A Probabilistic Framework to Improve microRNA Target Prediction by Incorporating Proteomics Data

Jingjing Li*

Department of Molecular Genetics, Donnelly Centre for Cellular and Biomolecular Research
University of Toronto, Toronto, Ontario, Canada

jj.li@utoronto.ca

Renqiang Min*

Department of Computer Science, Donnelly Centre for Cellular and Biomolecular Research
University of Toronto, Toronto, Ontario, Canada

minrq@cs.toronto.edu

Anthony Bonner

Department of Computer Science
University of Toronto, Toronto, Ontario, Canada

bonner@cs.toronto.edu

Zhaolei Zhang†

Department of Molecular Genetics, Donnelly Centre for Cellular and Biomolecular Research
Banting and Best Department of Medical Research, University of Toronto,
Toronto, Ontario, Canada

Zhaolei.Zhang@utoronto.ca

Received 16 July 2009
Accepted 11 August 2009

Due to the difficulties in identifying microRNA (miRNA) targets experimentally in a high-throughput manner, several computational approaches have been proposed. To this date, most leading algorithms are based on sequence information alone. However, there has been limited overlap between these predictions, implying high false-positive rates, which underlines the limitation of sequence-based approaches. Considering the repressive nature of miRNAs at the mRNA translational level, here we describe a probabilistic model to make predictions by combining sequence complementarity, miRNA expression level, and protein abundance. Our underlying assumption is that, given sequence complementarity between a miRNA and its putative mRNA targets, the miRNA expression level should be high and the protein abundance of the mRNA should be low. Having identified a set of confident predictions, we then built a second probabilistic model to trace back to the mRNA expression of the confident targets to investigate the mechanisms of the miRNA-mediated post-transcriptional regulation. Our results suggest that

*These authors contributed equally to this work.
†To whom correspondence should be addressed.

translational repression (which has no effect on mRNA level), instead of mRNA degradation, is the dominant mechanism in miRNA regulation. This observation explained the previously observed discordant correlation between mRNA expression and protein abundance.

Keywords: Gene Regulation; MicroRNA Target Prediction; Proteomics Data Analysis.

1. Introduction

MicroRNAs (miRNAs) are a class of small non-coding RNAs, typically about 22 nucleotides in length, and are known to block protein synthesis of their target genes by binding to the 3UTR of the mRNA transcripts with perfect (in plants) or imperfect (in animals and c. elegans) base pairing [4]. It was estimated that thousands of genes in the mammalian genome are under regulation by miRNA at the post-transcriptional level [13], and they have been shown to have many important functional roles [2].

Despite microRNAs importance and prevalence, it has proved to be difficult to experimentally identify and validate their target genes. To this date, only 40 miRNA targets have been confirmed in mouse and 200 in human [18]. As an alternative, a number of computational prediction programs had been developed and were widely used to predict miRNA targets in silico (see [5], [11], [13], and [15]). Most of these computational programs combine two types of data in making predictions: sequence complementarity between the miRNA and the putative target binding sites, and the evolutionary conservation of such sites (for a review, see [18]). Although great progress has been made in improving prediction accuracy, accurate prediction of miRNA targets remains challenging, the major difficulty being the lack of agreement among these algorithms. A recent benchmark study has compared the predicted targets of several leading algorithms and reported significant discrepancy among them [18]. The disagreement among these algorithms can be largely attributed to the different scoring schemes and weights given to imperfect base pairing between miRNA and binding sites and evolutionary conservation of the binding sites. Moreover, some of these sequence-based algorithms are known to be less robust as slight changes in parameters often result in very different predictions [26].

Because of the repressive nature of miRNAs regulatory roles and the availability of the genome-wide mRNA expression data, it was suggested that using gene expressions data could be helpful in predicting true miRNA targets [16]. The rationale of such approach is the following: if a miRNA is highly expressed and a putative target gene is lowly expressed in a particular tissue type, then it is considered as an additional evidence that the candidate gene is a true target. Although this algorithm produced encouraging results [10] and such a negative correlation between the expression levels of miRNA and target mRNA has been observed in various experiments (see [7], [16], and [19]), special cautions should be taken in this approach for the following reasons. (1) The reported instances of negative correlation between miRNA and mRNA expression levels are limited to a very small fraction of the known miRNAs with anecdotal evidence; it is uncertain whether such negative correlation can be

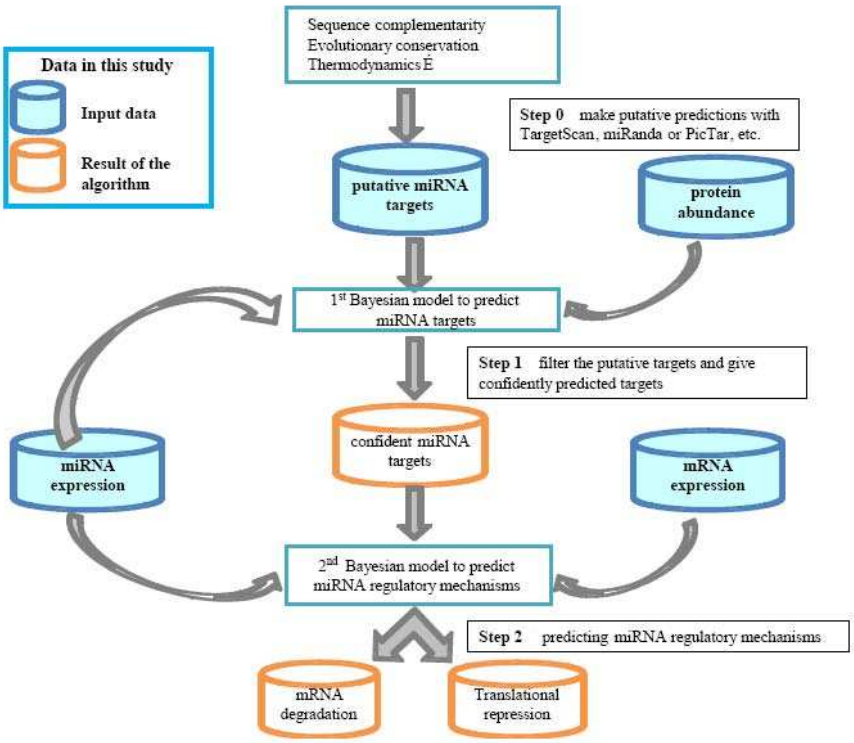

Fig. 1. A flowchart of the algorithms described in this work. The described algorithm takes four types of experimental data: (1) a set of putative miRNA targets, (2) protein abundance, (3) miRNA expression profiles, and (4) mRNA expression profiles. Details can be found in the main text.

extrapolated to other types of miRNAs and their targets. (2) It is known that, in contrast to the miRNAs in plants, the miRNAs in animals typically have imperfect sequence complementarity to their target sites and function mostly by binding to the target sites to inhibit the translation process, instead of causing degradation of the mRNA transcripts ([6], [24]). (3) In some cases, a strong positive correlation has also been observed between the expression levels of miRNA and their target mRNAs [23], which could be attributed to common regulators shared by the miRNA and their target genes. All the evidence described above suggest that in animals, the repression effect of miRNAs on their target genes is more obviously manifested at the translational level (i.e. protein abundance), thus, identifying miRNA targets solely based on transcriptional data might be less effective. In contrast, regardless of by degradation or by translational repression, the protein abundance of the target genes should be always negatively regulated. Motivated by this observation and by previous work, in this paper we propose a new Bayesian approach to predicting

miRNA targets in mouse using proteomic data in a high-throughput manner. In [22], researchers have used miRNA expression data and proteomics data to identify miRNA targets. However, their predictions are mainly based on biological experiments and no computational algorithm was proposed, and sequence-based methods such as TargetScan, PicTar and miRanda were only used to support their predictions. Besides, their experimental predictions were only performed in one tissue of rat, kidney. To our knowledge, our approach is the first computational one that incorporates miRNA expression data and proteomics data in multiple tissues to carry out high-throughput miRNA target predictions.

Our method consists of two steps. In the first step, it takes as input a set of putative miRNA target genes derived from sequence information alone; it then applies a probabilistic model using protein abundance data to assign confidence scores to the predicted miRNA-target pairs. In the second step, another probabilistic model is applied to the miRNA and mRNA expression data to predict whether the miRNA-mediated regulation is through translation repression or mRNA degradation. Figure 1 shows a flowchart of our approach.

2. Data Gathering

The mouse protein abundance data was derived from a recently published mass spectrometry study [12], in which the abundance of 4,768 proteins across 6 mouse tissues (brain, heart, liver, lung, placenta and kidney) was surveyed. After comparing with gene expression data from two microarray studies ([21], [27]), 1,758 proteins were confidently cross-mapped to their corresponding mRNAs. The incomplete coverage of the proteomic data was likely due to instrumentation bias, stringent filtering rules of database search or instability of low-level transcripts. The miRNA expression data was extracted from previous published microarray studies [3]. These authors also used TargetScan and miRanda separately to derive two lists of putative miRNA targets in mouse. The normalized mRNA expression profiling of 41,699 transcripts in 55 tissues was from [27]. Among these, 1,758 were confidently cross-mapped to the proteins (see above).

We chose to use full Bayesian formulism to make inference so as to take into account all possible uncertainties in our model. Inferences were made based on Gibbs sampling [8], which was performed in the WinBugs environment [20].

3. Method and Results

3.1. *Deriving a list of putative miRNA targets by sequence data alone*

We described the sources of the data in the above Methods section. As described in the Introduction, our current methods take as input a set of putative predictions from a sequence-based prediction algorithm. We decided to run our procedures twice using two different prediction algorithms: TargetScan [14] and miRanda [5]. The

general results and conclusions are unchanged. Based on the intersection among the four types of datasets (predicted miRNA targets, miRNA expression, protein abundance and mRNA expression), we retained 21,721 putative interactions for TargetScan predictions (75 miRNAs, 1,404 cross-mapped mRNAs) and 17,339 putative interactions for miRanda predictions (70 miRNAs, 1408 cross-mapped mRNAs).

After compiling the datasets, we investigated mRNA or protein expression profiles in 6 tissues (brain, heart, liver, lung, placenta and kidney), among which expression in 4 tissues (brain, heart, liver and lung) were used for model construction and making predictions, while the remaining 2 tissue types (placenta and kidney) were used for blind tests.

3.2. *Modeling protein abundance*

Instead of looking for negative correlation between miRNAs and mRNA transcripts as previously described in [10], our method relies on the correlation between miRNA and proteins (see Figure 1). We start with a set of miRNA and target genes as predicted from a sequence-based approach, we then model the protein abundance of the putative targets and the miRNA in individual tissues. If we observe a negative correlation between the miRNA and the putative target in one particular tissue, then the algorithm will assign a higher confidence score to this miRNA-protein pair. Conversely a positive correlation between miRNA expression and protein abundance, especially the cases where a high miRNA expression coincide with high protein abundance, will result in a low confidence score for the miRNA-protein pair.

We chose to use a probabilistic framework to model the rela-tionship between miRNA and the protein abundances. The first challenge in this approach is to find an appropriate background distribution to model the protein abundance data. Unlike mRNA expression profiles, which can be effectively modeled using a Gaussian distribution, the peptide counts are discrete values. A possible choice is to use Poisson to model the count events; however, a simple Poisson model is not suitable for modeling the peptide counts in this study since there are excessive zeros in the dataset and the non-zero counts are also over-dispersed (variance are much greater than the mean). After comparing with other possible models such as zero-inflated Poisson and transformed Gaussian, we chose to use negative binomial model (NB) to characterize the peptide counts, with which the mean and over-dispersion can be considered simultaneously with lower model complexity. Recent research also suggested NB is an optimal choice to model the abundance data with excessive zeros [25].

The protein abundance data has disrete integer values corresponding to protein counts and a lot of zeros corresponding to no protein abundance, and has very different sample variance from sample mean, which cannot be effectively modeled by a Possion distribution but can be fitted very well by a negative binomial distribution empirically. A negative binomial distribution with a positive real parameter r and

a real parameter γ ($0 < \gamma < 1$) is described in the following equation:

$$p(k|r,\gamma) = \binom{k+r-1}{r-1}\gamma^r(1-\gamma)^k = \frac{\Gamma(k+r)}{k!\Gamma(r)}\gamma^r(1-\gamma)^k. \quad (1)$$

In the above equation, k is an integer, and the mean of the distribution is $r\frac{1-\gamma}{\gamma}$. If we re-parametrize the negative binomial distribution using the mean parameter $\lambda = r\frac{1-\gamma}{\gamma}$ and the positive real parameter r, we have the following equation:

$$p(k|r,\gamma) = NB(k|\lambda,r) = \frac{\lambda^k}{k!}\frac{\Gamma(r+k)}{\Gamma(r)(r+\lambda)^k}\frac{1}{1+\frac{\lambda}{r}}. \quad (2)$$

In the above equation, $\Gamma(\cdot)$ is the Gamma function, r controls the over-dispersion of the distribution, and, when the over-dispersion paramter r approaches infinity, $NB(k|\lambda,r)$ approaches a Possion distribution with mean parameter λ. The above NB model uses r to adjust the variance independently of the mean parameter λ of the distribution, differing from a Possion distribution which has equal mean and variance. We model protein abundance data by NB using the parametrization in Equation 2. We assume the abundance of each protein i in tissue type t, W_{it}, follows NB distribution, with two parameters θ_{it} and r_t, $1 \le i \le N$ and $1 \le t \le T$, where N and T are the total number of genes (proteins) and total number of tissues types, respectively. Thus, the probability of protein i with peptide count k in tissue t can be modeled as the following in Equation 3.

$$p(W_{it} = k|\theta_{it}, r_t) = NB(k|\theta_{it}, r_t). \quad (3)$$

In Equation 3, θ_{it} represents the mean for protein i in tissue t and r_t represents the over-dispersion of the data, which was shared by all the proteins in the same tissue t. We then used hierarchical Bayesian negative binomial regression to regress the mean θ_{it} with miRNA expression in corresponding tissues, M_{jt}, $1 \le j \le J$, and $1 \le t \le T$, where J is the total number of miRNAs in the dataset. Equation 4 gives the regression of the mean in the model. Thus,

$$ln(\theta_{it}) = ln(\tau_t) - \rho_t \sum_{j=1}^{J} \omega_j \delta_{ij} M_{jt} \quad (4)$$

In Equation 4, τ_t stands for the background protein abundance shared by all the proteins in the same tissue t. As suggested in [10], we introduced δ_{ij} as a binary latent variable indicating whether or not the miRNA j regulates the gene i; ω_j is a regression coefficient associated with j-th miRNA expression shared by all the tissue types, and ρ_t is a scaling parameter for tissue t accounting for the measurement difference in different tissues. Since sequence complementarity is a necessary condition for true targets, we use a binary variable S_{ij} as the putative predictions

between miRNA j and protein i, which was derived from sequence-based predictions (TargetScanS, miRanda, or PicTar, etc); $S_{ij} = 1$ means that there is a putative predictions between i and j. The probability of a putative prediction being a true positive, p, is formally given in Equation 5.

$$\begin{aligned} p(\delta_{ij} = 1|S_{ij} = 1) = p, \\ p(\delta_{ij} = 1|S_{ij} = 0) = 0 \end{aligned} \tag{5}$$

To avoid over-fitting the data and to account for all possible uncertainties, we chose to use full Bayesian approach to infer δ_{ij} so that all the uncertainties and nuisance variables can be integrated out. Thus we assigned the priors to other parameters as follows (most were assigned flat priors):

$$\begin{aligned} p &\sim beta(1, 1), \\ \tau_t &\sim uniform(0, 50), \\ \rho_t &\sim gamma(\alpha, \alpha), \\ \alpha &\sim uniform(0, +\infty), \\ \omega_j &\sim exponential(\beta), \\ \beta &\sim uniform(0, 1000), \\ r_t &\sim exponential(a), \\ a &\sim uniform(0, 1000), \end{aligned} \tag{6}$$

With the likelihood and priors defined above, we then imple-mented Gibbs sampling [8] to compute marginal distribution of δ_{ij} conditioned on all evidence. All the inferences were made on drawing 5, 000 samples after 10,000 iterations.

3.3. *Apply Bayesian model to predict miRNA targets*

We applied the model described above to the 21,712 putative miRNA-protein interactions derived from TargetScan, and as-signed a confidence score to each putative interaction. Then we ranked the 21,712 putative interactions from the highest to the lowest confidence, and grouped them into 44 bins with each bin containing 500 ranked interactions. The results are shown in Figure 2A-D for 4 different tissue types.

As shown in Figure 2, our model can well capture the miRNAs repression effects in these four tissues. The miRNA-protein pairs that are predicted to have the highest confidence scores have lower protein abundance and higher miRNA expression; conversely the miRNA-protein pairs with the lowest confidence scores also have higher protein abundance and higher miRNA expression. For the interactions ranked with intermediate confidence scores, the miRNA expression is low, and the protein abundance can be either low or high. The fact that a large number of TargetScan predictions are located in the right side of the curve, i.e. low confidence

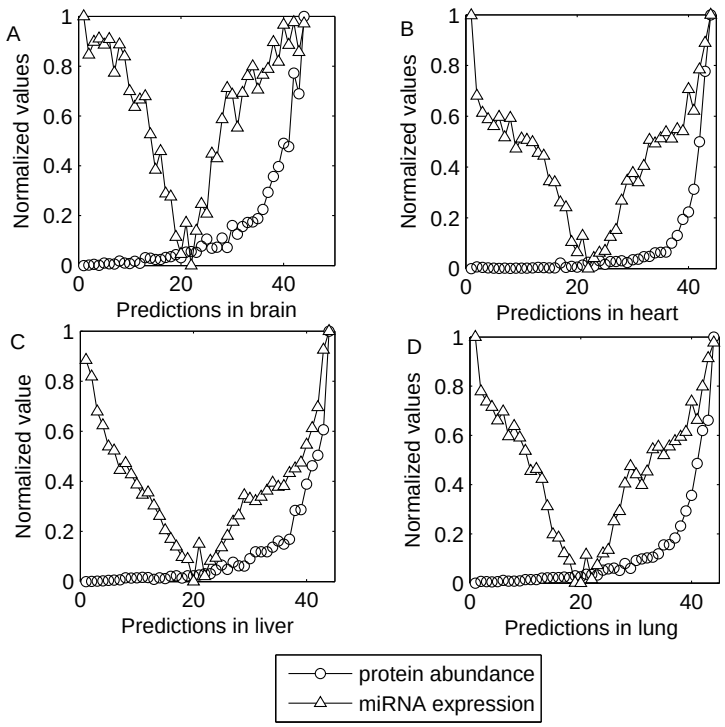

Fig. 2. miRNA targets prediction using miRNA expression and protein abundance. With our model, in the 4 tissues (panel A-D), the most confi-dent predictions (on the left) have the lowest protein abundance and the highest miRNA expression; while the least confident predictions (on the right) are high in both protein abundance and miRNA expression. All the data were scaled between 0 and 1. The putative predictions were from (Babak, et al, 2004) using TargetScan.

score with high miRNA expression and high protein abundance, indicates the extent of possible false-positives in the predictions made from sequence data alone. Because the Bayesian approach is intrinsically evidence-based, a prediction can only be made with high confidence if the miRNA is highly expressed in a certain tissue.

Note that the high-confidence miRNA-protein interaction pairs as shown in Figure 2 are predictions pooled from all 4 tissues. We do not explicitly model the tissue specificity of miRNAs in our formulism (see Equation 3); instead, the strengths of the miRNA regulation in specific tissues are inferred from the expression level of miRNAs. For instance, a miRNA can be interpreted as a functional regulator in a given tissue only if it is highly expressed and it has high confidence score with a potential target protein that is lowly expressed in that tissue.

3.4. *Blind tests for the Bayesian predictions method*

As described above, we only used the protein abundance and miRNA expression in brain, lung, heart and liver to train our model and make predictions; the data in the remaining two tissues (placenta and kidney) was left out during the model construction stage. To further validate our method, we subsequently conducted a blind test on the placenta and kidney data sets.

Figure 3A shows the results of the blind test. On the X-axis, we sorted the miRNA C protein pairs according to the confidence scores predicted by using the four training tissues; on the Y-axis, we plotted the protein abundance and miRNA expression level that are observed in placenta. The results indicate that, as a general trend, the predicted interactions can also reflect the desired tendency in placenta. The predicted interactions with high confidence usually have low protein abundance and high miRNA expression. The least confident predictions also have highly expressed proteins and highly expressed miRNAs, indicating those proteins are unlikely to be repressed by the miRNAs in placenta.

In kidney as the second blind test, shown in Figure 3B, al-though the miRNA expression data were not available for this tissue type, clearly our predictions were also effective and the most confident predictions have the lowest protein abundance and vice verse. All the above analyses were based on the sequence-based predictions from TargetScan. The same results also hold true after we repeated the analysis using predictions from another program miRanda [5]. Next, to test the robustness of our method, we shuffled the gene labels to randomize the data. The results from the shuffled data appear clearly random (Figure 2C), which strongly suggest that our prediction did not occur by chance.

3.5. *Comparison with TarBase and other methods*

We further searched for published experimental evidence for our predicted interactions in TarBase [18], which is a comprehensive database containing experimentally verified miRNA targets in a number of organisms. However, to this date, there are only 41 experimentally verified miRNA targets for mouse in the database. Since in the database all targets used gene symbol, we then converted the Swiss-Prot protein names in our study to corresponding gene symbols via http://idconverter.bioinfo.cnio.es/ [1]. However, except for the gene Arid3a (ARI3A_MOUSE), all other genes were not included in our dataset as they do not have protein abundance data compiled in this study. For Arida3a, in TarBase, it was annotated to be regulated by miR-125b. From our predictions based on miRanda predictions, the interactions between miR-125b and ARI3A_MOUSE was ranked among top 5% in all the 17,339 putative interactions, suggesting ARI3A_MOUSE is likely a true target. However, this interaction was not detected byTargetScan as compiled in [3].

We also compared our prediction results to the results obtained using the method in [10]. Although most of the predictions by both models are consistent, we

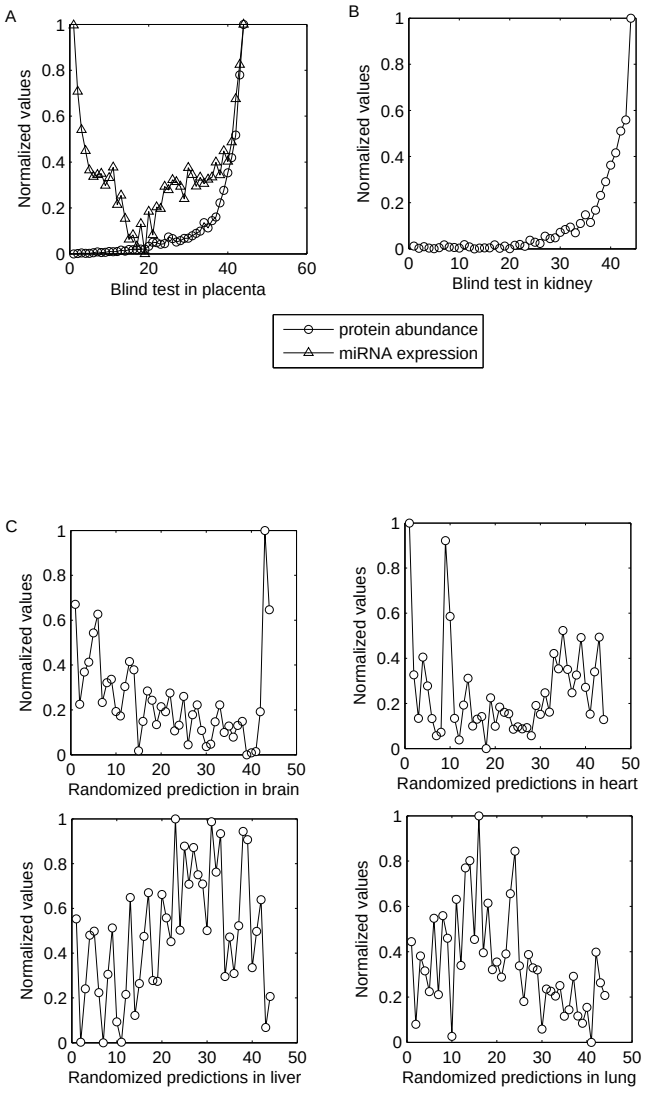

Fig. 3. Blind test of our predictions on placenta (A) and kidney (B). (C) Result from randomization while protein labels are shuffled.

found that proteomics data can further remove a lot of false negative predictions[a],

[a]Our Bayesian model filters sequence-based predictions and removes a lot of false positives pre-

and we believe that modeling proteomics data is the most reliable way of filtering miRNA target predictions when large-scale proteomics data become available. In details, the miRNA/target interactions such as mmu-mir-214/Q8R399_MOUSE, mmu-mir-211/Q8BYX4_MOUSE, mmu-miR-292-5p/KCNN3_MOUSE, and mmu-miR-298/RRAS2_MOUSE all ranked among top 10% in all the putative interactions in both models. However, the miRNA/target interactions such as mmu-miR-298/PLF3_MOUSE, mmu-miR-210/Q8BSZ8_MOUSE, and mmu-miR-92/8BZZ4_MOUSE all ranked among top 1% in all the putative interactions in our model, but they all ranked among bottom 15% in all the putative interactions in the model by [10]. We found that these miRNA/target pairs all have very good relatively high miRNA expression vs. relatively low protein abundance patterns, but they don't have very clear relatively high miRNA expression vs. relatively low mRNA expression patterns. Since miRNAs can either degrade mRNAs or repress mRNA translation, which will be discussed later in this paper, we believe that these interactions are very likely to be false negatives predicted by the model in [10].

3.6. *Discordant correlation between mRNA expression and protein abundance*

We next investigated whether these target genes are regulated by translational repression or by mRNA degradation. This can be achieved by comparing the protein abundance and mRNA expression data of predicted target genes. If miRNAs predominantly regulate their targets by degradation, then for target genes we would expect to see a good correlation between mRNA expression and their protein abundance (low mRNA concentration leads to low protein abundance). For all the putative targets, Figure 4 plots their average mRNA expression level and their total protein abundance across 4 tissues (brain, heart, liver and lung); the X-axis is the predicted targets sorted with the highest confidence on the left and lowest confidence on the right.

In Figure 4, interestingly, among the top ranked predicted targets, their mRNA expression fluctuates greatly; while among the bottom ranked predicted targets their mRNA expression is well correlated with their protein abundance. Such a lack of correlation between protein abundance and gene expressions suggested that degradation is unlikely to be the dominant mechanism for miRNA-mediated regulation, which is consistent with previous observations [3] and [9]. Such pronounced discordance between mRNA expression and protein abundance was also reported in a previous study [12]. In the next step, we describe another Bayesian model that allows us to trace back to the mRNA expression data, and calculate the probability of miRNA regulation by mRNA degradation or by transcriptional inhibition

dicted by sequence models.

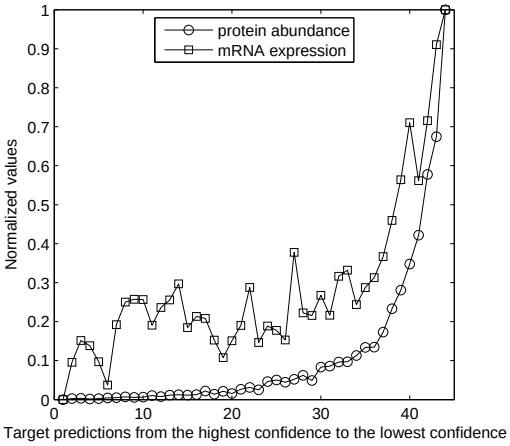

Fig. 4. mRNA expression and protein abundance for the gene targets of predicted interactions from the highest confidence (left) to the lowest con-fidence (right). The mRNA expression was averaged over the 5 tissue types (brain, heart, kidney, liver and lung) and protein abundance is the total peptide counts across 5 tissues (brain, heart, kidney, liver and lung). All the experimental data were scaled between 0 and 1.

3.7. *Two possible mechanisms: mRNA degradation vs. translational repression*

For the top miRNA-protein pairs that are predicted to be true regulator and targets, we can distinguish between these two possible regulatory mechanisms by analyzing the correlation between the miRNA expression and the mRNA expression. For example, if a top-ranked miRNA Cprotein pair has high miRNA expression and high mRNA expression, then it is a strong indicator that the protein target is regulated by translational expression. In contrast, if a predicted miRNA target has low mRNA expression, then it is likely regulated by mRNA degradation.

There are two common concerns in modeling the mRNA expression data: (1) the intrinsic low signal-to-noise ratio of microarray data, (2) the potential problem of missing values since a large number of the genes have expression levels measured as 0 [27]. To overcome these difficulties, we elected to discretize the mRNA expression data by using a cutoff of 0.1 to binarize the expression level to either low or high. For a given mRNA i in tissue t, $1 \leq i \leq L$ and $1 \leq t \leq T$, where L is the total number of mRNAs in the confident predictions derived from the first model, its mRNA expression R_{it} can be either low ($R_{it} = 0$) or high ($R_{it} = 1$). Let the probability of

degradation for mRNA i in tissue t be q_{it}, we assume,

$$P(R_{it} = k) = q_{it}^{(1-k)}(1 - q_{it})^k, k = 0 \ \ or \ \ 1. \tag{7}$$

We next used logistic regression to regress q_{it} with the expression of miRNAs that regulates gene i, in tissue t. Then we have the following equation,

$$logit(q_{it}) = log(\frac{q_{it}}{1 - q_{it}}) = \sum_{j=1}^{H} \Phi_j b_{ij} M_{jt} + c_t \tag{8}$$

in which H is the total number of miRNAs in the miRNA-mRNA interactions, M_{jt} is the expression of the j-th miRNA in tissue t, b_{ij} is a binary latent variable indicating whether or not the gene i is degraded by miRNA j , and Φ_j is a scaling parameter associated with the j-th miRNA, shared by all tissue types. The rationale behind Equation 8 is that for a given gene i, if its expression is low in tissue t, i.e. $R_{it} = 0$, then from the perspective of maximum likelihood, we need to maximize q_{it} so that the interactions between gene i and its regulating miRNAs that are highly expressed in tissue t should be assigned a higher degradation score. In this sense, the observed low expression of mRNA and high expression of miRNA together lead to the assignment of a high degradation probability. Similarly, if in tissue t, $R_{it} = 1$, then q_{it} needs to be minimized, implying those highly expressed miRNAs should be associated with a low degradation score, so the highly expressed miRNAs and mRNAs indicate such regulation is more likely to be through translational repression than through degradation.

Regarding the latent variable b_{ij}, we further required that:

$$\begin{aligned} p(b_{ij} = 1|\delta_{ij} = 1) = h, \\ p(b_{ij} = 1|\delta_{ij} = 0) = 0, \end{aligned} \tag{9}$$

in which, δ_{ij} indicates whether or not mRNA i is targeted by miRNA j. If $\delta_{ij} = 1$, then the miRNA j has a probability h to cause degradation to its target mRNA i. We then used a full Bayesian approach to estimate the parameters in the model to avoid overfitting the data and to account for all potential uncertainties. In the Bayesian framework, we then assigned priors to other parameters in the model as follows:

$$\begin{aligned} \Phi_j &\sim exponential(\Psi), \\ \Psi &\sim uniform(0, +\infty), \\ c_t &\sim uniform(-50, +50), \\ h &\sim beta(1, 1). \end{aligned} \tag{10}$$

Having defined the likelihood and the priors, we then inferred the posterior marginal distribution of $p(b_{ij} = 1|\mathbf{S}, W, M)$, conditioned on all the evidence. By implementing Gibbs sampling in the environment of WinBugs [20], all the inferences were based on drawing 5,000 samples after 10,000 iterations.

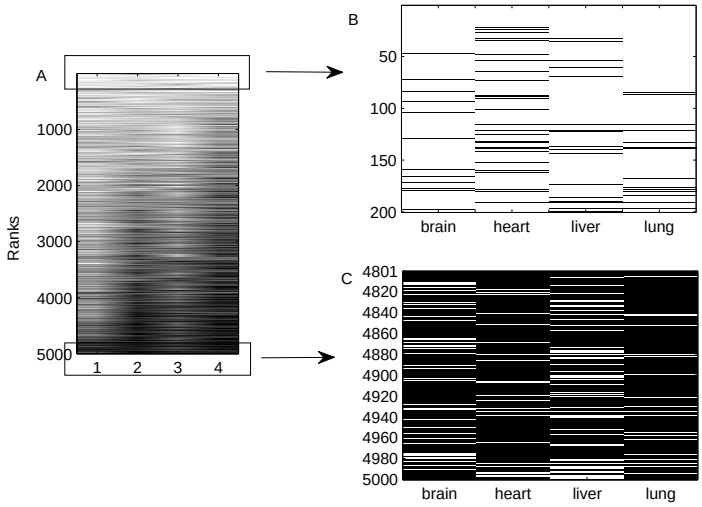

Fig. 5. (A) Ranking of miRNA targets according to the probability of being regulated by the mechanism of mRNA degradation; the targets were ranked from the highest degradation probability (top) to the lowest degradation probability (bottom). It also showed mRNA expression of the targets across 4 tissue types. Black color denotes high expression and white color denotes low expression. (B) The top ranked interactions have the highest degradation probability, and are associated with the low mRNA expression. (C) The bottom ranked interactions have the lowest degradation probability, and thus are associated with high mRNA expression.

3.8. *Apply the Bayesian model to mRNA data*

By implementing the model described above, we calculated the confidence scores for mRNA degradation for each miRNA-mRNA interaction pair, which indicated the likelihood that miRNA causes degradation to their mRNA targets. The lower degradation score implies higher probability of being translationally repressed. Then, we ranked the scores from the highest to the lowest, and grouped them into 50 bins, each containing 100 ranked interactions. Figure 5 shows the mRNA expression level of the ranked miRNA targets across 4 tissue types. The miRNA targets near the top of the Figure 5(A) have the highest probability of being regulated by mRNA degradation, as demonstrated by their low mRNA expression level (details shown in Figure 5(B)). Conversely the targets near the bottom have the highest probability of being regulated by translational repression (details shown in Figure 5(C)).

4. Discussion

4.1. *miRNA regulation by translational repression*

In this paper we described two novel formalisms in the computational analysis of miRNA regulation. We first introduced a Bayesian approach to identify miRNA targets based on protein abundance data. After having selected high confidence predictions, we then introduced a second Bayesian model to further distinguish the two possible regulatory mechanisms, i.e. mRNA degradation versus translational repression. We showed that our model is very effective in describing the three intertwining genomics data sets, i.e. miRNA expression, mRNA expression, and protein abundance. Our results demonstrated that protein abundance is a very useful resource in predicting miRNA targets, especially with the emerging evidence that translational repression is more prevalent than mRNA degradation as a regulatory mechanism in mammals. Our work also suggested that such repression mechanism likely contributed to the previously observed discordance between mRNA expression level and protein abundance. We would like to point out that although in this paper our model takes as input the predictions from TargetScan and miRanda, essentially results from any other sequence-based predictions can be used in our formulism.

4.2. *Potential limitations and future directions*

Even though our framework has obtained encouraging results, it certainly has limitations. We envision that it can be improved in the following area. (1) As we noted, miRNA is not the only mechanism of gene regulation. Some of the observed variations in protein abundance across tissues are likely the result of regulation at the transcription level by transcription factors, or at the post-transcriptional level by mRNA degradation pathways. (2) Although it has been reported that the mechanism of translational repression by miRNAs has little impact on mRNA level, the mRNA expression might be still helpful in predicting miRNA targets, and recent research suggested that targeted mRNA showing strong correlation (positive and negative) with miRNAs ([6] and [23]). In the future, we could incorporate the mRNA expression data with the proteomic data to build an integrated predictive model. (3) At this stage, our model takes as input the sequence-based predictions from another prediction programs such as TargetScan or miRANDA, therefore our algorithm does not explicitly consider the sequence complementarity and evolutionary conservation. As a future work, it would be interesting to extend our model to incorporate these properties into a unified probabilistic framework. (4) In our model, similar with [10], we assumed a single baseline distribution of protein abundance for all the genes in each tissue type. However, this is a significant simplification since different genes could have distinct baseline expression levels. The next step in this work is to take this into account and develop a more realistic expression baseline model. For example it would be possible to take into account the codon usage of

the genes to infer the possible baseline expression of a given gene [17].

Acknowledgments

ZZ acknowledges funding from Canadian Institutes of Health Research (CIHR). We thank Jim Huang, Quaid Morris, Yunchen Gong, and Ruth Isserlin for helpful discussions.

References

1. A. Alibes, P. Yankilevich, A. Canada, and R. Diaz-Uriarte. Idconverter and idclight: conversion and annotation of gene and protein ids. *BMC Bioinformatics*, 8:9, 2007.
2. V. Ambros. The functions of animal micrornas. *Nature*, 431(7006):350–5, 2004.
3. T. Babak, W. Zhang, Q. Morris, B. J. Blencowe, and T. R. Hughes. Probing micrornas with microarrays: tissue specificity and functional inference. *Rna*, 10(11):1813–9, 2004.
4. D. P. Bartel. Micrornas: genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–97, 2004.
5. A. J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, and D. S. Marks. Microrna targets in drosophila. *Genome Biol*, 5(1):R1, 2003.
6. Ana Eulalio, Eric Huntzinger, and Elisa Izaurralde. Getting to the root of mirna-mediated gene silencing. *Cell*, 132(1):9, 2008.
7. K. K. Farh, A. Grimson, C. Jan, B. P. Lewis, W. K. Johnston, L. P. Lim, C. B. Burge, and D. P. Bartel. The widespread impact of mammalian micrornas on mrna repression and evolution. *Science*, 310(5755):1817–21, 2005.
8. Andrew Gelman. *Bayesian data analysis*. Chapman and Hall, London, 1st edition, 1995.
9. L. He and G. J. Hannon. Micrornas: small rnas with a big role in gene regulation. *Nat Rev Genet*, 5(7):522–31, 2004.
10. J. C. Huang, Q. D. Morris, and B. J. Frey. Bayesian inference of microrna targets from sequence and expression data. *J Comput Biol*, 14(5):550–63, 2007.
11. M. Kiriakidou, P. T. Nelson, A. Kouranov, P. Fitziev, C. Bouyioukos, Z. Mourelatos, and A. Hatzigeorgiou. A combined computational-experimental approach predicts human microrna targets. *Genes Dev*, 18(10):1165–78, 2004.
12. T. Kislinger, B. Cox, A. Kannan, C. Chung, P. Hu, A. Ignatchenko, M. S. Scott, A. O. Gramolini, Q. Morris, M. T. Hallett, J. Rossant, T. R. Hughes, B. Frey, and A. Emili. Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling. *Cell*, 125(1):173–86, 2006.
13. A. Krek, D. Grun, M. N. Poy, R. Wolf, L. Rosenberg, E. J. Epstein, P. MacMenamin, I. da Piedade, K. C. Gunsalus, M. Stoffel, and N. Rajewsky. Combinatorial microrna target predictions. *Nat Genet*, 37(5):495–500, 2005.
14. B. P. Lewis, C. B. Burge, and D. P. Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets. *Cell*, 120(1):15–20, 2005.
15. B. P. Lewis, I. H. Shih, M. W. Jones-Rhoades, D. P. Bartel, and C. B. Burge. Prediction of mammalian microrna targets. *Cell*, 115(7):787–98, 2003.
16. L. P. Lim, N. C. Lau, P. Garrett-Engele, A. Grimson, J. M. Schelter, J. Castle, D. P. Bartel, P. S. Linsley, and J. M. Johnson. Microarray analysis shows that some micrornas downregulate large numbers of target mrnas. *Nature*, 433(7027):769–73, 2005.
17. A. C. McHardy, A. Puhler, J. Kalinowski, and F. Meyer. Comparing expression level-

dependent features in codon usage with protein abundance: an analysis of 'predictive proteomics'. *Proteomics*, 4(1):46–58, 2004.
18. P. Sethupathy, B. Corda, and A. G. Hatzigeorgiou. Tarbase: A comprehensive database of experimentally supported animal microrna targets. *Rna*, 12(2):192–7, 2006.
19. P. Sood, A. Krek, M. Zavolan, G. Macino, and N. Rajewsky. Cell-type-specific signatures of micrornas on target mrna expression. *Proc Natl Acad Sci U S A*, 103(8):2746–51, 2006.
20. Tomas A. Best N. Gilks W. Spiegelhalter, D.J. and D. Lunn. Bugs:bayesian inference using gibbs sampling. *MRC Biostatistics Unit, Cambridge, England. www.mrc-bsu.cam.ac.uk/bugs/*, 2003.
21. A. I. Su, T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. P. Cooke, J. R. Walker, and J. B. Hogenesch. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A*, 101(16):6062–7, 2004.
22. Zhongmin Tian, Andrew S. Greene, Jennifer L. Pietrusz, Isaac R. Matus, and Mingyu Liang. Micrornactarget pairs in the rat kidney identified by microrna microarray, proteomic, and bioinformatic analysis. *Genome Research*, 18(3):404–411, March 2008.
23. J. Tsang, J. Zhu, and A. van Oudenaarden. Microrna-mediated feedback and feedforward loops are recurrent network motifs in mammals. *Mol Cell*, 26(5):753–67, 2007.
24. M. Wakiyama, K. Takimoto, O. Ohara, and S. Yokoyama. Let-7 microrna-mediated mrna deadenylation and translational repression in a mammalian cell-free system. *Genes Dev*, 21(15):1857–62, 2007.
25. David Warton. Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. *Environmetrics*, 16:275–289, 2005.
26. X. Xu. Same computational analysis, different mirna target predictions. *Nat Methods*, 4(3):191; author reply 191, 2007.
27. W. Zhang, Q. D. Morris, R. Chang, O. Shai, M. A. Bakowski, N. Mitsakakis, N. Mohammad, M. D. Robinson, R. Zirngibl, E. Somogyi, N. Laurin, E. Eftekharpour, E. Sat, J. Grigull, Q. Pan, W. T. Peng, N. Krogan, J. Greenblatt, M. Fehlings, D. van der Kooy, J. Aubin, B. G. Bruneau, J. Rossant, B. J. Blencowe, B. J. Frey, and T. R. Hughes. The functional landscape of mouse gene expression. *J Biol*, 3(5):21, 2004.

Jingjing Li received M.Sc. degree in Electrical Engineering from Graduate School of Chinese Academy of Sciences in 2006. He is currently pursuing his Ph.D. degree in Department of Molecular Genetics at the University of Toronto since 2007.

Renqiang Min received his M.Sc. degree in Computer Science from Department of Computer Science at the University of Toronto in 2005, and he is currently pursuing his Ph.D. degree in Computer Science in the same department.

Anthony Bonner received his M.S. and Ph.D. degrees in Computer Science from Rutgers University, in 1990 and 1991, respectively. He was a post-doctoral fellow at INRIA-Rocquencourt, in France, and an assistant professor of computer science at Indiana University, Bloomington. Since 1991, he has been a faculty member at the University of Toronto, where he is now an associate professor of computer science and a member of the Collaborative Graduate Program in Genome Biology and Bioinformatics (CGPGBB). In 1999, he was a visiting professor in the Department of Computer and Information Science at the University of Pennsylvannia.

Zhaolei Zhang received his Ph.D. degree in biophysics from University of California, Berkeley in 2000. He joined Department of Molecular Genetics, University of Toronto in 2004. He is currently an associate professor affiliated with Department of Computer Science, Department of Molecular Genetics, Donnelly Centre for Cellular and Biomolecular Research, and Banting and Best Department of Medical Research, at the University of Toronto.