PREDICTION OF A GENE REGULATORY NETWORK IN CANCER CELLS

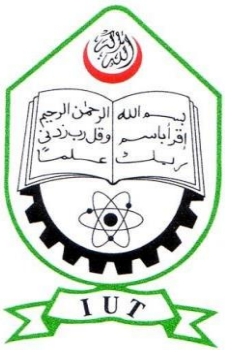
by

Mustadir Mahmood Anik (134437)

Nabil Farhan (134431)

A Thesis Submitted to the Academic Faculty in Partial Fulfillment of the Requirements for the Degree of

BACHELOR OF SCIENCE IN COMPUTER SCIENCE AND ENGINEERING



Department of Computer Science and Engineering

Islamic University of Technology (IUT)

Gazipur, Bangladesh

PREDICTION OF A GENE REGULATORY NETWORK IN CANCER CELLS

Approved by:

Tareque Mohmud Chowdhury

Supervisor and Assistant Professor,

Department of Computer Science and Engineering,

Islamic University of Technology (IUT)

Boardbazar, Gazipur-1704.

Date: ……

Candidates' Declaration

It is hereby declared that this thesis or any part of it has not been submitted elsewhere for the award of any degree or diploma.

Signature of the candidate

Mustadir Mahmood Anik

Student no. 134437

Department: CSE

Signature of the candidate

Nabil Farhan

Student no. 134431

Department: CSE

Signature of the Supervisor

Tareque Mohmud Chowdhury

Assistant Professor

Contents

Abstract:

The invention of high throughput technology like microarrays has enabled us to better understand how different cellular components interact. Thus created great interest in the field of Gene Regulatory Network(GRN) in particular. The interplay of interactions between DNA, RNA and proteins leads to genetic regulatory networks (GRN) and in turn controls the gene regulation. Directly or indirectly in a cell such molecules either interact in a positive or in
repressive manner therefore it is hard to obtain the accurate computational models through which the final state of a cell can be predicted with certain accuracy. A variety of models and methods have been developed to address different aspects of GRN. Using the Time series data and applying it to these models researchers generate meaningful results i.e. how genes interact with one another. However results found are not of much accuracy due to presence of intrinsic noise of the expression measurements. In order to produce more accurate GRNs using one of the many models available, a new technique is proposed here.

Motivation: Cancer is a complex disease, triggered by mutations in multiple genes and pathways. There is a growing interest in the application of systems biology approaches to analyze various types of cancer-related data to understand the overwhelming complexity of changes induced by the disease.

Results: We reconstructed a regulatory module network using gene expression, microRNA expression and a clinical parameter, all measured in lymphoblastoid cell lines derived from patients having aggressive or non-aggressive forms of prostate cancer. Our analysis identified several modules enriched in cell cycle-related genes as well as novel functional categories that might be linked to prostate cancer. Almost one-third of the regulators predicted to control the expression levels of the modules are microRNAs. Several of them have already been characterized as causal in various diseases, including cancer. We also predicted novel microRNAs that have never been associated to this type of tumor. Furthermore, the condition-dependent expression of several modules could be linked to the value of a clinical parameter characterizing the aggressiveness of the prostate cancer. Taken together, our results help to shed light on the consequences of aggressive and non-aggressive forms of prostate cancer.

Introduction:

Biological system has been traditionally studied by explaining behavior of individual cell components. Even though this knowledge is helpful it does not allow us to understand how complex cell components like gene work. Through the advent of high throughput technologies like microarrays, understanding Gene functionalities and therefore Gene Regulatory Network have become much more easier than it has been in the past.

To understand Gene Regulatory network, first we need to what gene really is. Gene is a section of DNA which contains instructions for making protein. This protein is then responsible for a particular characteristic like hair or eye color. To make protein messenger RNAs (mRNA) act as a template where the instruction from gene is transcribed or copied. From the mRNA strand the transcribed instructions are used to form a chain sequence of amino acids. This amino acid chain then twists and curls to form a complex 3 dimensional shape which is called a protein molecule. This protein molecule is then responsible for a particular characteristic.

During the past century, the basic strategy to decypher biological functions was essentially to concentrate efforts on a very limited set of molecules of interest. This reductive or gene-centric approach has had, and still has, an enormous success,

producing immediately applicable results in all areas of molecular biology knowledge. However, it has become clear that biological function can rarely be assigned to an individual molecule but is rather the result of the interactions among a discrete set of various types of molecules (proteins, RNA, metabolites, etc.). Those functional modules are a critical level of biological organization that cannot be identified by the study of their individual components (Hartwell *et al.*, 1999). One of the main goals of systems biology is to determine those modules and their components, by data-mining and integrating high-throughput 'omics' data.

Cancer is essentially a genetic disease, characterized by an uncontrolled proliferation and survival of damaged cells, resulting in tumor formation. Unlike other diseases, such as cystic fibrosis or muscular dystrophy, there is no single gene defect that directly 'causes' cancer. Cells have multiple safeguards to prevent

the effects of mutations appearing in various cancer genes, and it is only when several of those genes are affected that an invasive and potentially lethal tumor develops (Vogelstein and Kinzler, 2004). The picture is further complicated by the fact that new classes of molecules like microRNAs (miRNAs) have been shown to play a crucial role in tumorigenesis, and therefore should be taken into account (Esquela-Kerscher and Slack, 2006). Prostate cancer is the third most common cancer in men worldwide and occurs principally in the United States, Canada and northwestern Europe, but is uncommon in Asian countries and South America (Quinn and Babb, 2002). Prostate cancer is a complex disease, and finding the genetic causes of this disease has proven to be difficult, even if genome-wide association studies have recently detected a number of genetic variants, gene fusions and expression signatures associated with this disease (Witte, 2008). Furthermore, the progression of prostate cancer is also complex, with 'only' 10% of the patients being diagnosed with an aggressive form that can evolve to threaten their life. The determinants of this outcome are largely unknown (Lu-Yao *et al.*, 2002).

There is an increasing interest in systems biology approaches for the discovery of genes associated with cancer (Hood *et al.*, 2004; Hornberg *et al.*, 2006). Those approaches help to simplify the overwhelmingly complex picture that is often coming out of more traditional approaches by constructing more easily interpretable network representations of the underlying system and deriving concrete, experimentally verifiable hypotheses. The integration of clinical data in a robust framework that would allow the identification of modules that are pathologically altered in disease has been identified as one of the major challenges for network biology (Barabási and Oltvai, 2004).

Here, we used the LeMoNe algorithm to reconstruct a regulatory module network linked to prostate cancer using a large dataset of lymphoblastoid cells samples for which expression levels were measured for genes as well as miRNAs. LeMoNe uses ensemble-based probabilistic optimization techniques to identify clusters of co-expressed genes and their putative regulators (Joshi *et al.*, 2008, 2009; Michoel *et al.*, 2007). The algorithm has been validated and applied on various biological data sets (Michoel *et al.*, 2009; Vermeirssen *et al.*, 2009). Recently, we applied it to a set of cancer samples of various origins, for which expression data were available for both genes and a limited set of miRNAs. A couple of miRNAs were identified as high-scoring regulators for several modules of co-expressed genes, and a miRNA was validated experimentally as a regulator of a module linked to epithelial homeostasis, with a possible role in epithelial to mesenchymal transition (Bonnet *et al.*, 2010). So far, we used expression data measurements to assign regulators to clusters of co-expressed genes, but in this study we further extended the algorithm to simultaneously evaluate a heterogeneous set of candidate regulators which can be continuous-valued or discrete. In addition to combining transcription factors and miRNAs as regulators, we have also associated a clinical parameter to the condition-dependent expression levels of a module, gaining further insight in the regulatory processes.

Problem Statement:

Cancer is one of the most diseases faced by mankind. Every year thousands of people all over the world die due to this disease. Cancer is a form of tumor. A group of cells abnormally divide to form a lump of cells (tumor) which eventually leads to cancer. Although chemotherapy has proved to be helpful in cancer diagnosis in some cases in recent years, it may not cure the patient completely.

So what we are proposing is a technique that will look into the functionalities of genes in cancer cells. Cells have genes which are responsible for the regulation of cell division. Due to some mutations, these genes involved in the process are changed (Mutated), i.e. the instructions which these genes contain are changed. Mutation of genes causes the cell to divide abnormally (uncontrolled cell division and hence increase in cell number). This is cancer develops in humans.

We are suggesting the use of Gene Regulatory Network (GRN) to successfully identify how the genes of cancer cell are interacting with one another and therefore figure out which genes serve to influence the expression of genes that are responsible for the abnormal cell division in cancer cells.

Using data from our proposed technique diagnosis of cancer can be done at the gene level and hence might provide a more efficient approach to the treatment of cancer.

Gene Regulatory Network (GRN) is the study of how individual genes interact with one another. The expression of a particular gene results in a protein

and this protein molecule may in turn cause the expression of a completely different gene. Thus expression of a gene may be dependent on the expression level of a different gene. In cells, genes interact with one another to accomplish the complex functionalities of the cell, such as respiration, photosynthesis, cell division, etc.

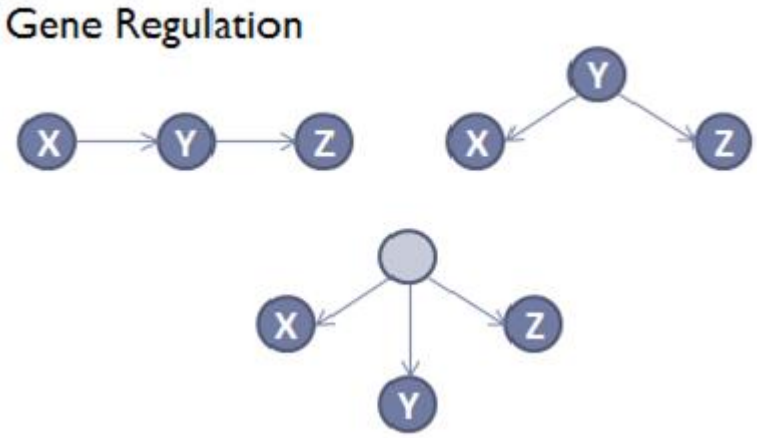


Figure: Correlation between genes

The advent of microarrays has made it easier for us to understand how genes work. Microarrays give us information about gene expression level during different time intervals. This is what we call the ***TIME SERIES DATA.*** These data are applied to one of several models developed for assessing Gene Regulatory

Network to find dependencies between different genes. Gene Regulatory Networks are needed to indicate the interrelation between genes in the genomic level. Such information is useful for disease treatment, drugs creation purposes and to understand the activity of living organisms in the molecular level.

REFERENCE FROM:

Identifying Gene Regulatory Networks from Gene Expression Data (by Vladimir Filkov *University of California, Davis***)**

Research Challenges

1. Analyzing & Understanding the whole methodology clearly.
2. Examined the previous works on this methodology.
3. Collecting The Dataset Which we want to work on.
4. Works with a new statistical software **R**.
5. Detection of hub genes.
6. Module Construction

How Gene Regulatory Network works:

.The discovery of gene regulatory networks (GRN) from time series data of gene expression observations can be used to:

1. Identify important genes in relation to a disease. Example: Cancer.

2. Gain an understanding of the dynamic interaction between genes.

3. Predict the gene expression values at future time points

A **gene regulatory network (GRN)** is a collection of molecular regulators that interact with each other and with other substances in the cell to govern the gene expression levels of mRNA and proteins. These play a central role in morphogenesis, the creation of body structures, which in turn is central to evolutionary developmental biology (evo-devo).

The regulator can be DNA, RNA, protein and complexes of these. The interaction can be direct or indirect (through transcribed RNA or translated protein). In general, each mRNA molecule goes on to make a specific protein (or set of proteins). In some cases this protein will be structural, and will accumulate at the cell membrane or within the cell to give it particular structural properties. In other cases the protein will be an enzyme, i.e., a micro-machine that catalyses' a certain reaction, such as the breakdown of a food source or toxin. Some proteins though serve only to activate other genes, and these are the transcription factors that are the main players in regulatory networks or cascades. By binding to the promoter region at the start of other genes they turn them on, initiating the production of another protein, and so on. Some transcription factors are inhibitory.

In single-celled organisms, regulatory networks respond to the external environment, optimizing the cell at a given time for survival in this environment. Thus a yeast cell, finding itself in a sugar solution, will turn on genes to make enzymes that process the sugar to alcohol.[1] This process, which we associate with wine-making, is how the yeast cell makes its living, gaining energy to multiply, which under normal circumstances would enhance its survival prospects.In multicellular animals the same principle has been put in the service of gene cascades that control body-shape.[2] Each time a cell divides, two cells result which, although they contain the same genome in full, can differ in which genes are turned on and making proteins. Sometimes a 'self-sustaining feedback loop' ensures that a cell maintains its identity and passes it on. Less understood is the mechanism of epigenetics by which chromatin modification may provide cellular memory by blocking or allowing transcription. A major feature of multicellular animals is the use of morphogen gradients, which in effect provide a positioning system that tells a cell where in the body it is, and hence what sort of cell to become. A gene that is turned on in one cell may make a product that leaves the cell and diffuses through adjacent cells, entering them and turning on genes only when it is present above a certain threshold level. These cells are thus induced into a new fate, and may even generate other morphogens that signal back to the original cell. Over longer distances morphogens may use the active process of signal transduction. Such signaling controls embryogenesis, the building of a bodyplan from scratch through a series of sequential steps. They also control and maintain adult bodies through feedback processes, and the loss of such feedback because of a mutation can be responsible for the cell proliferation that is seen in cancer. In parallel with this process of building structure, the gene

cascade turns on genes that make structura proteins that give each cell the physical properties it needs.

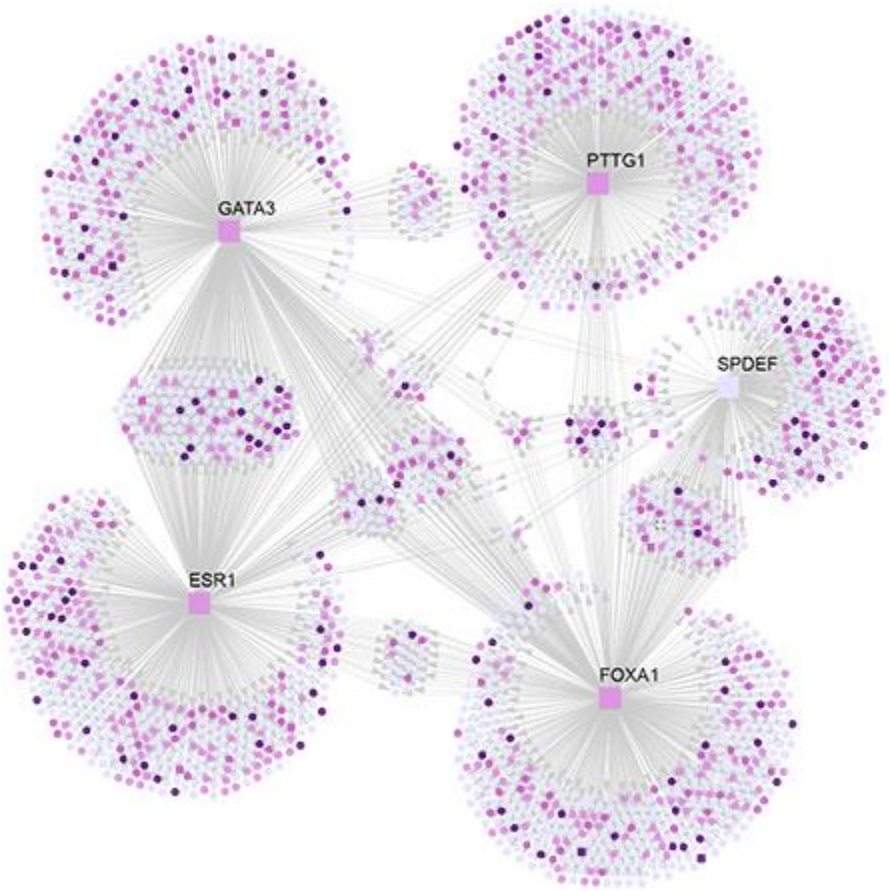


Fig: A simplified model of Gene Regulatory Network,.

Gene Regulatory Networks have two main processes.

1) The first is called **pre-processing**. Using MicroArrays (a high throughput technology) we gather expression levels of different inside a cell (cancer cells in our case) over a specified period of time. These gene expression levels for individual genes are arranged into a table which is called the **TIME SERIES DATA.**

2) 2)The next step is the modeling of gene regulatory networks using algorithms to derive a Gene Regulatory Network topology from the time series data found in the previous step. This GRN can be used to identify which genes are responsible for influencing the expression of mutated genes that result in abnormal cell division

The major approaches that deals with the modeling of gene regulatory networks involve

1. Differential equations.
2. Stochastic models
3. Evolving connectionist systems
4. Boolean networks
5. Generalized logical equations
6. Threshold models
7. Petri nets

8. Bayesian networks
9. Directed and undirected graphs
10. Learning Module Networking (LeMoNe)

Existing Techniques for pre-processing:

Previously Genetic Algorithm (GA) was applied to reduce the number of gene samples from the time series data. Then one of the above models is used to derive the Gene Regulatory Network for this reduced gene sample. Using Genetic Algorithm (GA) can have some limitations. These limitations reduce the accuracy of the derived Gene Regulatory Network. So in this paper we will be proposing an alternate approach to reducing the number of gene sample size to be used in any one of the networks to derive the Gene Regulatory Network.

The reason behind trying to find an alternate approach instead of using Genetic Algorithm is mentioned below using references from the following paper.

REFERENCE FROM:

Advantages and Disadvantages of Genetic algorithms for clustering (Abul Hashem Beg and Md Zahidul Islam)

There are some limitations of Genetic Algorithm (GA) based clustering technique.

Some of which are:

(1) Many existing Genetic Algorithm techniques randomly generate the number of genes. The quality of the genes is unlikely to be high due to random selection process.

(2) An existing technique called the GenClust technique generates high quality genes in the initial population. Thereby obtaining a good clustering result. However, the complexity of the initial population selection is very high.

(3)Many GA based clustering suffers from degeneracy. The degeneracy mainly occurs when multiple chromosomes represent the same solution. Degeneracy can lead to an inefficient solution. In order to avoid the degeneracy, an existing technique called GAGR [10] introduces a gene-rearrangement approach. However, the gene-rearrangement approach used in GAGR requires the same size (i.e. the same number of genes) of pair chromosomes, which participates in crossover operation. Moreover, GenClust also uses a gene-rearrangement operation that can re-arrange the chromosome pair with different sizes. However, the gene-rearrangement used in GenClust can handle a dataset with low dimensions. Therefore, the techniques that can handle the gene-rearrangement for the data set with high dimension are desirable.

(4) Moreover, the time complexities of some GA-based clustering techniques are $O(nm2+n2m)$,$O(n2+m2)$,$O(n2)$ respectively. Therefore, reducing the time complexity for GA-based clustering techniques is also highly desirable. Fitness value is calculated repeatedly which might be computationally expensive for some problems. Being stochastic, there are no guarantees on the optimality or the quality of the solution. If not implemented properly, the GA may not converge to the optimal solution.

Our Proposed Pre-Processing technique:

In this paper we are proposing an alternate way of pre-processing which does not require the use of genetic Algorithm (GA). Our pre-processing techniques will reduce the size of the gene sample to be implemented by the algorithms to derive the Gene Regulatory Network (GRN). This will enable us to remove the limitations of the existing technique (Genetic Algorithm) used to reduce the gene sample size.

We are suggesting the following technique:

1. Finding mean expression level of each gene from time series data and then calculating the perturbation level for that gene. Firstly we will assume a hypothetical threshold value (Tm). Then we will find the mean expression levels of each gene over a specified time interval. If the mean value is found to be less than the hypothetical threshold value (Tm), then we will discard

that gene. If the mean value found is greater than or equal to the hypothetical threshold value (Tm), then we will move on to the next step.

The next step to this technique is finding the perturbation values of each of the selected gene from the sample we found by comparing mean value with Tm. Then we will find the mean perturbation value (this will be our threshold Perturbation value ,Tp).

Now we will accept only those genes which have a perturbation value higher than or equal to the threshold Perturbation value, Tp, and implement these genes in any one algorithm we find suitable for deriving the Gene regulatory Network.

Using mean to find perturbation:

This technique allows us to find the perturbation level for each gene by first calculating the mean expression level of each gene.

WORKING PROCEDURE**:**

Using the following **time series data from microarray,** that shows the expression levels of genes at different time intervals:

	time(t1)	time(t2)	time(t3)	time(t4)	time(t5)
cg00000292	0.518	0.5537	0.5936	0.3373	0.5682
cg00002426	0.238	0.0344	0.3553	0.1159	0.2619
cg00003994	0.0875	0.1214	0.1615	0.063	0.3153
cg00005847	0.5204	0.5422	0.4792	0.2698	0.4583
cg00006414	0.0124	0.0101	0.0243	0.0185	0.0255
cg00007981	0.0199	0.0099	0.0134	0.0142	0.0084
cg00008493	0.9887	0.989	0.9801	0.9926	0.9864
cg00008350	0.0265	0.0199	0.027	0.025	0.0239
cg00009407	0.0419	0.0436	0.0465	0.0476	0.0547
cg00010193	0.5831	0.6	0.5765	0.589	0.5856

Step 1: Finding mean expression level of each gene from time series data

For cg00000292 we get,

$$\text{Mean} = \frac{\sum Xi}{n}$$

$$= \frac{.528+.5537+.55935+.3373+.5682}{5} = 0.51436$$

Finding mean expression values for all the above genes we get the following table:

	time(t1)	time(t2)	time(t3)	time(t4)	time(t5)	Mean
cg00000292	0.518	0.5537	0.5936	0.3373	0.5682	0.51416
cg00002426	0.238	0.0344	0.3553	0.1159	0.2619	0.35763
cg00003994	0.0875	0.1214	0.1615	0.063	0.3153	0.14974
cg00005847	0.5204	0.5422	0.4792	0.2698	0.4583	0.45398
cg00006414	0.0124	0.0101	0.0243	0.0185	0.0255	0.01816
cg00007981	0.0199	0.0099	0.0134	0.0142	0.0084	0.01316
cg00008493	0.9887	0.989	0.9801	0.9926	0.9864	0.98736
cg00008350	0.0265	0.0199	0.027	0.025	0.0239	0.02446
cg00009407	0.0419	0.0436	0.0465	0.0476	0.0547	0.04686
cg00010193	0.5831	0.6	0.5765	0.589	0.5856	0.58684
					Tm:	0.3

Here we have assumed the **hypothetical mean expression level, Tm to be 0.3**

We will select the **hypothetical mean expression level, Tm** via trial and error method. We can use different values for this hypothetical value, Tm until we get the most optimal solution.

Any mean value that is below this hypothetical value Tm will be discarded. In the above figure the mean values of gene expression with values greater than Tm are highlighted i.e.

mean value>=Tm

Step 2: Now we will find the perturbation value of each of the genes that have a mean value greater than Tm.

Perturbation is the summation of absolute differences between gene expression levels at intervals T_i , T_{i+1}, T_{i+2}, T_{i+3}…… T_{i+n}.

Formula for calculating perturbation:

$$\text{Perturbation} = \sum_{i=1}^{n-1} |X_i - X_{i+1}|$$

For cg00000292, we get the following perturbation,

Perturbation = |.518-.5537| + |.5537-.5936| + |.5936-.3373| + |.3373-.5682|

=.5628

By applying the above equation we get the following table of perturbation values.

Here we get the mean of the perturbation value, Tp= .40792

	time(t1)	time(t2)	time(t3)	time(t4)	time(t5)	Mean	Perturbation
cg00000292	0.518	0.5537	0.5936	0.3373	0.5682	0.51416	0.5628
cg00002426	0.238	0.0344	0.3553	0.1159	0.2619	0.35763	0.9099
cg00003994	0.0875	0.1214	0.1615	0.063	0.3153	0.14974	X
cg00005847	0.5204	0.5422	0.4792	0.2698	0.4583	0.45398	0.4827
cg00006414	0.0124	0.0101	0.0243	0.0185	0.0255	0.01816	X
cg00007981	0.0199	0.0099	0.0134	0.0142	0.0084	0.01316	X
cg00008493	0.9887	0.989	0.9801	0.9926	0.9864	0.98736	0.0279
cg00008350	0.0265	0.0199	0.027	0.025	0.0239	0.02446	X
cg00009407	0.0419	0.0436	0.0465	0.0476	0.0547	0.04686	X
cg00010193	0.5831	0.6	0.5765	0.589	0.5856	0.58684	0.0563
						Tm=.3	Tp= 0.40792

We have discarded all the genes with **perturbation value < .40792**

The selected gene sample has been highlighted in the perturbation column.

We have thus reduced the gene sample to only a few genes. Our technique finds out those genes that have dependencies with each other. The rest are discarded. We can now apply this reduced gene sample to an GRN algorithm which will derive the desired GRN topology.

Implementation of Learning Module Network (LeMoNe) Algorithm on the pre-processed data

We designed and tested the LeMoNe (Learning Module Networks) algorithm in previous studies (Joshi *et al.*, 2008, 2009; Michoel *et al.*, 2007). The algorithm extends the method of Segal *et al.* (2003) to infer regulatory modules and their specific regulators from gene expression data by using a more representative solution extracted from an ensemble of possible statistical models to explain the data. LeMoNe infers a module network in two major stages.

The first one is a two-way clustering of genes and conditions, using a Gibbs sampling procedure (Joshi *et al.*, 2008).

In order to avoid local optima, multiple clustering solutions are generated and subsequently integrated in a final set of so-called tight clusters, corresponding to sets of genes that are frequently associated across all the clustering solutions.

In the second stage, the algorithm infers a prioritized list of regulators for each cluster of co-expressed genes. More precisely, a hierarchical tree is build by grouping sets of conditions (corresponding in this case to samples taken from different patients) having similar means and standard deviation. Regulators are assigned to each node of the tree by logistic regression on the regulator expression values to predict the assignment of conditions to each side of the tree node (Joshi *et al.*, 2009). Regulators having a distinct expression pattern on each side of a given tree node will get a high probabilistic score. Multiple statistically equivalent partitions of conditions are generated for each cluster of co-expressed genes and an ensemble approach is used to sum the strength with which a regulator participates in each regulatory tree. A global score is calculated which reflects the overall statistical confidence, and which is used for prioritizing the whole list of regulators for a given set of co-expressed genes. The mathematical details of the algorithm can be found in Joshi *et al.*(2009).

Integrating discrete and heterogeneous continuous-valued regulators

As explained above, regulators are assigned to a co-expression cluster by using logistic regression on the binary splits of a set of hierarchically linked condition clusters. More precisely, let $\mathcal{C}0$ and $\mathcal{C}1$ be two disjoint sets of conditions. Given a regulator with expression value x in some condition, our model assumes there is a (hidden) binomially distributed random variable Y such that Y = 0 if the condition is assigned to $\mathcal{C}0$ and Y = 1 if it is assigned to $\mathcal{C}1$, with probability

$$p(Y=1\,|\,x)=\frac{1}{1+e^{-\beta(x-z)}}$$

For a continuous-valued regulator, the training data for a regulator *R* consists of a set of expression values *xm* across all measured conditions *m*. Furthermore, given the partition of conditions and their hierarchical tree, we know at each tree node which conditions *m* belong to $\mathcal{C}_0$ and which to $\mathcal{C}_1$. Hence, using Bayes' rule, we can determine the parameters β and *z* which maximize the posterior probability of

assigning regulator *R*. This posterior probability is then used as the score for *R* at this particular tree node and combined with the scores at other nodes to compute a global assignment score. The parameter *z* is interpreted as a *split value*, meaning if *R* is highly expressed ($xm > z$) the condition is assigned to one side of the split and if *R* is lowly expressed ($xm < z$) to the other side. The parameter β is determined by how well a regulator fits the separation of conditions: if $xm > z$ for all $m \in \mathcal{C}_1$and $xm < z$ for all $m \in \mathcal{C}_0$ (or vice versa), we can take $\beta = +\infty$ and obtain a maximal posterior probability. If there is no split value which achieves a good separation of conditions, β will be close to 0 leading to low values of the posterior probability. See Joshi *et al.* (2009) for more details.

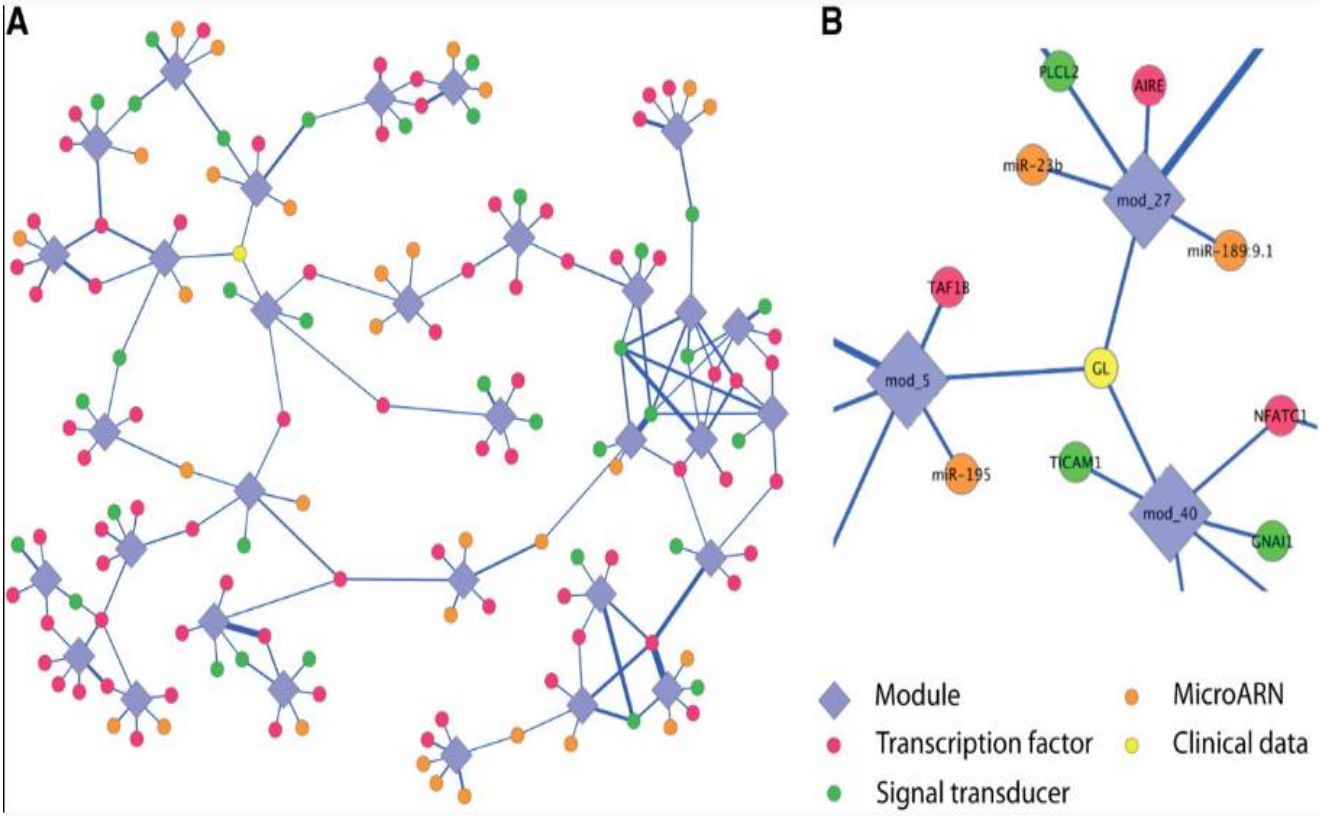
Clearly, there is no need for the values *x* to be comparable in absolute terms to the expression values determining the co-expression clusters. This is exploited to assign miRNA regulators. Furthermore, there is also no need for the values *x* to be continuous. In this article, we considered discrete regulators which can take two values, say 0 and 1. Then the parameter *z* becomes redundant and we set it $z = 0.5$, while β is determined as before by maximizing the posterior probability. As we are using a probabilistic model and the final regulator score is defined by a posterior probability, the scores of mRNA, miRNA and discrete regulators can all be integrated and compared on the same scale to determine the final module network with heterogeneous regulators.

Results and Discussions

The set of 43 tight clusters was used as input for the second stage of the algorithm, the assignment of regulators. The probabilistic score calculated for each regulator reflects how well its expression profile predicts the condition-dependent expression level of the genes in a cluster. Furthermore, we can use this score on heterogeneous types of regulators, including ones having discrete values (see Section 2). For this study, we have used three different types of regulators. First, we selected all transcription factors and signal transducers from the gene expression dataset, using the GO categories 'transcription factor activity' (GO:0003700) and 'signal transducer activity' (GO:0004871). This selection resulted in a set of 1558 genes. Second, we added a set of 735 microRNA expression profiles that were measured on the same samples, but using a distinct microarray platform (Wang *et al.*, 2009a, b). Third, we also used as a 'regulator' a clinical parameter, the Gleason grade, a discrete score assigned by a pathologist based on the microscopic appearance of prostate tissue biopsies. High values of Gleason grade are linked to more aggressive forms of prostate cancer characterized by a worse prognosis for the patient. The 90 samples in the dataset have been classified as 'high' or 'low' Gleason score.

A total of 77 374 regulator–module assignments were made by the algorithm, from which we selected the top 1% as high-scoring candidate regulators (774 regulator–module pairs). For each regulator assigned to a module, the algorithm is also selecting another one at random, thus defining a distribution of randomly assigned regulators. In this study, the distribution of all random regulators has a median score of 9.37, with a maximal score of 60.11. On the other hand, the top regulators (i.e. the top 1% of all assigned regulators) have a median score of 228, with a

minimum value of 107.47. Therefore the minimum score for a top regulator is still 3.8 times higher than the maximal score for a randomly assigned regulator, thereby demonstrating that the top regulators score is far greater from what could be expected by chance. There are 496 unique regulators in the top 1% selection. Most of the regulators are assigned to one cluster (68%), but some are assigned to two or more (Figs 1 and 2). Within this set, a total of 148 miRNAs have been selected (30% of all high-scoring regulators). Some miRNAs are also assigned to more than one cluster (Figs 1 and 2).



(**A**) Simplified representation of the module network inferred by the LeMoNe algorithm. Clusters of co-expressed genes have diamond shapes, while regulators are symbolized by circles. The color of the circle correspond to a given type of regulator. The thickness of the edges is proportional to the score of a regulator for a given module. For clarity, some clusters are not represented and we have limited the regulators to six per module. (**B**) Zoom on the module network representation. The yellow regulator labeled GL represent a clinical parameter, the Gleason score, which is connected to three different clusters.

Conclusion

In this study, we have applied a module network algorithm to a large expression data set measured on lymphoblastoid cell lines coming from patients having different forms of prostate cancer. Compared to our previous applications of the algorithm, we have further extended it to simultaneously evaluate a heterogeneous set of candidate regulators which can be continuous-valued or discrete.

We predicted a module network of 43 modules of co-expressed genes with their associated high-scoring regulators. Most of the modules show enrichment for specific GO categories. Several of those categories are related to cell cycle and

mitosis activities, which is consistent with previous studies on the same dataset. Almost 30% of the predicted regulators are miRNAs, and many of them have been characterized as causal in many diseases, including cancer. Our results also suggest novel miRNA candidates that could be linked to prostate cancer. This study also associate the Gleason score, a clinical parameter to modules enriched in cell growth and mitosis.

Our study clearly demonstrate the interest of systems biology approaches to study cancer and its consequences, more particularly by the integration of heterogeneous sets of candidate regulators. This type of analysis can be applied to various cancer types and tissues for which relevant expression data for mRNA, miRNA and various clinical parameters are available.

References:

- Bandres E, et al. ,microRNA-451 regulates macrophage migration inhibitory factor production and proliferation of gastrointestinal cancer cells, *Clin. Cancer Res.*, 2009
- Barabási A, Oltvai Z. Network biology: understanding the cell's functional organization, *Nat. Rev. Genet.* , 2004
- Bonnet E, et al. Module network inference from a cancer gene expression data set identifies microRNA regulated modules, *PLoS ONE* , 2010
- Michoel T, et al. Validating module network learning algorithms using simulated data, *BMC Bioinformatics*, 2011
- Michoel T, et al. Comparative analysis of module-based versus direct methods for reverse-engineering transcriptional regulatory networks, *BMC Syst. Biol.* , 2009
- PollardJ. , Tumour-educated macrophages promote tumour progression and metastasis, *Nat. Rev. Cancer* , 2008
- XiY et al. ,Differentially regulated micro-RNAs and actively translated messenger RNA transcripts by tumor suppressor p53 in colon cancer, *Clin. Cancer Res.* , 2006
- Joshi A, et al. Analysis of a Gibbs sampler method for model-based clustering of gene expression data, *Bioinformatics,*2008