Methods 69 (2014) 220–229

Contents lists available at ScienceDirect

Methods

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

Ensemble learning can significantly improve human microRNA target prediction

Seunghak Yu [a,b], Juho Kim [a], Hyeyoung Min [c], Sungroh Yoon [a,d,*]

[a] Department of Electrical and Computer Engineering, Seoul National University, Seoul 151-744, Republic of Korea
[b] Department of IT Convergence, Korea University, Seoul 156-713, Republic of Korea
[c] RNA Biopharmacy Laboratory, College of Pharmacy, Chung-Ang University, Seoul 156-756, Republic of Korea
[d] Bioinformatics Institute, Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 151-747, Republic of Korea

ARTICLE INFO

Article history:
Received 19 March 2014
Revised 16 July 2014
Accepted 18 July 2014
Available online 1 August 2014

Keywords:
MicroRNA
Target prediction
Sequence analysis
Algorithm
Machine learning

ABSTRACT

MicroRNAs (miRNAs) regulate the function of their target genes by down-regulating gene expression, participating in various biological processes. Since the discovery of the first miRNA, computational tools have been essential to predict targets of given miRNAs that can be biologically verified. The precise mechanism underlying miRNA–mRNA interaction has not yet been elucidated completely, and it is still difficult to predict miRNA targets computationally in a robust fashion, despite the large number of *in silico* prediction methodologies in existence. Because of this limitation, different target prediction tools often report different and (occasionally conflicting) sets of targets. Therefore, we propose a novel target prediction methodology called *stacking-based miRNA interaction learner ensemble* (SMILE) that employs the concept of stacked generalization (stacking), which is a type of ensemble learning that integrates the outcomes of individual prediction tools with the aim of surpassing the performance of the individual tools. We tested the proposed SMILE method on human miRNA–mRNA interaction data derived from public databases. In our experiments, SMILE improved the accuracy of the target prediction significantly in terms of the area under the receiver operating characteristic curve. Any new target prediction tool can easily be incorporated into the proposed methodology as a component learner, and we anticipate that SMILE will provide a flexible and effective framework for elucidating *in vivo* miRNA–mRNA interaction.

1. Introduction

MicroRNAs (miRNAs) are small non-protein coding RNAs that regulate gene function by down-regulating the expression of their target gene(s) [1,2]. As central players in post-transcriptional gene regulation, miRNAs are known to be involved in many biological processes and diseases [3,4], and miRNA research continues to rapidly increase. MicroRNAs exert their function through binding to target sites present in the 3′ untranslated region (UTR) of their cognate mRNAs. For a complete understanding of miRNA function, it is thus critical to identify the target mRNAs which a miRNA binds to and functions through.

However, the targets of most miRNAs remain unknown because of the lack of robust bioinformatics methods that predict potential mRNA targets precisely and the lack of non-laborious experimental verification systems. Since the base-pairings between miRNAs and their target mRNAs include bulges and non-canonical base pairs and are thus not perfect, it is difficult to predict miRNA targets with high accuracy. Efforts have been made to identify the molecular mechanism underlying miRNA–mRNA interaction, but the exact mechanism by which miRNAs select their targets and mediate translational repression has not yet been completely elucidated [5].

Nevertheless, studies have suggested that sequence complementarity, target site accessibility, and evolutionary conservation are important factors for target recognition [6]. The first-generation of miRNA target prediction tools focused on utilizing these factors to predict novel miRNA–target pairs. Examples include TargetScan [1], miRanda [7], RNAhybrid [8], DIANA-microT [9], PicTar [10], and PITA [11]. These first-generation tools significantly contributed to the field by reporting a large number of putative targets, some of which were later confirmed by experimental studies, thus expanding the library of known the miRNA–target interactions. However, these tools mostly consider only subsets of known factors that affect the binding of miRNAs to their targets, and they

* Corresponding author at: Department of Electrical and Computer Engineering, Seoul National University, Seoul 151-744, Republic of Korea.
E-mail address: sryoon@snu.ac.kr (S. Yoon).

http://dx.doi.org/10.1016/j.ymeth.2014.07.008
1046-2023/

often show unsatisfactory performance in terms of the rates of false positives and false negatives.

As the number of experimentally verified miRNA targets increased, second-generation tools emerged that commonly assumed the existence of 'training' data. These tools are based on machine-learning approaches that employ classifiers computationally trained using experimentally verified miRNA–target pairs. Examples include miRTarget2 [12], TargetBoost [13], TargetSpy [14], and TargetMiner [15]. These methods learn (occasionally subtle but critical) features appearing in real miRNA–target pairs and non-pairs and apply them to the prediction of unknown pairs, thus boosting the accuracy of prediction compared to the first-generation techniques. Recently, hybrid approaches that consider structural features and machine-learning features together have been proposed, *e.g.*, miREE [16], NBmiRTar [17], and DIANA-microT-ANN [18]. The performance of the second-generation methods relies on the quantity and quality of training data and the learning algorithm used, and these methods often exhibit limited abilities to reduce generalization error in the prediction of unknown interactions.

Various prediction tools have been proposed in the first and second generation, and the third-generation prediction methods collect the outcomes from different tools and combine them to obtain better results than individual tools can deliver. Some tools also integrate miRNA and mRNA expression data for target prediction, given miRNA expression and mRNA expression are inversely correlated. Examples of the third-generation methods include MiRror [19], MMIA [20], MAGIA2 [21], imiRTP [22], miRecords [23], miRGator ver2.0 [24], miRNAMap 2.0 [25], and miRGen [26]. Most of these tools do not account for the strengths and weaknesses of the unique features of each algorithm; rather, they merely incorporate the presence or absence of detection and target scores independently. Since each prediction method intrinsically possesses numerous false positives, the simple integration of multiple methods may result in increased false positives, making accurate target prediction more difficult.

To address the limitations of the existing approaches, here we propose a novel miRNA target prediction algorithm called *stacking-based microRNA interaction learner ensemble* (SMILE). We introduce this approach because existing prediction tools often produce inconsistent results even when given identical inputs [27–29] and the resulting diverse outputs from multiple prediction models can reduce the prediction error (see Eq. (1) in Section 2). The proposed SMILE method can be classified as a third-generation approach and works in two stages: each prediction tool is considered as an individual prediction model (1st stage), and the multiple models are then combined through ensemble learning (2nd stage). There exist multiple approaches for ensemble learning as described in Section 2, and our method is based on the idea of stacked generalization (stacking) [30,31], which summarizes the outcomes of individual learners for each object as a feature vector and classifies the feature vectors using a second-level learner. Using this idea, we can implicitly consider the features related to miRNA–mRNA interactions that individual tools miss, thus achieving improved prediction performance.

2. Background

2.1. Key idea behind ensemble learning

Ensemble learning consists of a set of prediction models and a method to combine these models. In ensemble learning, we train multiple models and combine the outputs from individual models to reduce generalization error. Depending on how the diversity of individual models is addressed, ensemble learning can be categorized into implicit and explicit methods [32]. An implicit method utilizes random subsets of training data when creating each individual model. Examples include *bagging*, which selects samples randomly and *random forests*, which selects samples and features randomly [33]. An explicit method obtains different individual models using a measurement. An example is *boosting*, which utilizes the error from one stage to the next, thus gradually increasing accuracy.

To reduce generalization error effectively, the diversity of the prediction models used in the first stage is important. Specifically, the outputs from individual prediction models had better be negatively correlated [34]. For example, in the case of a linear combiner, the classification error of a simple averaging ensemble denoted by E_{ave} is given in [34] as

$$E_{ave} = E_{add} \times \frac{1 + \delta \times (T-1)}{T} \tag{1}$$

where E_{add} is the added classification error of individual models, T is the number of prediction models used, and δ is the correlation coefficient among the outputs from individual models. If the individual models are equal ($\delta = 1$), then the ensemble error becomes identical to the individual error (*i.e.*, $E_{ave} = E_{add}$). If all of the models are uncorrelated ($\delta = 0$), the ensemble error becomes the average individual error (*i.e.*, $E_{ave} = E_{add}/T$). If the models are negatively correlated ($\delta < 0$), then the ensemble error becomes smaller than the average individual error (*i.e.*, $E_{ave} < E_{add}/T$).

2.2. Performance evaluation metrics

In this work, we formulate the problem of predicting miRNA–mRNA pairs as an instance of a binary classification task. To facilitate later discussion, we review the performance statistics widely used to evaluate different methodologies.

In the binary classification setting, evaluation metrics are based on the notion of true and false positives (TP and FP), and true and false negatives (TN and FN). TP (TN) refers to a positive (negative) instance that is correctly classified as positive (negative). FP (FN) is a negative (positive) instance that is incorrectly classified as positive (negative).

Based on these classifications, widely used performance metrics are defined [35]:

$$\text{sensitivity} = \text{true positive rate} = \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{specificity} = \text{true negative rate} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3}$$

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4}$$

$$\text{positive predictive value(PPV)} = \text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

$$\text{negative predictive value(NPV)} = \frac{\text{TN}}{\text{TN} + \text{FN}} \tag{6}$$

$$\text{F-measure} = 2\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{7}$$

2.3. Overview of the proposed approach

Fig. 1 shows the overview of our approach, which largely consists of three steps. First, we select six existing miRNA target prediction tools and let them produce the prediction results individually. Second, we preprocess the outcomes from the individual tools and then generate positive and negative training samples using a public database of experimentally verified miRNA–target pairs and various statistics obtained from the first stage results. Lastly, we train a binary classifier by using the