In the next section we show how we tackled this question.

Analysis of the use of simpler dataset to model more complex one
In this section, we investigate the improvement or deterioration of genes selected by Tomczak on the Sartorelli dataset. Figure 4 shows the average improvement or deterioration of ranks of myogenesis-related genes, top 100 genes (most informative), and 50 randomly selected genes (uninformative) in Sartorelli. We compared the original rank of each gene (which can be any number between 1 and 150 derived from its p-value comparing to others) with its rank based upon the ability of a model trained on Tomczak to predict gene's value in Sartorelli. Moreover, we evaluate the improvement or deterioration of genes rankings in our model with the ones generated using
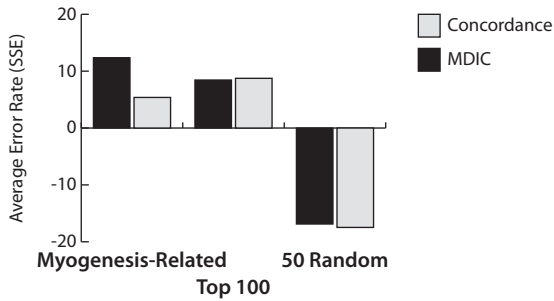
Figure 4 - The improvement or deterioration of genes ranking in Sartorelli. Firstly, we selected 100 informative and 50 uninformative genes using Tomczak dataset and extracted their ranks in Sartorelli. Secondly, we trained 1PB classifier on Tomczak and tested on Sartorelli. Finally, we ranked genes according to the average error rate of 1PB classifier in predicting their values in Sartorelli. This figure illustrates the average improvement or deterioration of Myogenesis-Related, Top 100, and 50 randomly selected genes in Sartorelli generated with our method and the gene rankings generated by concordance model.

the concordance model described by Lai *et al.* (2009). We can clearly see that the model learnt on Tomczak can capture the informative genes in Sartorelli and improve their rank whereas uninformative genes have been pushed down (almost 17 places in average) in the ranking by the classifier. Additionally, the improvement is even more pronounced for myogenesis-related genes with 12.33 places in average, which is significantly better than others with $P < 0.004$ generated using KS test, and as expected top 100 genes has been improved by 8.44 places. Even though both methods perform similarly on improving the ranks of top 100 and deteriorating the ranks of 50 randomly selected genes, the improvement of ranks for myogenesis-related genes are much more pronounced in our model than in the concordance model (improvement of 5.38 places).

Myh7 and *Tor3a* are two examples of significant improvements in Sartorelli dataset. *Myh7*, which originally ranked 101, improved 96 places to rank 5 (rank 55 in concordance model). During the learning phase it has been linked to four other genes of which three of them are myogenesis-related. These genes, in both datasets, have direct correlations and can represent each other in terms of prediction and validation. However, *Tor3a* has a very low rank in both dataset and yet improved 107 places from 128 to 21 (rank 31 in concordance model). It has been linked to *Prune* which also improved 106 places (from 131 to 25, 100 in concordance model). All three genes mentioned above have been selected as informative genes from Tomczak and yet placed into the bottom 50 due to the quality of Sartorelli dataset. These were some examples of the ability of model to pull out informative genes from a distribution (**Figures S10A** and **S10B**).

Although the overall improvement on myogenesis-related genes is significantly high, we were concerned why this model failed to improve the rank of some genes like *Id3* which dropped from rank 1 in Sartorelli to 133 (rank 51 in concordance model). In the learning process, *Id3* has been linked to 4 genes which are: *Fabp3*, *Rbm38*, *X99384*, and *Slco3a1*. Now in order to answer the question, firstly, we validate the relatedness of these genes to *Id3* in Tomczak dataset to investigate if they are significant and can represent *Id3*. Secondly, we study the expression level of these genes in Sartorelli to identify the reason why this model failed dramatically in predicting the *Id3* value.

Additional file 1, Figure S11 demonstrates the expression level of *Id3* along with its parent/children in both Tomczak and Sartorelli datasets. In Tomczak we can clearly see that there is an inverse relationship between *Id3* and the other 4 genes which is very significant. While the differentiation state changes, *Id3* drops from the expression level of approximately 11 to 8.5 and similarly its relatives show an increase of about 2 points in their expression values. This supports the assumption of the relatedness of these genes to *Id3* in the learning process on Tomczak dataset. However, considering that *Id3* is still very significant in Sartorelli, *Id3* parent/children show no

variation and simply are not significant. As a conclusion, this model failed to predict *Id3* expression value and as a result the rank of *Id3* dropped 132 places most probably due to the quality and biological variation of Sartorelli dataset. Since we aim to overcome the lack of overlap on the gene regulatory network studies across species and platforms, the natural extension of the work in this paper would be to explore how this model can be used on datasets from multiple biological systems with increasing complexity. Moreover, it would be valuable to consider methods such as model averaging (Madigan and Raftery, 1994) that has been shown better generalization in classifier's accuracy. Consequently, it improves the performance of classifiers in identifying the most informative genes and avoids deterioration of cases like *Id3*. Furthermore, dynamic Bayesian networks can be adopted when learning from time-series data in order to handle auto-regulation and feedback loops, two key components of regulatory networks in biological data (Shen-Orr et al., 2002; Lee et al., 2002).

CONCLUSIONS

In this study, we have investigated a number of different Bayesian classifiers and datasets for identifying firstly, subsets of genes that are related to myogenesis and muscle differentiation, and secondly the use of cleaner and more informative datasets in modelling more biologically complex datasets. We have shown that an appropriate combination of simpler and more informative datasets produce very good results, whereas models learnt on genes selected from more complex datasets performed poorly. We concluded that simpler datasets can be used to model more complex ones and capture the interactions among genes. Moreover, we have described that highly predictive and consistent genes, from a pool of differentially expressed genes, across independent datasets are more likely to be fundamentally involved in the biological process under study. In three published datasets, we have demonstrated that these models can explain the myogenesis-related genes (genes of interest) significantly better than others ($P < 0.004$) since the improvement in their rankings is much more pronounced. These results imply that gene regulatory networks identified in simpler systems can be used to model more complex biological systems. In the example of muscle differentiation, a myogenesis-related gene network may be difficult to derive from in vivo experiments directly due to the presence of multiple cell types and inherently higher biological variation, but may become evident after initial training of the network on the cleaner in vitro experiments. In order to validate our approach, firstly, we evaluated our model on synthetic datasets and secondly we performed comparisons between our approach and the method of Lai *et al.* (2009) which we call concordance model. It is shown that our model performs comparably in improving the ranks of informative genes and deteriorating the ranks of uninformative ones, but that the improvement of ranks for myogenesis-related genes is much more pronounced whilst additionally modelling the interactions among genes. However, it is necessary to develop other statistical measures so that the model can be quantified to distinguish different degrees of complexities and platforms whilst handling the auto-regulation and feedback loops within the network.

Authors' contributions
SYA, PACH and AT contributed equally to methods development and drafting the paper. PACH provided the biological insight on the datasets related to muscle differentiation. AT designed the algorithms and SYA developed the codes. PACH and AT supervised the study. All authors analyzed the data, read and approved the final manuscript.

Reference List

Bockhorst,J., Craven,M., Page,D., Shavlik,J., and Glasner,J. (2003). A Bayesian network approach to operon prediction. Bioinformatics. *19*, 1227-1235.

Buntine,W.L. (1996). A guide to the literature on learning probabilistic networks from data. IEEE Transactions on Knowledge and Data Engineering *8*, 195-210.

Cao,Y., Kumar,R.M., Penn,B.H., Berkes,C.A., Kooperberg,C., Boyer,L.A., Young,R.A., and Tapscott,S.J. (2006). Global and gene-specific analyses show distinct roles for Myod and Myog at a common set of promoters. EMBO J. *25*, 502-511.

Chickering,D.M., Heckerman,D., and Meek,C. (2004). Large-sample learning of Bayesian networks is NP-Hard. Machine Learning Research *5*, 1287-1330.

Fielding,A.H. (2007). Introduction to classification. In Cluster and classification techniques for the Biosciences, Cambridge University Press), p. 86.

Friedman,N., Geiger,D., and Goldszmidt,M. (1997). Bayesian network classifiers. Machine Learning *29*, 131-163.

Friedman,N. and Koller,D. (2003). Being Bayesian about network Structure. A Bayesian approach to structure discovery in Bayesian networks. Machine Learning *50*, 95-125.

Friedman,N., Linial,M., Nachman,I., and Pe'er,D. (2000). Using Bayesian networks to analyze expression data. J. Comput. Biol. *7*, 601-620.

Grossman,D. and Domingos,P. (2004). Learning Bayesian network classifiers by maximizing conditional likelihood. Proceedings of the 21st International Conference on Machine Learning *68*, 46-54.

Haynes,B.C. and Brent,M.R. (2009). Benchmarking regulatory network reconstruction with GRENDEL. Bioinformatics *25*, 801-807.

Heckerman,D. (1998). A tutorial on learning with Bayesian networks. In Learning in graphical models, (Dordrecht: Kluwer Academic Publishers), p. 301.

Iezzi,S., Di,P.M., Serra,C., Caretti,G., Simone,C., Maklan,E., Minetti,G., Zhao,P., Hoffman,E.P., Puri,P.L., and Sartorelli,V. (2004). Deacetylase inhibitors increase muscle cell size by promoting myoblast recruitment and fusion through induction of follistatin. Dev. Cell *6*, 673-684.

Irizarry,R.A., Hobbs,B., Collin,F., Beazer-Barclay,Y.D., Antonellis,K.J., Scherf,U., and Speed,T.P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. *4*, 249-264.

Jelier,R., Schuemie,M.J., Veldhoven,A., Dorssers,L.C., Jenster,G., and Kors,J.A. (2008). Anni 2.0: a multipurpose text-mining tool for the life sciences. Genome Biol. *9*, R96.

Kohavi, R. Wrapper for performance enhancement and oblivious decision graphs. 1995. Stanford University, Computer Science Department.

Ref Type: Thesis/Dissertation

Lai,Y., Eckenrode,S.E., and She,J.X. (2009). A statistical framework for integrating two microarray data sets in differential expression analysis. BMC. Bioinformatics. *10 Suppl 1*, S23.

Lam,W. and Bacchus,F. (1994). Learning Bayesian belief networks (an approach based on the MDL principle). Computational Intelligence *10*, 1-31.

Lee,T.I., Rinaldi,N.J., Robert,F., Odom,D.T., Bar-Joseph,Z., Gerber,G.K., Hannett,N.M., Harbison,C.T., Thompson,C.M., Simon,I., Zeitlinger,J., Jennings,E.G., Murray,H.L., Gordon,D.B., Ren,B., Wyrick,J.J., Tagne,J.B., Volkert,T.L., Fraenkel,E., Gifford,D.K., and Young,R.A. (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. Science *298*, 799-804.

Ma,H.W., Kumar,B., Ditges,U., Gunzer,F., Buer,J., and Zeng,A.P. (2004). An extended transcriptional regulatory network of Escherichia coli and analysis of its hierarchical structure and network motifs. Nucleic Acids Res. *32*, 6643-6649.

Mac Nally,R. (2000). Regression and model-building in conservation biology, biogeography and ecology: the distinction between - and reconciliation of - 'predictive' and 'explanatory' models. Biodiversity and Conservation *9*, 655-671.

Madigan,D. and Raftery,A.E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. Journal of the American Statistical Association *89*, 1535-1546.

Miron,M., Woody,O.Z., Marcil,A., Murie,C., Sladek,R., and Nadon,R. (2006). A methodology for global validation of microarray experiments. BMC. Bioinformatics *7*, 333.

Murphy,K.P. (2001). The Bayes Net toolbox for Matlab. Computing Science and Statistics: Proceedings of the Interface *33*, 331-350.

Pearl,J. (1986). Fusion, propagation, and structuring in belief networks. Artificial Intelligence *29*, 241-288.

Peña,J.M., Björkegren,J., and Tegnér,J. (2005). Learning dynamic Bayesian network models via cross-validation. Pattern

Recognition Letters *26*, 2295-2308.

Rhodes,D.R., Barrette,T.R., Rubin,M.A., Ghosh,D., and Chinnaiyan,A.M. (2002). Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. Cancer Res. *62*, 4427-4433.

Schwarz,G. (1978). Estimating the dimension of a model. The Annals of Statistics *6*, 461-464.

Segal,E., Shapira,M., Regev,A., Pe'er,D., Botstein,D., Koller,D., and Friedman,N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet. *34*, 166-176.

Shen-Orr,S.S., Milo,R., Mangan,S., and Alon,U. (2002). Network motifs in the transcriptional regulation network of Escherichia coli. Nat Genet. *31*, 64-68.

Stone,M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). Journal of the Royal Statistical Society *36*, 111-147.

Su,J. and Zhang,H. (2006). Full Bayesian network classifiers. Proceedings of the 23rd International Conference on Machine Learning *148*, 897-904.

Tobler,J.B., Molla,M.N., Nuwaysir,E.F., Green,R.D., and Shavlik,J.W. (2002). Evaluating machine learning approaches for aiding probe selection for gene-expression arrays. Bioinformatics. *18 Suppl 1*, S164-S171.

Tomczak,K.K., Marinescu,V.D., Ramoni,M.F., Sanoudou,D., Montanaro,F., Han,M., Kunkel,L.M., Kohane,I.S., and Beggs,A.H. (2004). Expression profiling and identification of novel genes involved in myogenic differentiation. FASEB J. *18*, 403-405.

Van den Bulcke,T., Van,L.K., Naudts,B., van,R.P., Ma,H., Verschoren,A., De,M.B., and Marchal,K. (2006). SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. BMC. Bioinformatics. *7*, 43.

Xu,X., Wang,L., and Ding,D. (2004). Learning module networks from genome-wide location and expression data. FEBS Lett. *578*, 297-304.

Zhang,C., Lu,X., and Zhang,X. (2006). Significance of gene ranking for classification of microarray samples. IEEE Transactions on Computational Biology and Bioinformatics *3*, 312-320.
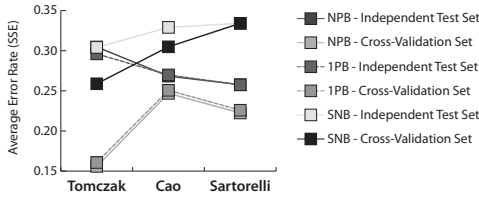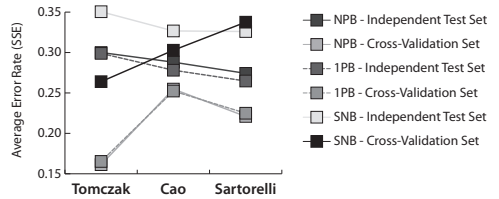
APPENDIX

Figure S1 - The comparison of classifiers with increasing complexity. Three Bayesian network models (SNB, 1PB, and NPB) have been trained using cross-validation set and validated on independent datasets. An average error rate of the classifiers' prediction has been calculated for each gene (selected from Tomczak dataset) and an overall SSE on cross-validation set and independent test set are illustrated in this figure. These models have been trained on each dataset and validated on the other two datasets.
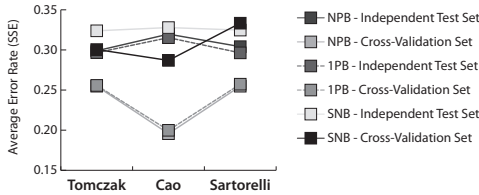
Figure S2 - The comparison of classifiers with increasing complexity. Three Bayesian network models (SNB, 1PB, and NPB) have been trained using cross-validation set and validated on independent datasets. An average error rate of the classifiers' prediction has been calculated for each gene (selected from Tomczak dataset) and an overall SSE on cross-validation set and independent test set are illustrated in this figure. These models have been trained on each dataset and validated on the other two datasets.
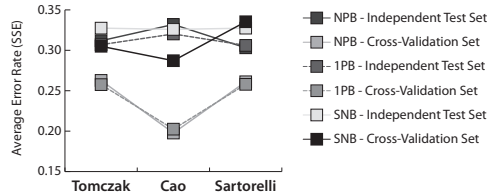
Figure S3 - The comparison of classifiers with increasing complexity. Three Bayesian network models (SNB, 1PB, and NPB) have been trained using cross-validation set and validated on independent datasets. An average error rate of the classifiers' prediction has been calculated for each gene (selected from Cao dataset) and an overall SSE on cross-validation set and independent test set are illustrated in this figure. These models have been trained on each dataset and validated on the other two datasets.
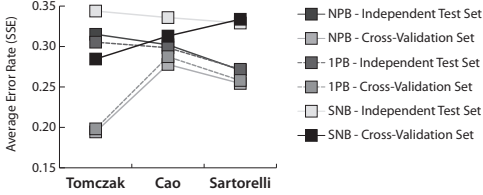
Figure S4 - The comparison of classifiers with increasing complexity. Three Bayesian network models (SNB, 1PB, and NPB) have been trained using cross-validation set and validated on independent datasets. An average error rate of the classifiers' prediction has been calculated for each gene (selected from Cao dataset) and an overall SSE on cross-validation set and independent test set are illustrated in this figure. These models have been trained on each dataset and validated on the other two datasets.
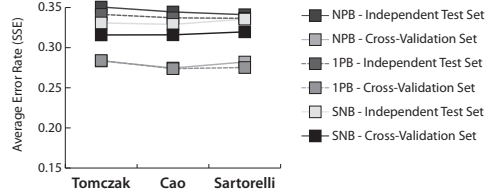
Figure S5 - The comparison of classifiers with increasing complexity. Three Bayesian network models (SNB, 1PB, and NPB) have been trained using cross-validation set and validated on independent datasets. An average error rate of the classifiers' prediction has been calculated for each gene (selected from Sartorelli dataset) and an overall SSE on cross-validation set and independent test set are illustrated in this figure. These models have been trained on each dataset and validated on the other two datasets.

Figure S6 - The comparison of classifiers with increasing complexity. Three Bayesian network models (SNB, 1PB, and NPB) have been trained using cross-validation set and validated on independent datasets. An average error rate of the classifiers' prediction has been calculated for each gene (selected from Sartorelli dataset) and an overall SSE on cross-validation set and independent test set are illustrated in this figure. These models have been trained on each dataset and validated on the other two datasets.
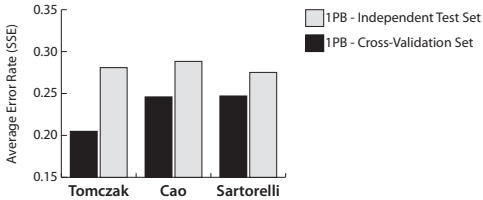

Figure S7 - The comparison of the differences between cross-validation set and independent test set on average error rates of 1PB classifier (extracted from figure 1).
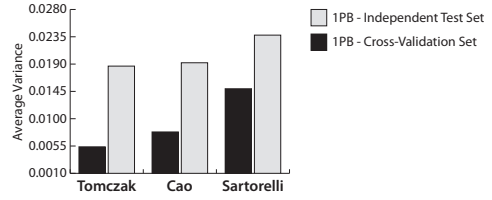

Figure S8 - The investigation of inference of adding more complexity to the model by adding 50 randomly selected genes as uninformative on 1PB classifier performance. In this figure we compare the average variance of 1PB classifier after adding 50 uninformative genes to the model.
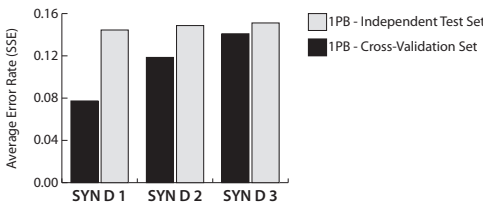

Figure S9 - This figure illustrates the performance of 1PB classifier on modeling three synthetic datasets generated using SynTReN application by manipulating the biological and experimental complexity. There is an increase of the biological variability on three datasets which matches an increase on the average error rate of models learnt.
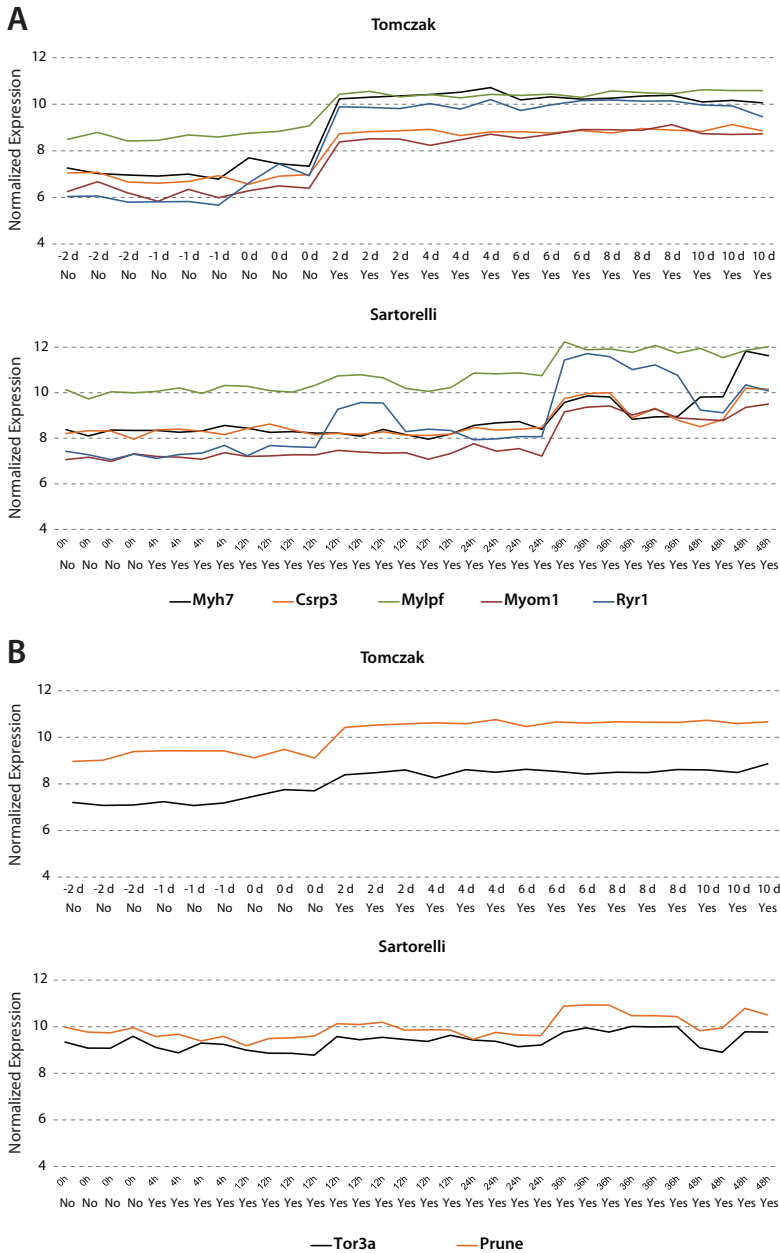
Figure S10 - A).The expression level of *Myh7* along with its parent/children in both Tomczak and Sartorelli datasets. In Tomczak we can clearly see that there is a strong relationship between *Myh7* and the other 4 genes. Moreover, in Sartorelli dataset the correlation still exists between *Myh7* and *Csrp3*, *Mylpf*, *Myom1*, and *Ryr1* even though it is not as strong as Tomczak. **B)** The expression level of *Tor3a* along with its parent in both Tomczak and Sartorelli datasets. In Tomczak we can clearly see that there is a good relationship between *Tor3a* and *Prune*. Moreover, in Sartorelli dataset the correlation still exists between *Tor3a* and *Prune*. This figure is an example of a large improvement of rank of a given gene after training on Tomczak. The x-axis represents both the time points and the differentiation status.
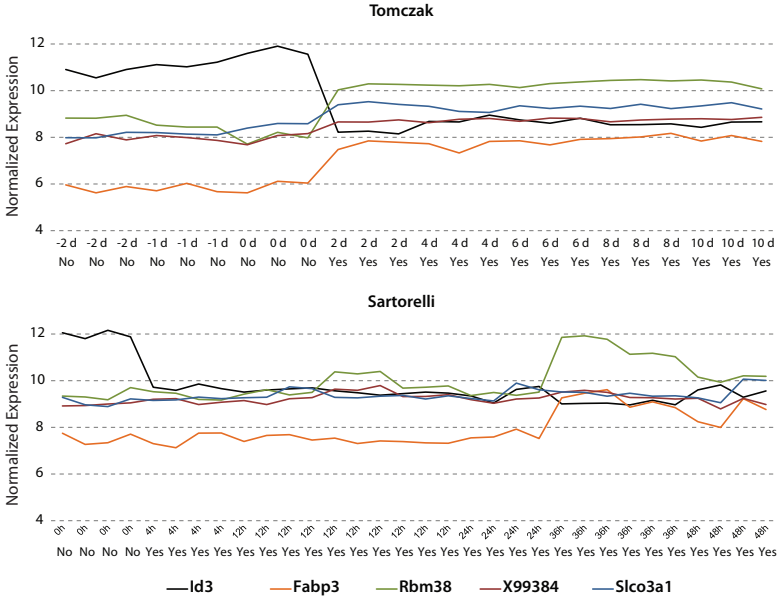
Figure S11 - The expression level of *Id3* along with its parent/children in both Tomczak and Sartorelli datasets. In Tomczak we can clearly see that there is an inverse relationship between *Id3* and the other 4 genes while Sartorelli dataset shows no significant correlations between *Id3* and *Fabp3*, *Rbm38*, *X99384*, and *Slco3a1*. This figure is an example of a large deterioration of rank of a given gene after training on Tomczak. The x-axis represents both the time points and the differentiation status.

Table S1 - Differentiation Hypothesis. Investigating how well the models can separate the informative and uninformative genes from each other. Firstly, we ranked genes according to their average error rate and variance. Secondly, using Kolmogorov-Smirnov test and original ranking list, we explored which model can separate the informative genes from uninformative genes the best.

Gene Selection		Error Rate (SSE)		Variance	
		Cross-Validation Set	Independent Test Set	Cross-Validation Set	Independent Test Set
» Tomczak	Differentiation Hypothesis	TRUE	TRUE	TRUE	TRUE
	P-value	5.02E-24	9.77E-10	5.02E-24	3.68E-05
	Kolmogorov-Smirnov Test	0.880198	0.552871	0.880198	0.394257
	Average Performance	0.165259	0.298921	0.00537	0.018667
Cao	Differentiation Hypothesis	TRUE	TRUE	TRUE	TRUE
	P-value	1.89E-22	6.16E-06	1.91E-20	0.004314
	Kolmogorov-Smirnov Test	0.850297	0.425347	0.810693	0.295842
	Average Performance	0.202472	0.320211	0.007819	0.019219
Sartorelli	Differentiation Hypothesis	FALSE	TRUE	FALSE	FALSE
	P-value	0.443901	0.007507	0.527435	0.104457
	Kolmogorov-Smirnov Test	0.145941	0.282178	0.136832	0.205149
	Average Performance	0.275287	0.336551	0.014939	0.023772

Table S2 - The specification of three synthetic datasets generated for the purpose of the validation and reproduction of the result of applying our model on real microarray datasets used for this study. Three datasets have been generated on the well-described network structure of *E. coli* (Ma *et al.*, 2004) which contains 1330 number of nodes and 2724 interactions. Average performance is measured based on SSE/Variance.

	SYN D 1	SYN D 2	SYN D 3
Burnin point	2000	2000	2000
Number of Experiments	15	15	15
Number of Samples per experiment	2	2	2
Number of Nodes	1000	1000	1000
Number of Background nodes	0	0	0
Probability for complex 2-regulator interactions	0.3	0.5	0.7
Biological noise	0.1	0.3	0.5
Experimental noise	0.1	0.3	0.5
Noise on correlated inputs	0.1	0.3	0.5
Number of External nodes	0	0	0
Number of Correlated external nodes	0	0	0
Sub network selection method	Cluster Addition		
Random seed	13	13	13