Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/bbe

Original Research Article

# MicroRNA expression prediction: Regression from regulatory elements

*Hasan Oğul* *, *M. Emre Tuncer*

*Department of Computer Engineering, Başkent University, Ankara, Turkey*

ARTICLE INFO

*Article history:*
Received 7 January 2015
Received in revised form
7 April 2015
Accepted 25 October 2015
Available online 7 November 2015

*Keywords:*
MicroRNA
Microarray
Gene expression

ABSTRACT

MicroRNAs are known as important actors in post-transcriptional regulation and relevant biological processes. Their expression levels do not only provide information about their own activities but also implicitly explain the behaviors of their targets, thus, in turn, the circuitry of underlying gene regulatory network. In this study, we consider the problem of estimating the expression of a newly discovered microRNA with known promoter sequence in a certain condition where the expression values of some known microRNAs are available. To this end, we offer a regression model to be learnt from the expression levels of other microRNAs obtained through a microarray experiment. To our knowledge, this is the first study that evaluates the predictability of microRNA expression from the regulatory elements found in its promoter sequence. The results obtained through the experiments on real microarray data justify the applicability of the framework in practice.



## 1. Introduction

In the last decade, there has been a drastic change in understanding of gene regulation due the discovery of tiny non-recoding regulatory entities; called microRNAs. Many evidences suggest that they are effective in various cellular processes such as development, aging and apoptosis [1–3] by playing role as regulatory factors in post-transcriptional level. It is already proven that various diseases are associated with the abnormal behaviors of specific microRNAs [4–6]. As they are mediating the translational behavior of mRNAs, they are also regulated by other transcription factors during transcription, which in turn determines their expression levels. Therefore, whole regulatory process contains many cycles at several stages. Elucidating the activity of a miRNA thus requires the quantification of its expression which will pivot the translational repression.

As in other RNA expression measurements, a common way for miRNA expression profiling is to use microarray technology [7]. It can allow the quantification of the expression levels of several miRNAs simultaneously. The number of public miRNA expression datasets has been steadily increased in last few years. On the other hand, the number of known miRNAs has also been increasing in a comparable speed. Therefore, there is a remarkable difference between the numbers of profiled miRNAs in experiments done in increasing years. While the researchers can significantly

* *Corresponding author at*: Department of Computer Engineering, Başkent University, Ankara, Turkey.
E-mail address: hogul@baskent.edu.tr (H. Oğul).

http://dx.doi.org/10.1016/j.bbe.2015.10.010

benefit from previous microarray datasets, the lack of the expression levels of the miRNAs that are discovered later than the date of the experiment done makes it uncomfortable in inferring new hypotheses or designing novel experiments. This problem may also arise when any of the measurement is missing due to the potential signal noise or experimental errors [8].

Knowing the fact that the replication of a whole profiling experiment by a microarray having the probes for all currently known miRNAs is not feasible pertaining to the cost or environmental constrains, we suggest here the use of an *in silico* means of quantifying the expression levels of novel (or missed) miRNAs. To this end, we introduce a computational method that relies on the prediction of the exact value of the expression of a miRNA from the expression levels of known miRNAs that were measured over a previous microarray experiment. In this study, a regression formulation is done to model the expression through the regulatory elements observed along the promoter of input miRNA. Similar attempts have been done before to predict mRNA expression from the information of transcription factor binding sites. A pioneering work by Beer and Tavazoie used Bayesian networks to identify whether a given mRNA is expressed or not [9]. This work was then re-examined by a naive Bayes approach and shown that this prediction can be realized with a less complicated model [10]. Our study has two important contributions. First, we consider the problem as prediction of the exact value of expression rather than binary classification of high or low expression. Second, the problem is addressed in the context of miRNA experiments instead of mRNA microarrays. The results that we obtain using Relevance Vector Machines as regression model encourages the use of the *in silico* miRNA expression profiling concept as a complementary tool in gene regulation analysis.

## 2. Methods

### 2.1. *General framework*

The framework that we offer to predict microRNA expression is shown in Fig. 1. The input arguments to the system involve the promoter sequence of microRNA in question and the binding sites of all transcription factors in specified organism. A statistical model, which is defined as a regression model in our case, is learnt from measured expression values and the regulatory elements of other known microRNAs. The regulatory elements of a microRNA are found by sequential search of TF binding sites on its promoter. Each element is characterized by a binary value that represents the condition of binding or non-binding of corresponding TF, which in turn built a composition of regulatory elements.

Considering its recent success in several contexts, including gene regulation studies [11,12], we adopted Relevance Vector Machine (RVM) as a regression model in our framework. On the other hand, since this is the first attempt for prediction of miRNA expression from sequence information, there is no previous method in the literature that can be compared against our approach in terms of the aptness of prediction. Therefore, we also used two simple regression methods; k-Nearest Neighbors and linear regression, to make baseline predictors.

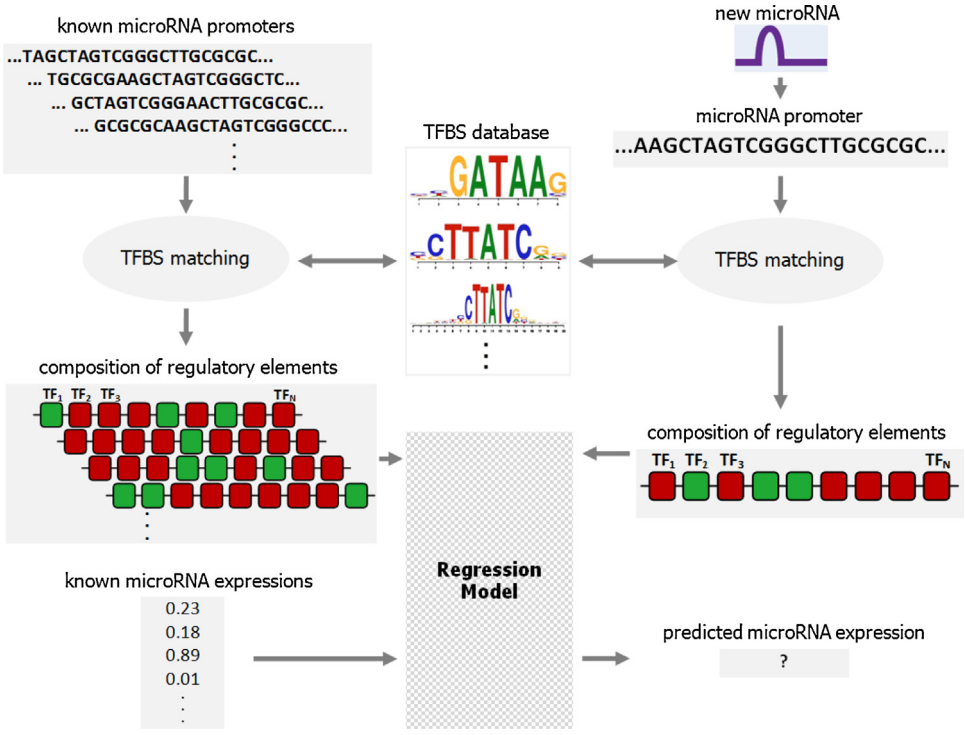


**Fig. 1 – The general framework for predicting microRNA expression (the regulatory elements are shown by green or red, which specifies the existence or non-existence of TF binding respectively. TFBS stands for transcription factor binding site). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)**