Identifying Transcription Factors and microRNAs as Key Regulators of Pathways Using Bayesian Inference on Known Pathway Structures

Damian Roqueiro, Lei Huang and Yang Dai
Department of Bioengineering
University of Illinois at Chicago
Chicago, USA
{droque1, lhuang7, yangdai}@uic.edu

Abstract—Transcription factors and microRNAs are both known to regulate gene expression in eukaryotes in a sequence-specific manner. This has led to the creation of numerous computational approaches that aim at predicting what genes are the targets of certain transcription factors and microRNAs. These methods, although powerful, provide a static snapshot of how genes may be regulated and are often plagued by the presence of false positives. We propose a method that combines: a) transcription factors and microRNAs that are predicted to target genes in pathways, with b) microarray expression profiles, in conjunction with c) the known structure of molecular pathways. These elements are integrated in a Bayesian network that allows the identification of the main regulators in different pathways, based on probability inference, of ER+ and ER– tumors.

Keywords—Bayesian network; probability inference; transcription factor; microRNA; random forest; breast cancer microarray data

I. INTRODUCTION

Transcription factors (TFs) and microRNAs are well-known regulators of gene expression. The former bind directly to the regulatory regions of genes whereas the latter regulate the expression of genes at a post-transcriptional stage. Although they have different mechanisms of regulation, evidence suggests that TFs and microRNAs regulate target genes in a coordinated way. In order to facilitate the elucidation of these regulation mechanisms, we propose an integrative approach to analyze mRNA and microRNA expression data together with a probabilistic framework based on Bayesian inference.

Bayesian networks [1] have been extensively used to analyze microarray experiments. In this context, the main goal was to use expression data to infer interactions and statistical dependencies among genes. These dependencies were, in turn, used to learn the dynamic structure of a regulatory network [2]. This methodology has been the foundation for a large amount of algorithmic approaches. In all these cases, the Bayesian network (BN) –or its more generic dynamic counterpart (DBN)– was used as a tool to reverse engineer the gene network, i.e., the interactions between genes were inferred from observational data.

In this work, we do not focus on the task of learning the structure of the BN from expression data. Our goal is to use a known network structure, describing interactions between genes and proteins, for Bayesian inference. The network structure can be any experimentally confirmed interaction network (for example, pathways obtained from KEGG [3] or from the Pathway Interaction Database [4]). We extend the network to include TFs and microRNAs that are predicted to target nodes in the network. We further compute conditional probabilities between the nodes on the extended network using expression data. Finally, we create a BN that receives as evidence a list of differentially expressed genes and provides as output a ranked list of TFs and microRNAs that best explain the expression level of genes in the network. As a result of this, the output TFs and microRNAs are hypothesized to be putative regulators of the pathway.

We applied this methodology to the analysis of mRNA and microRNA expression data generated from seven breast tumor studies [5-11]. The patients in these studies were divided into two groups: estrogen receptor positive (ER+) and estrogen receptor negative (ER–). ER+ and ER– breast tumors display different molecular patterns in terms of cell differentiation, proliferation, survival, invasion and angiogenesis. Understanding the distinct molecular mechanisms in tumors with different ER status (ER+ and ER–) will provide insight into potential novel targets for breast cancer treatment [12]. Here we demonstrate the usefulness of our integrative method in uncovering the relationships among TFs, microRNAs and pathway genes that are associated with these tumors.

II. METHODS

We started our integrative approach by discretizing the expression data of mRNAs and microRNAs from ER+ and ER– tumor microarray profiles. We subsequently obtained the known structure of 31 KEGG pathways and pre-processed them to guarantee that: a) there were no cycles and b) all nodes in the pathway had expression data. For the nodes that passed this pre-processing step we proceeded to obtain lists of TFs and microRNAs that are predicted to target the nodes. We then ranked the different TFs and microRNAs based on their ability to predict the expression level of a target gene. We obtained one ranking list per gene and extended the pathways to include the top 5 TFs and top 3 microRNAs for each gene in the pathway. Finally, a Bayesian network was created for each extended pathway. Inference was done by entering, as evidence, the statuses (discrete values) of differentially expressed genes in the pathway. The marginal probabilities were approximated for all unobserved nodes. From these, the TFs or microRNAs with the largest marginals are considered the most probable

978-0-7695-4574-5/11 $26.00 © 2011 IEEE
DOI 10.1109/BIBM.2011.120

2011 IEEE International Conference on Bioinformatics and Biomedicine
TABLE I. ANALYZED ER+/ER– EXPRESSION DATASETS

| Dataset name         | Source          | Number of samples |
|----------------------|-----------------|-------------------|
| Boersma (mRNA)       | GSE5847 [5]     | 41 52             |
| Desmedt (mRNA)       | GSE7390 [6]     | 107 51            |
| Miller (mRNA)        | GSE3494 [7]     | 213 34            |
| Minn (mRNA)          | GSE2603 [8]     | 57 42             |
| Sotiriou (mRNA)      | GSE2990 [9]     | 74 24             |
| Wang (mRNA)          | GSE2034 [10]    | 209 77            |
| Enerly (mRNA)        | GSE19783 [11]   | 60 35             |
| Enerly (microRNA)    | GSE19536 [11]   | 60 35             |

regulators of expression in the pathway. An overview of the entire methodology is illustrated in Fig. 1.

A. Pre-processing of raw microarray data

The raw data from seven studies of ER+/ER– breast tumors [5-11] were downloaded from the Gene Expression Omnibus. Table 1 provides details of the source of the data and the number of samples for each tumor type. The first six studies contain only mRNA expression profiles whereas the last one (Enerly) has concurrent mRNA and microRNA expression profiles on ER+/ER– breast tumors. Herein, we will refer to the datasets using the name provided in Table 1.

The profiling in the Enerly study was performed using duplicate hybridizations on Agilent “Human microRNA Microarray Kit (V2)” and human genome 4X44K one-color oligo array. For the rest of the studies, gene profiling was carried out using Affymetrix high-density oligonucleotide arrays (HG-U133A). Gene expression analysis was performed using packages in Bioconductor [13]. The Robust Multichip Averaging algorithm (RMA) with quantile normalization was used for normalization. Additionally, to minimize the noise level in the subsequent task of data discretization, Affymetrix detection calls were obtained to identify probesets with low or no level of expression.

The raw microRNA data were normalized with the RMA algorithm using the AgiMicroRna package [14]. The mRNA data in the Enerly study were already normalized.

B. Differential expression analysis of the mRNA-Enerly dataset

Differentially expressed genes in the mRNA-Enerly dataset were used as evidence in the Bayesian inference process. Since the mRNA data were already normalized, our first pre-processing step was to identify the least variable mRNA probes. Probes with a coefficient of variability of less than 50% were filtered out. This left a total of 8,589 probes, all of them with a unique Entrez Gene ID, that were subsequently analyzed for differential expression between the ER+ and ER– tumors. We used the limma package with the Benjamini–Hochberg correction for multiple tests. The adjusted p-value threshold was set to 0.05.

C. Microarray data discretization

In order to integrate the expression values of genes from different experiments, a procedure for data discretization was implemented on the normalized mRNA and microRNA expression data. We discretized these values based on the proximity to the mean expression value in the microarray so that each gene had one discrete value per microarray.

Probesets from different microarrays were mapped to unique Entrez IDs. The common genes between the first six datasets and the mRNA-Enerly dataset were kept. The first six datasets, all of them of the same microarray platform, had 12,025 unique gene Entrez IDs. The mRNA-Enerly dataset had 17,177 unique genes. A total of 11,514 unique Entrez IDs were common across all mRNA datasets.

Figure 1. Flowchart of data integration methodology.
Data normalization of each dataset forced its microarrays to have the same empirical distribution of intensities. As an example, the density of expression values of all the microarrays in the Boersma dataset is shown in Fig. 2a. In contrast, when only the probesets marked as Present or Marginal are considered, the density function adopts a shape closer to that of a normal distribution (Fig. 2b).

For each microarray, the mean and standard deviation of the expression values of probesets, marked as either Present or Marginal (P/M), were obtained. Then, the P/M probesets were assigned a discrete value depending on how many standard deviations away they were from the mean (Fig. 2b).

The five discrete values considered in this study were: very low, and very high (≥2σ from μ); medium low and medium high (≥1σ from μ); and medium (<1σ from μ). Probesets marked as Absent were given a discrete value of very low.

D. Structure pre-processing for KEGG pathways

The KEGG database [3] provides experimental knowledge in many forms, one of them being molecular networks called KEGG pathway maps. For our work, the pathway maps were analyzed as networks, with directed edges between the nodes representing a known interaction. The pathways analyzed were related to signaling (KEGG Ids 04150-04920) and cancer (05200-05223).

The structure of a pathway, nodes and edges was used as the backbone of a Bayesian network (BN). Before the BN could be constructed, a pre-processing step was implemented on the pathway. This pre-processing yielded a new network, based on the original pathway, with the following properties:

- No cycles: The KEGG pathway was transformed into a directed acyclic graph (DAG). Edges that created a loop were discarded.
- Nodes with expression data: The Entrez ID of each node in the pathway was checked against the list of genes that had expression data (11,514 Entrez IDs from our microarray analysis). Nodes with no expression data were removed. The parents and children (if any) of a removed node were updated to include new edges linking them.
- Interaction filtering: Only the following interactions annotated in a KEGG pathway were taken into consideration: a) gene expression relations: expression, repression and indirect effect; and b) protein-protein interactions: activation, inhibition and indirect effect.

The parsing of the raw KEGG Markup Language files was done using the package KEGGgraph [15] in R.

E. The targets of transcription factors and microRNAs

Our goal in implementing a BN for a known pathway was to identify the set of TFs and microRNAs that are putative regulators of nodes in the pathway. Therefore, the new network obtained from the previous pre-processing step needed to be expanded to include the TFs and microRNAs that are predicted to target the nodes in the pathway. We followed two different approaches to obtain which TFs and microRNAs may target a node.

1) TF target prediction. bindSDb [16] is a database we developed to store experimentally proven and predicted transcription factor binding sites (TFBSs). For the prediction portion, the database returns a set of TFs that are predicted to bind to the promoter region of a gene. The prediction of binding is based solely on sequence analysis. It uses the MATCH [17] algorithm to determine if a TF, represented by one or more position weight matrices (PWMs) from TRANSFAC (ver. 2010.1) [18], may bind to the promoter of the gene. In our work, for each gene in a pathway –or protein encoded by a gene– we obtained from bindSDb all the TFs that are predicted to bind to the promoter region of the gene (defined as +2Kb, −2Kb from the transcription start site).

Additionally, we obtained from TRANSFAC the information about the genes that encode the predicted TFs (when available). In this way, each gene, in the pathway would be associated to a set of genes whose protein products (TFs) are predicted to target gene,. If one of the predicted TFs was already present in the pathway, then it was not included as a putative regulator of gene,. 2) microRNA target prediction. All microRNA-gene predictions were downloaded from ExprTarget [19]. ExprTarget is an online human microRNA target prediction database which integrates several microRNA target prediction algorithms. The downloaded microRNA-gene predictions were further filtered by their scores. Those with a score of less than 3 were discarded, yielding a total of 18,937 unique microRNA-gene pairs.

F. Random Forest and variable importance

The previous step provided, for each gene in a pathway, a list of TFs and microRNAs that were predicted to target that gene. Ideally, we would have expanded our pathway by adding incoming edges to a gene for every TF and microRNA that targeted the gene. As it will become clear later, this was infeasible especially because of the large number of TFs that may target a gene. Table 2 shows the number of TFs and microRNAs that were predicted to target the genes of three signaling pathways. For a given node in a BN we could not afford to have 100+ parents because the conditional probability table used to maintain the state of the
node with respect to its parents would have been too large. Therefore, we needed to find a way to limit the number of TFs and microRNAs that targeted a node. To that respect, we used a machine learning approach that ranked the TFs and microRNAs for each gene. This ranking allowed us to choose, say, the top $p$ elements thus limiting the number of extra parents a node could have.

The random forest (RF) classification algorithm [20] was used as a supervised classification method on each gene of a pathway. Two RF classifiers were created to predict the expression level of a gene: one using the TFs and the other using the microRNAs that targeted the gene. For each classifier, the values of the predictor variables were the discretized expression levels of the TFs (mRNAs) or microRNAs. Our ultimate goal was not to find a classifier to predict the expression level of genes but to use RF to measure the importance of each predictor variable. In this manner, we obtained for each gene a group of TFs and microRNAs that could help differentiate the expression level of the gene across different microarrays. We used the information of the TFs predicted to regulate a gene and the microRNAs predicted to target that gene, as described in section E. Fig. 3 illustrates the layout of the data.

The supervised learning predictor for gene $g$ is defined as $T_k = (v_i, x_i)$ with $i = 1$ to $M$ where $M$ is the total number of microarrays used in the classifier. For TFs, $M = 980$ (the first six studies listed in Table 1) and for microRNAs $M = 95$ (Enerly dataset). The multi-class response vector $y$ contains the $M$ discrete expression levels of gene $g$ in the microarrays. Each vector $x_i$ has $k$ predictor variables (TFs or microRNAs that target gene $g$). $x_{ij}$ for $j = 1$ to $k$ contains the discrete expression value of predictor $j$ in microarray $i$. The values were coded according to the data discretization: from 1 through 5, where 1 = very low and 5 = very high. For each gene, an ensemble of 2,000 trees (for TFs) and 500 trees (for microRNAs) was created. One third of the variables were randomly chosen at each tree level and one third of the samples were left as out of bag. Variable importance was determined after performing permutations on the trees to assess the change in their predicting power. Each variable was assigned a mean decrease of accuracy score and the ranking of variables for the gene was based on this score. The analysis was implemented with the R package randomForest.

G. Pathway extension

At this stage, we had all the information we needed to create a BN for a pathway. The modified pathway obtained after pre-processing in section D had to be extended to accommodate the TFs and microRNAs ranked in section F. Our RF analysis output two rankings for each gene: the variable importance ranking for the TFs and for the microRNAs. A merged pathway was then created by reading the pre-processed pathway and adding as nodes the top 5 TFs and top 3 microRNAs from each gene’s ranking. The same TF could target more than one gene in the pathway. Therefore, the node for the TF was added just once with multiple edges going from this node to different target genes. The same consideration applied to the microRNAs. This newly merged pathway was then fed to the BN process.

H. Bayesian network construction

Simply put, a BN can be characterized as [21][22]:

- A directed acyclic graph $G = (V, E)$ where $V$ is a set of variables and $E$ is a set of directed edges between the variables.
- Each variable in $V$ has a finite set of mutually exclusive states.
- For each variable $B$ with parents $A_1, A_2, \ldots, A_p$ there is a set of parameter probabilities $\Gamma$ in the form of conditional probability tables (CPTs) that capture $P(B | A_1, A_2, \ldots, A_p)$.

The first two items have been addressed in previous sections (pre-processing and discretization). The creation of the CPT for a given node in the pathway was implemented in the following way:

1) If the node did not have any parents, the CPT was basically a vector representing the node’s prior. It was computed by obtaining the frequencies of each discrete value across all the appropriate microarrays (TFs and genes used the first six datasets of Table 1 whereas microRNAs used the Enerly dataset).

2) If the node had parents $A_1, A_2, \ldots, A_p$ the CPT reflected the probability of all possible combinations of states between the node and its parents. The probability of each possible combination was obtained by counting and then dividing by the total number of observations. A high-dimensional matrix $C$ of $5$-by-$5$-by-...($p + 1$)-times was used to compute the CPT. The matrix $C$ was initialized with 1s to assume that each possible combination of states was possible. Then, for each microarray, the discrete expression values of the node and its parents were obtained as a vector $v = [v_{d1}, v_{d2}, \ldots, v_{dn}, v_{node}]$. The contents of matrix $C$ at the cell $C[v_{d1}, v_{d2}, \ldots, v_{dn}, v_{node}]$ were then incremented by one. At the end, each position of $C$ was divided by the sum of all elements in $C$. The matter of what set of microarrays to use was resolved in the following way:
• If any of the node’s parents $A_1, A_2, \ldots, A_p$ was a microRNA, the Enerly dataset was used.
• Otherwise, the first six datasets listed in Table 1 were used.

This distinction was necessary because to compute the CPT of a node that had at least one microRNA as parent, we needed to process microarrays that had both, expression values for genes/TFs as well as microRNAs. Evidently, the CPT of nodes with a microRNA targeting them was created from fewer observations than nodes whose parents were only TFs or other pathway nodes.

I. Evidence and inference

An important aspect of a BN is the evidence, i.e., the values assigned to observed nodes. For evidence, the differentially expressed genes between ER+ and ER– samples in the mRNA-Enerly dataset were used. Only those differentially expressed genes that were part of a pathway (not as TFs but as KEGG pathway nodes) were used as evidence. If a gene was over-expressed in ER+, its evidence was entered as very high. Conversely, if the gene was underexpressed in ER+, the evidence value was set to very low.

Finally, for a BN with variables $X_1, X_2, \ldots, X_n$, where the evidence $e = \{X_{n+1}, X_{n+2}, \ldots, X_{n+s}\}$ and the values of variables $X_1, X_2, \ldots, X_n$ are unobserved, we would like to obtain $P(X_1, X_2, \ldots, X_n | e)$. This joint probability is defined as:

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | parents(X_i)) \quad (1)$$

Because the size of the CPT for each variable $X_i$ is exponential on the number of parents of $X_i$, this computation is prohibitive for large networks. To complicate matters further, we would like an answer to the question: what is the probability of $X_i = x$ given the evidence $e$? This requires the marginalization of $X_i$ from equation (1). Since exact inference is computationally infeasible, we have to find an approximation to the marginal probability $P(X_i | e)$.

In our work, this was achieved by using a Gibbs sampler. The marginal probabilities for all unobserved nodes were sampled at a rate of $Q \times$ number of nodes in the BN, with $Q = 250$. The value of $Q$ is an empirical number obtained from the analysis of two BNs of size 16 and 33 nodes. Because of their small size, all marginals were computed in an exact manner and then approximated using a Gibbs sampler. The approximation error was computed for different number of iterations of the sampler and $Q$ was obtained (data not shown).

The BN creation, Gibbs sampler, inference engine and marginalization of nodes were implemented with the Bayes Net toolbox (BNT) for Matlab [23].

III. EXPERIMENTAL RESULTS

When analyzing the results, we decided to focus on nodes whose marginals had one state with a probability larger than 0.8. We then compared these nodes with the parents of all differentially expressed genes in the pathway.

In the Enerly dataset, the genes RPS6KA1 and MAPK3 were under- and over-expressed in ER+ breast tumors. In the mTOR signaling pathway, both genes shared two parent nodes: IGF1 and INS (Fig. 4). The marginal probability for IGF1 put its expression value at medium with a probability of 0.98. INS had an expression value of very low with a marginal probability of 0.92.

It has been shown [24] that upon ligand binding, IGF1 activates the PI3K/Akt/mTOR pathway via IRS-1 (insulin receptor substrate-1) maintaining the proliferation of breast cancer cells. Of course, these high marginal probabilities were inferred because we knew that IGF1 targeted RPS6KA1 and MAPK3 from the pathway structure.

In addition to this, we were interested in finding if any of the TFs had any effect in determining the expression value of their targets. Upon looking at the TFs that targeted RPS6KA1 and MAPK3 we found that ELF4 had a marginal of 0.91 for the expression level medium (Table 3). ELF4 is a member of the Ets family of proteins that negatively regulates quiescence of hematopoietic stem cells [25]. In a recent study conducted in mice [26] it was shown that the mTOR pathway inhibits ELF4 in activated cytotoxic T cells. In our study, the most probable expression level for ELF4, given the evidence of differentially expressed genes in the pathway, remained at a medium level of expression.

In a similar way, microRNAs hsa-miR-26a/b were found to have high marginals for the expression level very high. Although the marginals for their target gene RPS6KA6 did not favor a particular state, we could not conclude that there was no direct effect of hsa-miR-26a/b on RPS6KA6. Nevertheless, it has been shown that hsa-miR-26a affects growth and tumorigenesis [27].

We therefore hypothesize that the TF ELF4 and the microRNAs hsa-miR-26a/b are key regulators of the mTOR pathway in ER+ breast cancer tumors. This conclusion is based on 1) inference from a BN created from the known structure of the pathway, 2) the evidence entered as differential expressed genes in ER+/ER– breast tumors and

### TABLE III. SELECTED MARGINALS FOR THE mTOR PATHWAY

| Node          | very low | medium low | medium | medium high | very high |
|---------------|----------|------------|--------|-------------|-----------|
| ELF4          | 0.09     | 0.91       |        |             |           |
| IGF1          | 0.01     | 0.01       | 0.98   |             |           |
| INS           | 0.92     | 0.01       | 0.07   |             |           |
| hsa-miR-26a   | 0.02     | 0.01       | 0.05   | 0.91        |           |
| hsa-miR-26b   | 0.01     | 0.01       | 0.15   | 0.83        |           |
| RPS6KA6       | 0.24     | 0.22       | 0.20   | 0.17        | 0.16      |
| SPI1          | 0.67     | 0.15       | 0.18   |             |           |
| SREBF1        | 0.07     | 0.68       | 0.24   |             |           |
3) the breast cancer microarrays used to populate the CPTs of the nodes in the BN.

IV. CONCLUSION AND FUTURE WORK

We proposed an integrative bioinformatics methodology that combines a) the TFs and microRNAs that are predicted to target pathway genes, with b) microarray expression profiles, in conjunction with c) the known structure of molecular pathways. All these elements were integrated into a probabilistic framework (a BN) that was used to make inferences about key TFs and microRNAs as regulators of pathways in ER+ and ER− breast tumors.

Another important use of our framework is to test hypotheses about the expression levels of TFs or microRNAs and their effect on genes. We foresee the researcher posing questions of the form: “What would the expression level of genes $g_1$ and $g_2$ be if microRNA$_A$ is expressed at a very high level?”

Several technical issues deserve further investigation. When making inference about the expression level of a gene, TF or microRNA, we would ideally want to obtain the most probable explanation (MPE) given the evidence at hand. This evidence can be tangible, i.e., obtained from a microarray experiment, or, as it was mentioned before, it can be a set of hypotheses that interest us. In either case, an exact solution to the MPE problem in Bayesian inference has proven to be impossible in general [21]. Thus, in this work we have decided to use the marginals as a proxy for MPE. In turn, we approximated the marginals for the unobserved nodes using a stochastic sampling algorithm (Gibbs sampler). We plan to improve our methodology by thoroughly examining different importance sampling algorithms that will minimize the variance between the drawn samples and the target distribution [28].

Finally, a self-imposed limitation of our model was the removal of edges that would create cycles in the network. Our next step will be to improve our probabilistic framework to use a dynamic Bayesian network (DBN) that allows for cycles and that better reflects the positive feedback present in many molecular pathways.

REFERENCES

[1] Pearl J. “Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning (UCLA Technical Report CSD-850017).” Proceedings of the 7th Conference of the Cognitive Science Society, UC Irvine. pp. 329–334, 1985
[2] Friedman N., Linial M., Nachman I., and Pe’er D. “Using Bayesian networks to analyze expression data.” J. Comput. Biol. 7, 601–620, 2000
[3] Kanehisa M., Goto S., Furumichi M., Tanabe M., and Hirakawa M. “KEGG for representation and analysis of molecular networks.” Nucleic Acids Res. 38, 355-360, 2010
[4] Schaefer C.F., Anthony K., Krupa S., Buchoff J., et al. “PID: the Pathway Interaction Database.” Nucleic Acids Res. 37:674-679, 2009
[5] Boersma B.J., et al., “A stromal gene signature associated with inflammatory breast cancer.” Int J Cancer, vol. 122, 1324-32, 2008
[6] Desmedt C., et al., “Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes,” Clin Cancer Res, vol. 14, 5158-65, 2008
[7] Miller L.D., et al., “An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival,” Proc Natl Acad Sci USA, vol. 102, 13555-5, 2005
[8] Minn A.J., et al., “Genes that mediate breast cancer metastasis to lung.” Nature, vol. 436, pp. 518-24, 2005
[9] Sotiriou C., et al., “Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis,” J Natl Cancer Inst, vol. 98, pp. 262-72, 2006
[10] Wang Y., et al., “Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer,” Lancet, vol. 365, pp. 671-9, 2005
[11] Enerly E., et al., “miRNA-mRNA Integrated Analysis Reveals Roles for miRNAs in Primary Breast Tumors,” PLoS ONE, vol. 6, p. e16915, 2011
[12] Osborne C.K. and Schiff R., “Mechanisms of Endocrine Resistance in Breast Cancer,” Annu. Rev. Med., vol. 62, pp. 233-247, 2011
[13] Gentleman R.C., Carey V.J., Bates D.M., Bolstad B., Dettling M., et al. “Biocductor: open software development for computational biology and bioinformatics” Genome Biology.5:R80, 2004
[14] Lopez-Romero P., “Pre-processing and differential expression analysis of Agilent microRNA arrays using the AgiMicroRNA Bioconductor library,” BMC Genomics, vol. 12, p. 64, 2011
[15] Zhang J.D., Wiemann S. “KEGGigraph: a graph approach to KEGG PATHWAY in R and bioconductor”. Bioinformatics (Oxford, England), Vol. 25, No. 11, 2009
[16] Roqueiro D., Frasar J., and Dai V. “bindISDb: A Binding-information Spatial Database”. Proc. of the 2010 IEEE BIBM Workshop on Data-mining of Next-Generation Sequencing Data, Hong Kong, 2010
[17] Kel A.E., Gossling E., Reuter I., Cheremushkin E., Kel-Margoulis O.V, et al. “MATCH: A tool for searching transcription factor binding sites in DNA sequences.” Nucleic Acids Res 31: 3576-3579, 2003
[18] Matsy V., Fricke E., Geffers R., Gossling E., Haubrock M., et al. “TRANSFAC: transcriptional regulation, from patterns to profiles.” Nucleic Acids Res 31:374-378, 2003
[19] Gamazon E.R., Im H.-K., Duan S., Lussier Y.A., Cox N.J., Dolan M.E., and Zhang W., “ExpreTarget: An Integrative Approach to Predicting Human MicroRNA Targets,” PLoS ONE, vol. 5, p. e13534, 2010
[20] Breiman L. “Random Forests,” Mach. Learn., vol 45, pp 5-32, 2001
[21] Kivesthout J. “Most Probable Explanations in Bayesian networks: complexity and tractability”. Technical Report ICIS–R10001, Radboud University Nijmegen, 2010
[22] Jensen F.V., Nielsen T.D. “Bayesian Networks and Decision Graphs”. Second edition. Information Science and Statistics. Springer Science. 2007
[23] Murphy K. “The Bayes Net Toolbox for Matlab.” In Computing Science and Statistics: Proceedings of the Interface, volume 33, 2001
[24] Ghayad S.E., Cohen P.A. “Inhibitors of the PI3K/Akt/mTOR pathway: new hope for breast cancer patients”. Recent Pat Anticancer Drug Discov. 5:29–57, 2010
[25] Lacorazza H. D., Yamada T., Liu Y., Miyata Y., Sivina M., Nunes J., et al. “The transcription factor MEF/ELF4 regulates the quiescence of primitive hematopoietic cells”. Cancer Cell 9: 175–187, 2006
[26] Yamada T., Gierach K., Lee P-S, Wang X., and Lacorazza H.D. “Heterogeneous Akt cell growth and tumorigenesis of nasopharyngeal carcinoma through repression of EZH2”.” Cancer Res 71: 225–233, 2011
[27] Cheng J. and Druzdzel M.J.”AIS-BN: An Adaptive Importance Sampling Algorithm for Evidential Reasoning in Large Bayesian Networks”, Journal of Artificial Intelligence Research. 13: 155-188, 2000