Supervised Method for Construction of microRNA-mRNA Networks: Application in Cardiac Tissue Aging Dataset

Georgios N. Dimitrakopoulos - *IEEE Student Member*, Konstantina Dimitrakopoulou, Ioannis A. Maraziotis, Kyriakos Sgarbas -*IEEE Member*, and Anastasios Bezerianos, *IEEE Senior Member*

Abstract— **MicroRNAs play an important role in regulation of gene expression, but still detection of their targets remains a challenge. In this work we present a supervised regulatory network inference method with aim to identify potential target genes (mRNAs) of microRNAs. Briefly, the proposed method exploiting mRNA and microRNA expression trains Random Forests on known interactions and subsequently it is able to predict novel ones. In parallel, we incorporate different available data sources, such as Gene Ontology and Protein-Protein Interactions, to deliver biologically consistent results. Application in both benchmark data and an experiment studying aging showed robust performance.**

I. Introduction

Over the last decade, microRNAs (miRNAs) have emerged as important and evolutionarily conserved regulators of various physiopathological processes, from development to cancer [1]. MiRNAs are small non coding RNAs, typically consisted of 21-25 nucleotides, that play important role in gene regulation and their role is to suppress gene expression by binding to mRNAs preventing them from being translated. A single miRNA can target hundreds of mRNAs, thus contributing significantly in gene expression regulation. Many algorithms have been proposed to detect transcripts targeted by miRNAs, usually relying on sequence analysis, however combining mRNA and miRNA expression data can reveal disease mechanisms or cellular processes[2]. Therefore, it is of great interest to unravel the miRNA-mRNA regulatory network governing a disease state or cellular process and explore the synergic or additive effects when multiple miRNAs target the same mRNA. In this way, miRNAs associated with a specific condition could be used as indicators or even as candidate therapeutic targets [3].

In Systems Biology, network notation has been widely used to model gene-gene interactions. In the recent literature several Gene Regulatory Network (GRN) inference algorithms have been developed based on various mathematical and computational methods, such as Machine Learning, Information Theory, Bayesian Networks and Ordinary or Stochastic Differential Equation models [4].

A main limitation of the majority of proposed GRN algorithms is that they rely solely on gene expression data. However, it has been shown that including other kinds of information leads to biologically more accurate results [5]. Of great interest is the Machine Learning category, under which are included algorithms based on Random Forest (GENIE3 [6]), Support Vector Machines (SVM) (SIRENE [7]) and Neural Networks (ENFRN [8]), because by construction these methods can embed a priori knowledge.

Random Forest (RF) [9] is an ensemble method, based on Classification and Regression Trees and has been successfully used for a wide variety of classification problems in Systems Biology; for example to determine a set of genes able to predict a disease [10] or to detect Single Nucleotide Polymorphisms (SNPs) related with certain diseases [11] and predict targets of miRNAs based on their sequence [12]. There are few cases that RF have been used for regression problems, mostly in an unsupervised way, with the scope to evaluate the association of variables to a condition; for example in [13] SNPs are ranked according to their association with Alzheimer disease and in [14] miRNAs are associated with glioblastoma based on fold change values of gene expression. With regard to gene regulatory network inference, only GENIE3 used regression with RF [6].

In the road for deciphering the miRNA-mRNA network, the proposed method uses regression RF with supervision, i.e. known miRNA-mRNA interactions in order to predict new potential targets. For this, we exploit microarray experiments measuring simultaneously under the same experimental settings miRNA and mRNA expression. First, a training phase is performed upon a priori knowledge and then test phase follows upon the remaining data, utilizing optionally heterogeneous biological data, such as Gene Ontology and Protein-Protein Interactions to ensure biologically relevant results. The proposed method was evaluated on benchmark datasets generated by the DREAM 5 network inference challenge [15] and compared to SIRENE algorithm in terms of accuracy. Finally, application of our method on a microarray experiment (recording both mRNA and miRNA expression profiles) investigating cardiac tissue aging mechanisms predicted miRNA-mRNA interactions, recently supported as cardiac age-related.

II. Methods

We developed a supervised method based on RF, which is able to detect potential target genes of miRNAs by

This research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Thalis. Investing in knowledge society through the European Social Fund.

G.N.D. and K.S. are with the Department of Electrical and Computer Engineering, University of Patras, Patras, 26500, GR (geodimitrak@upatras.gr , sgarbas@upatras.gr).

K.D. and I.A.M. are with the Medical School, University of Patras, Patras, 26500, GR (kondim@upatras.gr, imaraziotis@gmail.com)

A.B. is with the Medical School, University of Patras, Patras, 26500, GR and SINAPSE, National University of Singapore, 117456, Singapore (correspondence author: +302610969147, tassos.bezerianos@nus.edu.sg).

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

exploiting known miRNA-mRNA interactions. This method can optionally exploit different available biological data sources to increase outcome accuracy. Also, we present in detail Random Forests and two other algorithms of similar nature with our proposed method.

The cornerstone of our method is RF, which is an ensemble method that trains a large number of Classification and Regression Trees and aggregates their result by majority voting for classification and averaging for regression. Briefly, each tree is constructed on a random subset of the data and while growing, a variable to split is selected randomly. Thereby it is able to deliver accurate results and it does not suffer from overfitting to the training data. A RF-based GRN algorithm is Gene Network Inference with Ensemble of Trees (GENIE3). Considering each gene as target and all other genes as candidate regulators, it trains a Random Forest and subsequently uses the Variable Importance metric of the trained model to evaluate the rank of the potential regulators for each gene. This algorithm performs only training of RF based on the complete gene expression data without testing, hence, it operates in an unsupervised way. Supervised Inference of Regulatory Networks (SIRENE) is a supervised algorithm which predicts target genes of Transcription Factors (TF). In detail, SIRENE solves a classification problem using SVM and for each transcription factor it determines if genes are targets or not. It uses known relationships between TFs and targeted genes as positive examples, while in the absence of negative examples uses a cross-validation scheme on the unknown genes. In essence, to define a gene as a target, its profile should be similar to profile of other targets, while the regulator profile is not utilized.

We achieve embedding a priori knowledge by performing first a training phase based on this knowledge and then proceeding to test phase with the rest of the data. Specifically, for each known miRNA-gene interaction, we train a RF, using the gene profile as input and the miRNA profile as output. Intuitively, we train a RF to learn the function "a gene is regulated by a miRNA". Next, we use the trained model to test if other genes can be targets of the same miRNA. For each gene, we provide as input its profile and then we calculate the mean square error (MSE) between the predicted output and the miRNA profile. In case we obtain many error values for a candidate target, since a miRNA has usually more than one known targets, we keep the minimum error as final prediction score. In order to make comparable the error values derived from different RFs, miRNA expression values are normalized with zero mean and unit variance. An advantage of using regression instead of classification in comparison to SIRENE is that we overcome the absence of negative examples. Moreover, when a miRNA-mRNA interaction is given, the described scheme can take advantage of both miRNA and mRNA expression profiles.

Additionally, in order to reduce search space and at the same time deliver more biologically consistent results, we can exploit additional data sources, such as Gene Ontology (GO) biological process terms and Protein-Protein Interactions (PPIs). After training a model on a known miRNA-mRNA relationship, we restrict testing to genes belonging to the same GO biological processes as the gene under investigation. Moreover, PPI can be used complementary, so as to include in test set only genes that are in close proximity in the network topology. In our experiments, we limited our analysis up to second order neighbors.

```
Pseudocode of our method
Input: gene expression matrix $g$, miRNA expression matrix $m$, list of known interactions $L$, optional biological data $B$
Output: scores
Initialize scores = $+\infty$
for each interaction in $L$ between i-th miRNA and j-th gene
    $\mathrm{RF}_{ij}$ = train_RandomForest($g_j$, $m_i$)
    if B == $\varnothing$ , $\mathrm{test\_set}_j$ = get_all_genes()
    if B == GO, $\mathrm{test\_set}_j$ = get_genes_from_same_GO($B$, $g_j$)
    if B == PPI, $\mathrm{test\_set}_j$ = get_neighbors($B$, $g_j$)
    for each gene $g_k$ in $\mathrm{test\_set}_j$
        $\mathrm{prediction}_{ik}$ = test_ RandomForest ($\mathrm{RF}_{ij}$, $g_k$)
        error = MeanSquareError($m_i$, $\mathrm{prediction}_{ik}$)
        $\mathrm{scores}_{ik}$ = min(error, $\mathrm{scores}_{ik}$)
```

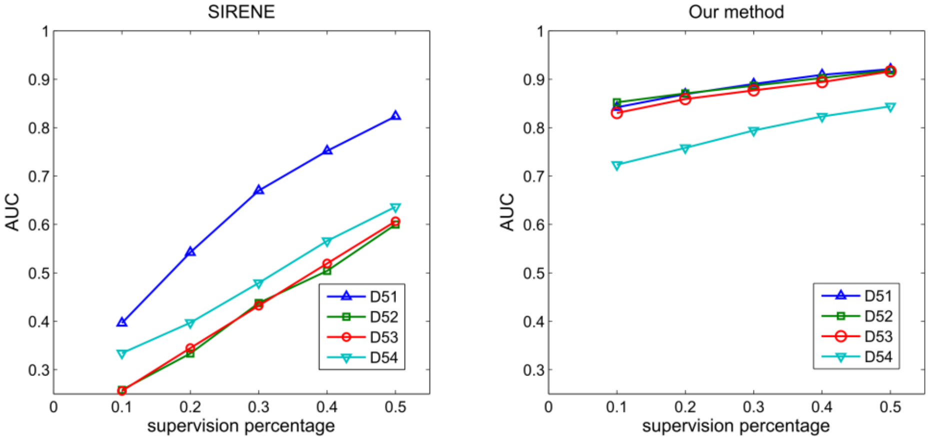

Figure 1. Performance of (A) SIRENE and (B) our method on the four DREAM 5 datasets for various suprvision percentages.

III. Results And Discussion

A. Data

To test the efficiency of our method, we used the benchmark datasets provided by DREAM 5 network inference challenge [15]. The DREAM project organizes annual challenges for Systems Biology problems, such as network inference, providing gene expression datasets along with the real network topology derived from validated biological data. We used the 4 datasets with averaged experimental conditions, containing 1,643 genes - 487 samples (*in silico*), 2,677 genes - 53 samples (S. *aureus*), 4,511 genes - 487 samples (*E. coli*) and 5,950 genes - 321 samples (*S. cerevisiae*), which hereafter will be referred as D51, D52, D53 and D54 respectively.

Next, we applied our method in a microarray dataset studying cardiac aging accessible with GSE43556 series number in NCBI's Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) [16]. In this dataset the expression of 22,406 genes and 566 microRNAs was measured across 8 mice of different age. Data were log transformed and Z-transformed. We tested the hypothesis that the expression of a given gene is associated with age. For each gene we performed linear regression, using:

$$Y_{ij} = \beta_{0j} + \beta_{1j} Age_{i} + \varepsilon_{ij}$$

where Y_{ij} is the signal intensity of gene *j* in sample *i*, Age_i is the age of the specimen from which sample *i* was obtained, and ε_{ij} is an error term. Coefficients β_0 and β_1 were estimated by least squares. A two-tailed F-test was performed on the differential expression to estimate statistical significance of the slope of the curve, which would indicate an association between expression and age. Considering profiles with $|slope| > 0.005$ and $P\text{-value} < 0.05$ as putatively age dependent resulted in a set of 155 miRNAs and 2,995 genes.

In parallel, we extracted the GO biological process terms related to each gene from the corresponding platform file GPL1261. PPIs were collected from iRefIndex (http://irefindex.org/) and MiMI (http://mimi.ncibi.org/) databases, resulting in 28,037 interactions among 4,399 proteins. The set of miRNA-mRNA interactions was compiled from Tarbase (http://www.microrna.gr/tarbase) (experimentally verified interactions) and miRecords (http://mirecords.biolead.org/) (predicted interactions), including in total 73,932 relations among 410 microRNAs and 10,151 genes. With respect to miRecords, we included only relations supported by at least four miRNA target prediction tools.

TABLE I. Performance on Dream 5 Datasets

Algorithm	Dataset			
	D51	*D52*	*D53*	*D54*
GENIE3	**0.815**	0.622	0.617	0.518
Pearson Correlation	0.609	**0.631**	0.580	0.517
Anova-based	0.780	0.608	**0.671**	0.519
Meta-predictor	0.695	0.538	0.602	**0.540**
Our Method (10%)	0.842	0.852	0.830	0.724
Our Method (50%)	0.921	0.919	0.916	0.844

In bold the maximum performance of the unsupervised methods per dataset in DREAM 5 contest is highlighted.

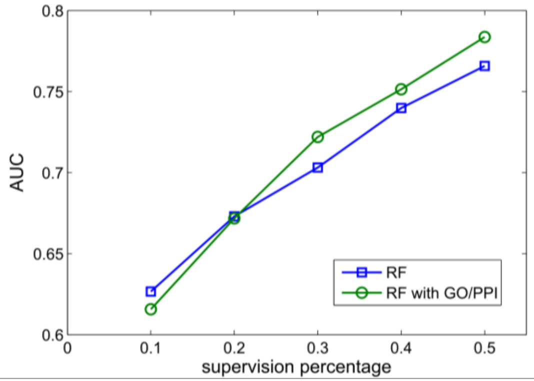

Figure 2. Performance of our method on the aging dataset was improved when additional biological data sources were included.

B. Results

In our experiments, to construct the training set for each regulator, we kept randomly a percentage (10%-50%) of the known interactions, so remaining interactions would be used for evaluation. Reported results are mean values over 100 repetitions. In all cases, the number of trees in RF was set to 100. To evaluate the accuracy of the inferred networks, Area Under Curve (AUC) was used, which is computed as the area under the curve of the True Positive Rate versus the False Positive Rate at various values of threshold, which overcomes the need to search for an optimal threshold.

In Table 1, the top performing methods in DREAM 5 contest are shown [15]. It is important to note that with the exception of D51 dataset, which is artificial and is the smallest of the four, AUC scores were very low and in the case of D54, marginally better than random prediction (0.50). This indicates that based only on gene expression, detecting most of the real interactions is very hard task for any method.

Initially, we applied SIRENE (Fig. 1A) and our method (Fig. 1B) on DREAM datasets, providing as input various percentages of the real interactions. Our method showed superior performance against all unsupervised methods and SIRENE in all cases. Additionally, it is evident that it provided robust results and outperformed the unsupervised methods even when a small fraction of real relationships was available. Also, for SIRENE, a very large percentage (50%) of the known relationships was required as input to provide acceptable performance, similar to unsupervised methods.

Next, we applied our method in the cardiac aging dataset, with the scope to detect the age-related miRNA-mRNA interactions (Fig. 2). We observed that our method was able to provide robust results, with AUC significantly over 0.5, which was increased proportionally with supervision. Moving forward, we utilized GO and PPI data during the test phase. Despite testing was restricted to about 1/3 of all possible targets, results remained stable or improved. This is explained by the notion that genes regulated by the same miRNA are functionally associated and using this a priori knowledge we discarded successfully unrelated genes. Finally, in order to detect potential novel interactions, we used as input to our method the highly confident miRNA-mRNA relations. In Table 2, example miRNA-mRNA predicted relations are provided, all of which were not listed in our initial interaction pool but are supported by at least one target prediction tool (Diana - http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=microtv4, Miranda - www.mirbase.org, Targetscan - www.targetscan.org). In many cases, the reported KEGG pathways are associated with cardiac aging or cardiovascular diseases according to literature-based evidence.

IV. Conclusion

Our proposed a method showed solid performance in benchmark datasets and was proven able to take advantage of previous biological knowledge to deliver accurate predictions. In future work, additional features can be used for detecting association between microRNAs and genes, such as sequence similarity. Currently there are relatively few experiments measuring both mRNA and miRNA expression, but we expect their number to rise in the future, which will increase the demand for such methods.

TABLE II. Aging Related Predicted Targets

MicroRNA	Gene	KEGG Pathways
mmu-let-7a	Fam126b	-
	Gnb1	mmu04014-Ras signaling pathway [17]
mmu-mir-106a	Atp1b2	mmu04261-Adrenergic signaling in cardiomyocytes [19]
	Prkar2a	mmu04210-Apoptosis [18]
	Rnf207	-
	Tdrkh	-
mmu-mir-298	Cul3	mmu04120-Ubiquitin mediated proteolysis [20]
	Glyr1	-
	Grn16515	-
mmu-mir-351	Cep	mmu04020-Calcium signaling pathway [21]
	Cgn	mmu04530-Tight junction
	A830080D01Rik	-
mmu-mir-494	Wwc2	-
	Dhx35	-
	Acsl4	mmu03320-PPAR signaling pathway [22]
	1810013224Rik	-
mmu-mir-503	Ak4	mmu04014-Ras signaling pathway [17]
	Cnot6l	mmu03018-RNA degradation
	Etnk1	mmu01100-Metabolic pathways
	Bcl11a	-
	Stat5b	mmu04630-Jak-STAT signaling pathway [23]
	Hif1a	mmu04150-mTOR signaling pathway [24]
	Srsf1	mmu03040-Spliceosome

Indicative examples of age-related miRNAs and their predicted targets. The reported references provide evidence that the respective miRNA-mRNA pairs are associated with cardiac aging or cardiovascular diseases. No KEGG Pathway annotation was available for genes marked with '-'.

References

[1] M. H. Schulz, K. V. Pandit, C. L. Lino Cardenas, N. Ambalavanan, N. Kaminski, and Z. Bar-Joseph, “Reconstructing dynamic microRNA-regulated interaction networks”, *Proceedings of the National Academy of Sciences*, 110:39, pp. 15686-15691, 2013.

[2] C. L. Plaisier, M. Pan, and N. S. Baliga, “A miRNA-regulatory network explains how dysregulated miRNAs perturb oncogenic processes across diverse cancers”, *Genome Res.*, 22, 2302-14, 2012.

[3] J. Fu, W. Tang, P. Du, G. Wang, W. Chen, J. Li, Y. Zhu, J. Gao, and Long Cui, “Identifying MicroRNA-mRNA regulatory network in colorectal cancer by a combination of expression profile and bioinformatics analysis”, *BMC Systems Biology*, 6:68, 2012.

[4] M. Hecker, S. Lambeck, S. Toepfer, E. van Someren, and R. Guthke, “Gene regulatory network inference: data integration in dynamic models-a review”, *Biosystems,* 96:1, pp. 86-103, Apr 2009.

[5] P. B. Madhamshettiwar, S. R Maetschke, M. J. Davis, A. Reverter, and M. A. Raga, “Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets”, *Genome Medicine*, 4:41, 2012.

[6] V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, P. Geurts, “Inferring Regulatory Networks from Expression Data Using Tree-Based Methods”, *PLoS ONE*, 5(9): e12776, 2010.

[7] F. Mordelet, and J.- P. Vert, “SIRENE: supervised inference of regulatory networks”, *Bioinformatics*, 24(16):i76-i82, 2008.

[8] I. A. Maraziotis, A. Dragomir, and D. Thanos, “Gene Regulatory networks modeling using a dynamic evolutionary hybrid”, *BMC Bioinformatics*, 11:140, Mar. 2010.

[9] L. Breiman, “Random forests”, *Machine Learning,* 45:1, 5–32, 2001.

[10] R. Díaz-Uriarte, S. A. De Andres, “Gene selection and classification of microarray data using random forest”, *BMC Bioinformatics*, 7:3, 2006.

[11] H. J. Cordell , “Detecting gene–gene interactions that underlie human diseases”, *Nature Reviews Genetics*, 10:6, pp. 392-404, 2009.

[12] M. R. Mendoza, G. C. da Fonseca, G. Loss-Morais, R. Alves, R. Margis, and A. L. C. Bazzan, “RFMirTarget: Predicting Human MicroRNA Target Genes with a Random Forest Classifier”, *PLoS ONE*, 8(7): e70153, 2013.

[13] Y. Wang, W. Goh, L. Wong, G. Montana and the Alzheimer's Disease Neuroimaging Initiative, “Random forests on Hadoop for genome-wide association studies of multivariate neuroimaging phenotypes”, *BMC Bioinformatics*, 14(Suppl 16):S6, 2013.

[14] S. Wuchty, D. Arjona, A. Li, Y. Kotliarov, J. Walling, et al. “Prediction of Associations between microRNAs and Gene Expression in Glioma Biology”, *PLoS ONE*, 6(2): e14681, 2011.

[15] D. Marbach, J. C. Costello, R. Küffner, N. M. Vega, R. J. Prill, D. M. Camacho, and K. R. Allison, “Wisdom of crowds for robust gene network inference”. *Nature Methods*, 9:8, pp. 796-804, 2012.

[16] A. R. Boon, K. Iekushi, S. Lechner, T. Seeger, et al., “MicroRNA-34a regulates cardiac ageing and function”, *Nature,* 495:7439, pp. 107-10, Mar 2013.

[17] A. T. Naito, I. Shiojima, and I. Komuro, “Wnt signaling and aging-related heart disorders”, *Circ Res*, 107:11, pp. 1295-303, Nov 2010

[18] M. Pollack, S. Phaneuf, A. Dirks, and C. Leeuwenburgh, “The role of apoptosis in the normal aging brain, skeletal muscle, and heart”, *Ann N Y Acad Sci*, 959, pp.93-107, Apr 2002.

[19] L. Barki-Harrington, C. Perrino, and H. A. Rockman, “Network integration of the adrenergic system in cardiac hypertrophy”, *Cardiovasc Res*, 63:3, pp. 391-402, Aug 2004.

[20] A. L. Portbury, M. S. Willis, and C. Patterson, “Tearin' up my heart: proteolysis in the cardiac sarcomere”, *J Biol Chem*, 286:12, pp. 9929-34, 25 Mar 2011

[21] Q. Lou, A. Janardhan, and I. R. Efimov, “Remodeling of calcium handling in human heart failure”, *Adv Exp Med Biol*, 740, pp. 1145-74, 2012.

[22] B. N. Finck, “The PPAR regulatory system in cardiac physiology and disease”, *Cardiovasc Res*, 73:2, pp. 269-77, Jan 15 2007.

[23] Y. T. Xuan, Y. Guo, H. Han, Y. Zhu, and R. Bolli, “An essential role of the JAK-STAT pathway in ischemic preconditioning”, Proc Nat Acad Sci, 98:16, pp. 9050-5, Jul 31 2001.

[24] B. J. North, and D. A, Sinclair, “The intersection between aging and cardiovascular disease”, *Circ Res*, 110:8, pp. 1097-108, Apr 13 2012.