Detecting MicroRNA Targets by Linking Sequence, MicroRNA and Gene Expression Data

Jim C. Huang[1], Quaid D. Morris[2], Brendan J. Frey[1,2]

[1] Probabilistic and Statistical Inference Group, University of Toronto
10 King's College Road, Toronto, ON, M5S 3G4, Canada
[2] Banting and Best Department of Medical Research, University of Toronto
160 College St, Toronto, ON, M5S 3E1, Canada
jim@psi.toronto.edu

Abstract. MicroRNAs (miRNAs) have recently been discovered as an important class of non-coding RNA genes that play a major role in regulating gene expression, providing a means to control the relative amounts of mRNA transcripts and their protein products. Although much work has been done in the genome-wide computational prediction of miRNA genes and their target mRNAs, two open questions are how miRNAs regulate gene expression and how to efficiently detect *bona fide* miRNA targets from a large number of candidate miRNA targets predicted by existing computational algorithms. In this paper, we present evidence that miRNAs function by post-transcriptional degradation of mRNA target transcripts: based on this, we propose a novel probabilistic model that accounts for gene expression using miRNA expression data and a set of candidate miRNA targets. A set of underlying miRNA targets are learned from the data using our algorithm, GenMiR (**Gen**erative model for **miR**NA regulation). Our model scores and detects 601 out of 1,770 targets obtained from TargetScanS in mouse at a false detection rate of 5%. Our high-confidence miRNA targets include several which have been previously validated by experiment: the remainder potentially represent a dramatic increase in the number of known miRNA targets.

1 Introduction

Recent results show that there may not be many more mammalian protein-coding genes left to be discovered [9]. As a result, one of the main goals in genomics is now to discover how these genes are regulated. In the basic model for gene regulation, transcription factors act to enhance or suppress the transcription of a gene into messenger RNA (mRNA) transcripts. Recent evidence points to the existence of an alternative, post-transcriptional mechanism for gene regulation in which the abundances of transcripts and/or their protein products are reduced. In particular, microRNAs (miRNAs), a subclass of so-called non-coding RNA genes [8], have been identified as such a component of the cell's regulatory circuitry. miRNA genes do not go on to produce proteins, but instead produce short, 22-25 nt-long mature miRNA molecules. These then target mRNA transcripts through complementary base-pairing to short target sites.

miRNAs are believed to either trigger the degradation of their targets [3, 20] or repress translation of the transcript into protein [1]. There is substantial evidence that miRNAs are an important component of the cellular regulatory network, providing a post-transcriptional means to control the amounts of mRNA transcripts and their protein products. Previous work has focused primarily on the genome-wide computational discovery of miRNA genes [5, 19, 23] and their corresponding target sites [13, 16, 17, 24]. Experiments have shown that multiple miRNAs may be required to regulate a targeted transcript [16] and that miRNAs can regulate the expression of a substantial fraction of protein-coding genes with a diverse range of biological functions [1, 4].

Although many miRNA genes and target sites have been discovered by computational algorithms [5, 6], there remain two open problems in miRNA genomics. One is to determine whether miRNAs regulate their targets through the post-transcriptional degradation mechanism, through the translational repression mechanism, or possibly both. Another problem is the fact that there are relatively few miRNA targets which have experimental support [13, 16]. The computational algorithms used to find targets have limited accuracy [16, 17] due to the short lengths of miRNA target sites and thus empirical methods are needed to tease out true miRNA targets from false ones. Experimental validation of targets is currently done through *in vitro* reporter assays [13, 18] which provide some measure as to whether the miRNA binds to a target site. One concern with this type of assay is that a miRNA-target pair validated *in vitro* might not be biologically relevant inside the cell [1]. In addition, assays performed on a single miRNA-target pair might also erroneously reject the pair given that the combinatorial nature of miRNA regulation isn't taken into account and many miRNAs may be required to observe down-regulation of the targeted transcript. Finally, such assays are relatively expensive and time-consuming to conduct, so that only a handful of targets have been validated using this method. Expression profiling has been proposed as an alternative method for validating miRNA targets [20], but this has the problem of becoming intractable due to the combinatorial nature of miRNA regulation in which the action of many miRNAs must be taken into account.

While computational sequence analysis methods for finding targets and expression profiling methods have their own respective limitations, we can benefit from the advantages of both by combining the two methods [11] to detect miRNA targets. Given the thousands of miRNA targets being output by target-finding programs [13, 16, 17] and given the ability to profile the expression of thousands of mRNAs and miRNAs using microarrays [12, 21], we motivate a high-throughput computational technique for detecting miRNA targets in which both sequence and gene expression data are combined. The pipeline for detecting targets is shown in Fig. 1: a set of candidate miRNA targets is first generated using a target-finding program. Our model uses this set of candidates to account for gene expression using miRNA microarray expression data while taking into account the combinatorial nature of miRNA regulation. In this paper, we first address the question as to how miRNAs regulate gene expression: we

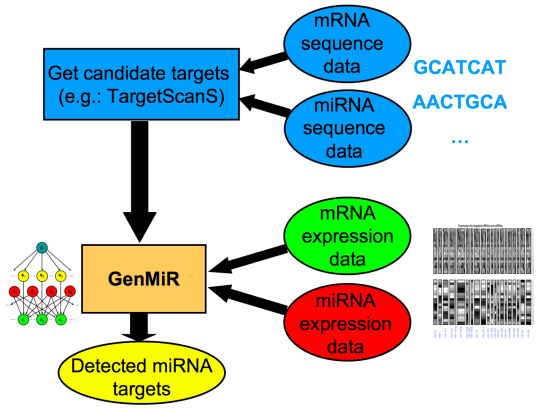


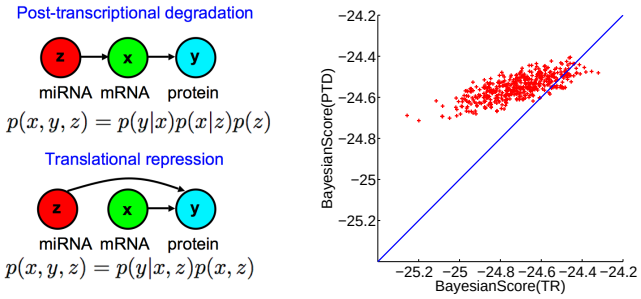
Fig. 1. Pipeline for detecting miRNA targets using GenMiR: a set of candidate targets is generated using a target-finding program (e.g.: TargetScanS). The candidates, along with expression data for mRNAs and miRNAs, are input into the GenMiR probability model. The output of the model consists of a set of miRNA targets which are well-supported by the data.

will present evidence in favor of the post-transcriptional degradation model for miRNA regulation. From this, we will formulate a probabilistic graphical model in which miRNA targets are learned from expression data. Under this model, the expression of a targeted mRNA transcript can be explained through the regulatory action of multiple miRNAs. Our algorithm, GenMiR (**Gen**erative model for **miR**NA regulation), learns the proposed model to find a set of miRNA targets which are biologically relevant. We will show that our model can accurately identify miRNA targets from expression data and detect a significant number of targets, many of which provide insight into miRNA regulation.

2 Post-transcriptional degradation (PTD) VS. translational repression (TR)

We will begin by addressing the question of whether miRNAs regulate gene expression by post-transcriptional degradation of target mRNAs [3] or by repressing translation of a targeted transcript [1] into proteins. In the first scenario, we expect that both mRNA expression levels and protein abundances will be decreased through the action of a miRNA. In the second scenario, protein abundances would be decreased without any necessary change in the expression of their parent mRNA. To determine which of the two mechanisms of miRNA regulation is most likely given biological data, we will present two simple *Bayesian networks* for the proposed mechanisms, shown in Figure 2a. Each network consists of a directed graph where nodes representing both miRNA and mRNA expression measures as well as protein abundances are linked via directed edges representing dependencies between the 3 variables. Thus, each network encodes a different set of dependencies between miRNA, mRNA and protein measures:

our aim here is to see which set of dependencies best describes biological reality.



(a) (b)

Fig. 2. (a) Bayes nets for degradation and repression regulatory mechanisms: each network presents a particular set of dependencies between miRNA, mRNA and protein measures (b) Scatter plot of scores obtained from both models for each miRNA-mRNA-protein triplet: most of the data is better accounted for by the PTD model than by the TR model.

To do so, we examined data profiling the expression of 78 mouse miRNAs [2] with mRNA expression data [25] paired to protein mass-spectrometry data [15] consisting of the measurements of $3,080$ mRNA-protein pairs across 5 tissues common to the 3 data sets. All the measured values were then ranked across tissues to get discrete rank values. We then used a set of human miRNA targets output from the target-finding program TargetScanS [17, 26]. These consisted of a total of $12,839$ target sites in human genes which were identified based both on miRNA-target site complementarity as well as conservation in 3'-UTR regions across 5 mammalian species (human, mouse, rat, dog and chicken). After mapping these targeted transcripts to the mouse mRNAs and miRNAs in the above data using the Ensembl database and BLAT [7], we were left with 473 candidate miRNA-target interactions involving 211 unique mRNA-protein pairs and 22 unique miRNAs.

With the above data in hand, we gathered statistics over mRNA, protein and miRNA measurements x, y, z across tissues $t = 1, \cdots, 5$ for each putative miRNA targets. We then scored the two regulatory models for each miRNA/mRNA/protein triplet using Bayesian scores [10] computed as

$$\begin{aligned} \text{BayesianScore}(PTD) &= \sum_t \log\Big(p(y_t|x_t)p(x_t|z_t)p(z_t)\Big) \\ \text{BayesianScore}(TR) &= \sum_t \log\Big(p(y_t|x_t,z_t)p(x_t,z_t)\Big) \end{aligned} \tag{1}$$

where each conditional probability term is a multinomial probability averaged over a Dirichlet prior distribution. These scores correspond to the log-likelihood

of a miRNA/mRNA/protein data triplet given one of the two models for miRNA regulation. Figure 2b shows a scatter plot of the 2 scores for each mRNA-miRNA-protein triplet: we can see that in the vast majority of cases, the PTD model offers a far better fit to our data than the TR model, providing good evidence in favor of the PTD model within the current debate of whether miRNAs act by degrading targets or by repressing translation [1, 3]. With this result in hand, we will now motivate the use of both mRNA and miRNA expression data to detect miRNA targets.

3 Exploring miRNA targets using microarray data

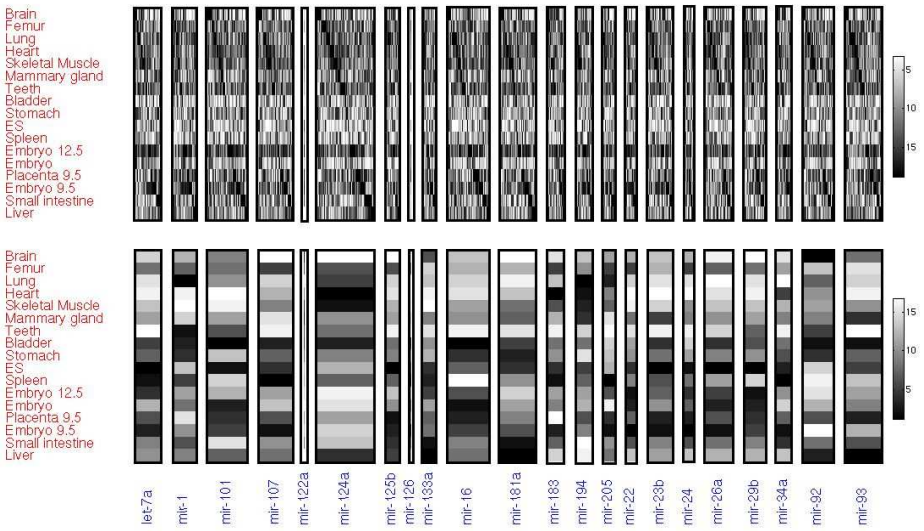


Fig. 3. Rank expression profiles of targeted mRNAs and corresponding miRNAs: each profile measures expression across 17 mouse tissues. For targeted mRNA transcripts (top row), a rank of 17 (black) denotes that the expression in that tissue was the highest amognst all tissues in the profile whereas a rank of 1 (white) denotes that expression in that tissue was lowest amongst all tissues. miRNA intensities are shown using a reverse colormap (bottom row), with a rank of 17 (white) denoting that the expression was highest and a rank of 1 (black) denotes that expression in that tissue was lowest. Each miRNA targets and down-regulates multiple mRNA transcripts and a given mRNA transcript may be targeted by multiple miRNAs.

To explore putative relationships between mRNAs and miRNAs, we used the above microarray expression data profiling the expression of a total of $41,699$ mRNA transcripts and 78 miRNAs across 17 tissues common to both data sets: expression values consisted of arcsinh-normalized intensity values in the same range, with negative miRNA intensities were thresholded to 0. From the above

set of 12,839 TargetScanS targets, 1,770 are represented across this set of miRNAs and mRNAs in the form of 788 unique mRNAs and 22 unique miRNAs: the set of putative miRNA-mRNA pairs are shown in Fig. 3. Given the expression data and a set of putative targets, we looked for examples of down-regulation in which the expression of a targeted mRNA transcript was low in a given tissue and the targeting miRNA was highly expressed in that same tissue.

Among miRNAs in the data from [2], miR-16 and miR-205 are two that are highly expressed in spleen and embryonic tissue respectively (Fig. 3). The cumulative distribution of expression of their targeted mRNAs in these two tissues is shown in Fig. 4. The plots show that the expression of targeted mRNAs is negatively shifted with respect to the background distribution of expression in these two tissues ($p < 10^{-7}$ and $p < 0.0015$ using a one-tailed Wilcoxon-Mann-Whitney test, Bonferroni-corrected at $\alpha = 0.05/22$). This result suggests that regulatory interactions predicted on the basis of genomic sequence can be observed in microarray data in the form of high miRNA/low targeted transcript expression relationships. While it is feasible to find such relationships for a single miRNA using an expression profiling method [20], to test for the more realistic scenario in which mRNA transcripts are down-regulated by multiple miRNAs, we would require a large number of microarray experiments for a large number of miRNAs. Additional uncertainty would be introduced by miRNAs that are expressed in many tissues. An alternative is to use data which profiles the expression of mRNAs and miRNAs across many tissues and formulate a statistical model which links the two using a set of candidate miRNA targets. A sensible model would account for negative shifts in tissue expression for targeted mRNA transcripts given that the corresponding miRNA was also highly expressed in the same tissue. By accounting for the fact that miRNA regulation is combinatorial in nature [4, 16], we will construct such a model which will hopefully capture the basic mechanism of miRNA regulation. The model takes as inputs a set of candidate miRNA targets and expression data sets profiling both mRNA transcripts and miRNAs: it then accounts for examples of down-regulation in the expression data to output a subset of the candidate miRNA targets which are well-supported by the data.

4 A probabilistic graphical model for miRNA regulation

In this section, we describe our model of miRNA regulation. Under this model, the expression of a targeted transcript can be reduced by the action of one or many miRNAs, each of which will be allowed to reduce it by some fixed amount. Conditioned on observing high expression of one or many miRNAs, the expression of a targeted transcript is negatively shifted with respect to a background level which is to be estimated. The particular miRNAs that will participate in targeting a transcript will be selected using a set of unobserved binary indicator variables; the problem of detecting miRNA targets will therefore consist of inferring which of these indicator variables are turned on and which are turned off given observed data.

Consider two separate expression data sets profiling N mRNA transcripts and M miRNAs across T tissues. Let indices $i = 1, \cdots, N$ and $j = 1, \cdots, M$

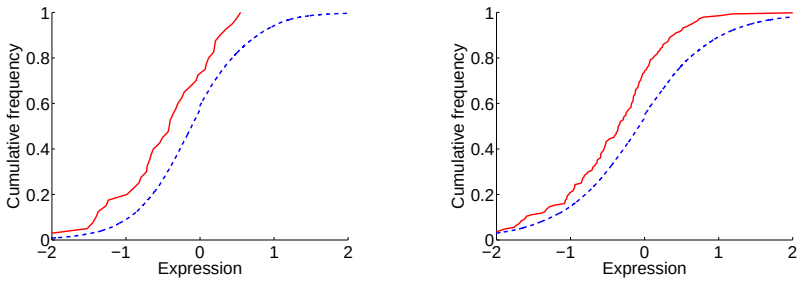


Fig. 4. Effect of miRNA negative regulation on mRNA transcript expression: shown are cumulative distributions for (a) Expression in embryonic tissue for mRNA transcripts targeted by miR-205 (b) Expression in spleen tissue for mRNA transcripts targeted by miR-16. A shift in the curve corresponds to down-regulation of genes targeted by miRNAs. Targets of miR-205 and miR-16 (solid) show a negative shift in expression with respect to the background distribution (dashed) in tissues where miR-205 and miR-16 are highly-expressed.

denote particular mRNA transcripts and miRNAs in our data sets. Let $\mathbf{x}_i = [x_{i1} \cdots x_{iT}]^{\mathrm{T}}$ and $\mathbf{z}_j = [z_{j1} \cdots z_{jT}]^{\mathrm{T}}$ be the expression profiles over the T tissues for mRNA transcript i and miRNA j such that x_{it} is the expression of the i^{th} transcript in the t^{th} tissue and z_{jt} is the expression of the j^{th} miRNA in the same tissue. Suppose now that we are given a set of candidate miRNA-target interactions in the form of an $N \times M$ binary matrix $\mathbf{C}$ where $c_{ij} = 1$ if miRNA j putatively targets transcript i and $c_{ij} = 0$ otherwise. The matrix $\mathbf{C}$ therefore contains an initial set of candidate miRNA targets for different miRNAs: these are putative miRNA-mRNA regulatory relationships within which we will search for cases which are supported by the microarray data.

Due to noise in the sequence and expression data, the limited accuracy of computational target-finding programs as well as incomplete knowledge of the regulatory network of the cell, there is uncertainty as to which miRNA targets are in fact biologically relevant. We can represent this uncertainty using a set of unobserved binary random variables indicating which of the candidate miRNA targets are well supported by the data. We will assign an unobserved random variable s_{ij} to each candidate miRNA-target interaction such that $s_{ij} = 1$ if miRNA j genuinely targets mRNA transcript i. Then, the problem of detecting miRNA targets can be formulated in terms of finding a subset $\{(i,j) \in \mathbf{C} | s_{ij} = 1\}$ such that miRNA-target interactions in this subset are supported by the observed expression data.

We can now describe a relationship between the expression of a targeted mRNA transcript and a miRNA in tissue t:

$$\begin{aligned} E[x_{it}|s_{ij} = 1, z_{jt}, \Theta] &= \mu_t - \lambda_j z_{jt}, \quad \lambda_j > 0 \\ E[x_{it}|s_{ij} = 0, z_{jt}, \Theta] &= \mu_t \end{aligned} \tag{2}$$

where λ_j is some positive regulatory weight that determines the relative amount of down-regulation incurred by miRNA j, μ_t is a background expression parameter and Θ consists of the λ_j and μ_t parameters. Thus, the above explicitly models the relationship observed in Fig. 4 in which the expression of a targeted transcript is negatively shifted with respect to the background given that a miRNA is highly expressed. Thus, miRNAs can never directly increase the expression of their target transcripts, in accordance with current evidence about their functions [1, 20].

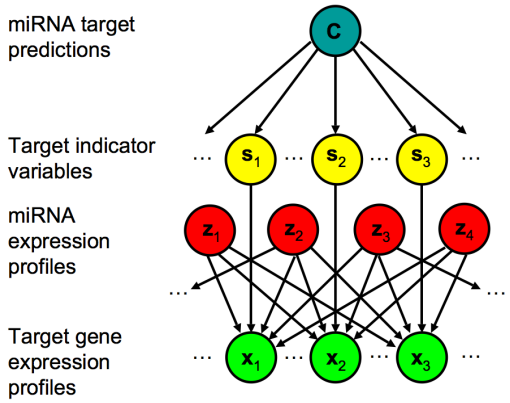


Fig. 5. Bayesian network used for detecting miRNA targets: each mRNA transcript is assigned a set of indicator variables which select for miRNAs that are likely to regulate it given the data.

We can extend the above to allow for multiple miRNAs to cooperate in tuning the expression of a targeted transcript. Here, each miRNA is allowed to decrease the expression of its target transcript by some relative amount λ_j such that

$$E[x_{it}|\{s_{ij}\},\{z_{jt}\},\Theta] = \mu_t - \sum_j \lambda_j s_{ij} z_{jt}, \quad \lambda_j > 0$$

We are now able to present our model for miRNA regulation: we will do so using the framework offered by probabilistic graphical models, where we can explicitly represent dependencies between observed and unobserved variables as well as model parameters using a directed graph. If we denote the prior targeting probabilities as $p(s_{ij} = 1|c_{ij} = 1) = \pi$ and $\mathbf{S}$ as the set of s_{ij} variables, we can write the probabilities in our model given the expression of the miRNAs and a set of candidate miRNA targets as

$$\begin{aligned} p(\mathbf{x}_i|\mathbf{Z},\mathbf{S},\Theta) &= \mathcal{N}(\mathbf{x}_i;\boldsymbol{\mu} - \textstyle\sum_j \lambda_j s_{ij}\mathbf{z}_j, \boldsymbol{\Sigma}), \quad \lambda_j \geq \epsilon > 0 \\ p(\mathbf{S}|\mathbf{C},\Theta) &= \textstyle\prod_{(i,j)} p(s_{ij}|\mathbf{C},\Theta) = \prod_{(i,j)\notin\mathbf{C}} \delta(s_{ij},0) \prod_{(i,j)\in\mathbf{C}} \pi^{s_{ij}} (1-\pi)^{1-s_{ij}} \\ p(\mathbf{X},\mathbf{S}|\mathbf{C},\mathbf{Z},\Theta) &= \textstyle\prod_i p(\mathbf{x}_i|\mathbf{Z},\mathbf{S},\Theta) \prod_j p(s_{ij}|\mathbf{C},\Theta) \end{aligned} \tag{3}$$

where ϵ is a lower bound on the regulatory weights λ_j, $\delta(\cdot,\cdot)$ is the Dirac delta function, $\mathbf{X}$ and $\mathbf{Z}$ are the sets of observed expression profiles for mRNAs and miRNAs, $\mathbf{C}$ is the set of candidate miRNA targets and $\Theta = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \pi, \{\lambda_j\}_{j=1}^{M}\}$ is the set of model parameters containing the background expression $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, the prior targeting probability π and the miRNA regulatory weights λ_j. The parameter ϵ addresses the constraint that miRNAs cannot make a null contribution to the expression of its target given that the interaction between the miRNA and its target is valid.

The above model links the expression profiles of mRNA transcripts and miRNAs given a set of candidate miRNA targets: within these, we will search for relationships which are supported by the data by inferring the settings for the unobserved s_{ij} variables. Fig. 5 shows the Bayesian network corresponding to Equation 3: each mRNA transcript in the network is assigned a set of indicator variables which select for miRNAs that are likely to regulate it given the data.

Having presented the above probabilistic model for miRNA regulation, we are now faced with the task of inference, or computing the posterior probability $p(s_{ij}|\mathbf{x}_i, \mathbf{Z}, \mathbf{C}, \Theta) \propto \sum_{\mathbf{S}\setminus s_{ij}} p(\mathbf{x}_i, \mathbf{S}|\mathbf{Z}, \mathbf{C}, \Theta)$ that a given miRNA target is valid conditioned on the data. Exact inference would require summing over a number of terms exponential in the number of miRNAs. Unfortunately, this summation will be computationally prohibitive for transcripts which are targeted by a large number of miRNAs. Thus, we will turn instead to an approximate method for inference which will make the problem tractable.

5 Variational learning for detecting miRNA targets

For variational learning [14, 22] in a graphical model with latent variables $\mathcal{H}$ and observed variables $\mathcal{E}$, the exact posterior $p(\mathcal{H}|\mathcal{E})$ is approximated by a distribution $q(\mathcal{H};\phi)$ parameterized by a set of variational parameters ϕ. Variational inference therefore consists of an optimization problem in which we optimize the fit between $q(\mathcal{H};\phi)$ and $p(\mathcal{H},\mathcal{E})$ with respect to the variational parameters ϕ. This fit is measured by the Kullback-Leibler (KL) divergence $D(q||p)$ between the q and p distributions, which can be written as

$$D(q||p) = \int_{\mathcal{H}} q(\mathcal{H};\phi) \log \frac{q(\mathcal{H};\phi)}{p(\mathcal{H},\mathcal{E})}\, d\mathcal{H} = \sum_{\mathbf{S}} q(\mathbf{S}|\mathbf{C}) \log \frac{q(\mathbf{S}|\mathbf{C})}{p(\mathbf{X},\mathbf{S}|\mathbf{C},\mathbf{Z},\Theta)}$$

where $\mathbf{X}$, $\mathbf{Z}$, $\mathbf{C}$ and $\mathbf{S}$ have been substituted as the observed and latent variables $\mathcal{E}$ and $\mathcal{H}$ respectively.

The approximating distribution can be further simplified via a mean-field decomposition of the q-distribution in which all the latent s_{ij} variables are assumed to be independent and thus

$$q(\mathbf{S}|\mathbf{C}) = \prod_{i,j} q(s_{ij}|\mathbf{C}) = \prod_{(i,j)\in\mathbf{C}} \beta_{ij}^{s_{ij}} (1-\beta_{ij})^{1-s_{ij}} \tag{4}$$

where the variational parameters β_{ij} will be fitted to the observed data $\mathbf{X}$, $\mathbf{Z}$, $\mathbf{C}$. We will therefore approximate the intractable posterior $p(s_{ij}|\mathbf{x}_i, \mathbf{Z}, \mathbf{C}, \Theta)$ with

a simpler distribution $q(s_{ij}|\mathbf{C})$ which will make inference tractable. If we write the expected sufficient statistics $\mathbf{u}_i$, $\mathbf{W}$ and $\mathbf{V}$ as

$$\begin{aligned}
\mathbf{u}_i &= \sum_{j:(i,j)\in\mathbf{C}} \lambda_j \beta_{ij} \mathbf{z}_j \qquad (5)\\
\mathbf{W} &= \frac{1}{N}\sum_i \Big(\mathbf{x}_i - (\boldsymbol{\mu} - \mathbf{u}_i)\Big)\Big(\mathbf{x}_i - (\boldsymbol{\mu} - \mathbf{u}_i)\Big)^T \\
\mathbf{V} &= \frac{1}{N}\sum_i \sum_{j:(i,j)\in\mathbf{C}} (\beta_{ij} - \beta_{ij}^2)\lambda_j^2 \mathbf{z}_j \mathbf{z}_j^T
\end{aligned}$$

then the KL divergence $D(q||p)$ can be written simply as

$$\begin{aligned}
D(q||p) = \sum_{(i,j)\in\mathbf{C}} \Big(\beta_{ij}\log\frac{\beta_{ij}}{\pi} + (1-\beta_{ij})\log\frac{1-\beta_{ij}}{1-\pi}\Big) + \frac{N}{2}\log|\boldsymbol{\Sigma}| \\
+ \frac{N}{2}\mathrm{tr}\Big(\boldsymbol{\Sigma}^{-1}(\mathbf{W}+\mathbf{V})\Big) + \text{const.} \qquad (6)
\end{aligned}$$

Approximate inference and parameter estimation will be accomplished via the variational EM algorithm [14, 22], which iteratively minimizes $D(q||p)$ with respect to the set of variational parameters (E-step) and the model parameters (M-step) until convergence to a local minimum. Thus, taking derivatives of $D(q||p)$ and setting to zero yields the following updates:

Variational E-step:

$$\forall (i,j)\in\mathbf{C},\quad \frac{\beta_{ij}}{1-\beta_{ij}} = \frac{\pi}{1-\pi}\exp\left[-\lambda_j \mathbf{z}_j^T \boldsymbol{\Sigma}^{-1}\Big(\mathbf{x}_i - \big(\boldsymbol{\mu} - (\sum_{k\neq j:(i,k)\in\mathbf{C}} \lambda_k\beta_{ik}\mathbf{z}_k + \frac{\lambda_j}{2}\mathbf{z}_j)\big)\Big)\right] \qquad (7)$$

Variational M-step:

$$\begin{aligned}
\boldsymbol{\mu} &= \frac{1}{N}\sum_i (\mathbf{x}_i + \mathbf{u}_i) \qquad (8)\\
\boldsymbol{\Sigma} &= \mathrm{diag}\Big(\mathbf{W}+\mathbf{V}\Big) \\
\forall j,\quad \lambda_j &= \max\left(-\frac{\sum_i \beta_{ij}\mathbf{z}_j^T\boldsymbol{\Sigma}^{-1}\Big(\mathbf{x}_i - (\boldsymbol{\mu} - \sum_{k\neq j:(i,k)\in\mathbf{C}}\lambda_k\beta_{ik}\mathbf{z}_k)\Big)}{\sum_i \beta_{ij}\mathbf{z}_j^T\boldsymbol{\Sigma}^{-1}\mathbf{z}_j}, \epsilon\right) \\
\pi &= \frac{\sum_{(i,j)\in\mathbf{C}}\beta_{ij}}{\mathrm{card}(\mathbf{C})} \qquad (9)
\end{aligned}$$

where the expected sufficient statistics $\mathbf{u}_i$, $\mathbf{W}$ and $\mathbf{V}$ are obtained from the E-step. Now that we have defined the update equations for performing inference and estimating model parameters, we will use the above algorithm to learn a subset of candidate miRNA targets which are biologically relevant.

6 Results

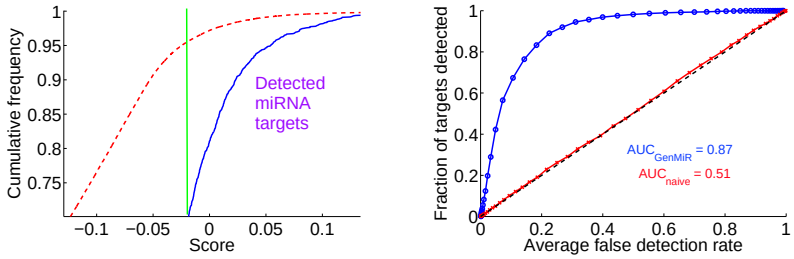
We can now turn to the problem of detecting miRNA targets. We start from the above set of $1,770$ TargetScanS 3'-UTR targets and use the microarray data from [2] and [25] to learn our model. The GenMiR algorithm was initialized with $\beta_{ij} = \pi = 0.5 \; \forall (i,j) \in \mathbf{C}$, $\lambda_j = 0.01 \; \forall j = 1, \cdots, M$ and the parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ were initialized to the sample mean and covariance of the mRNA expression data. The algorithm was run for 200 iterations or until convergence with $\epsilon = 0.01$. Once our model has been learned from the data, we assign a score to each of the candidate interactions $(i,j) \in \mathbf{C}$ according to

$$Score(i,j) = \log_{10} \left(\frac{\beta_{ij}}{1 - \beta_{ij}} \right) \tag{10}$$

Thus a miRNA-mRNA pair is awarded a higher score if it was assigned a higher probability of being *bona fide* under our model given the data.

To assess the accuracy of our model, we performed a series of permutation tests in which we learned scores from data where the mRNA transcript labels were permuted, under the null hypothesis is that there are no regulatory interactions between mRNAs and miRNAs. We generated 100 data sets in which transcript labels were permuted and we learned our model on each data set. The resulting empirical cumulative distributions over scores for both the permuted and unpermuted data are shown in Fig. 6a. The plot indicates that many of the candidate miRNA targets may be *bona fide*, as significantly more miRNA targets can be learned from the unpermuted data than from the permuted data ($p < 10^{-24}$, WMW).

To make a set of predictions, we can threshold the score: for different values of this threshold, we get a certain number of false detections. We can estimate the sensitivity and specificity of each threshold value by comparing the number of miRNA targets with score above the threshold to the average number of targets corresponding to the permuted data which also have a score above the threshold. By varying the threshold, we obtain the curve shown in Fig. 6b which relates the fraction of candidate targets detected ({# of candidate targets detected}/{# of candidate targets}) to the average false detection rate ({Average # of permuted targets detected}/{# of candidate targets}) for different threshold values, where the average false detection rate is computed for each threshold value using the average fraction of permuted miRNA-targets that are detected over the 100 permutations. Setting the threshold score to -0.022 to control for an average false detection rate of 5%, we have detected a total of 601, or 34% of the $1,770$ TargetScanS candidates. This suggests that many biologically relevant miRNA targets can be found in our expression data and that our model is able to accurately find them.



(a) (b)

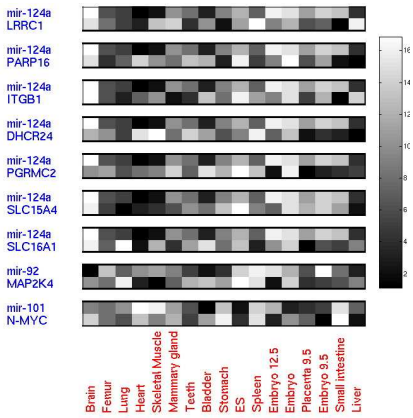
Fig. 6. (a) Empirical cumulative distribution of scores for the permuted data (dashed) and unpermuted data (solid); all scores above the threshold score correspond to detected miRNA targets (b) Fraction of candidate targets detected ({# of candidate targets detected}/{# of candidate targets}) VS. average false detection rate ({Average # of permuted targets detected}/{# of candidate targets}) using both GenMiR scoring (circle) and naive Pearson correlation scoring (cross), with areas under the curves (AUC) shown.

While the above results are encouraging, we might wonder as to whether our model offers any real advantage over naively detecting miRNA targets using Spearman correlation, where we would expect that the expression profiles corresponding to valid miRNA-mRNA pairs are anti-correlated across tissues. By looking at candidate miRNA targets independently of one another using this score, we obtain the curve shown in Fig. 6b. The plot shows that by looking at a single miRNA-mRNA pair and ignoring the action of other miRNAs, the naive method leads to poor performance. In contrast, the GenMiR algorithm can detect a higher number of candidate miRNA targets for a given number of false detections by taking into account multiple miRNAs per targeted transcript, obtaining a good overall fit to the data.

6.1 Biologically relevant targets detected by GenMiR

Within our set of high-confidence miRNA targets, we observe some of the small number of targets that have experimental support (Fig. 7). In particular, we correctly predict the interaction between miR-101 and the mouse homolog of the human N-MYC gene [18], as well as the relationship between miR-92 and MAP2K4 [18], a gene involved in signal transduction. In addition, we recovered 7 mouse homologs of human transcripts that were shown to be downregulated [20, 26] in brain by miR-124a.

The remainder of our miRNA targets potentially represent a dramatic increase in the number of known targets. The full list of miRNA targets detected using GenMiR and their corresponding scores is available on the project web page (see Appendix), along with GO annotations. The broad range of GO annotations for our miRNA targets further reinforces the prevalent hypothesis [1, 4] that miRNAs indeed regulate a wide variety of biological processes. Given the above results, we believe that most of these targets are biologically relevant and provide insight into miRNA regulation.



miRNA	Human Gene ID	GenMiR score
miR-124a	LRRC1	-0.0210
miR-124a	PARP16	0.0170
miR-124a	ITGB1	0.0108
miR-124a	DHCR24	-0.0132
miR-124a	PGRMC2	-0.0022
miR-124a	SLC15A4	0.0142
miR-124a	SLC16A1	-0.0026
miR-92	MAP2K4	0.0133
miR-101	N-MYC	-0.0064

Fig. 7. Rank expression profiles of experimentally validated miRNA targets across 17 mouse tissues which are also detected by our model. For targeted mRNA transcripts, a rank of 17 (black) denotes that the expression in that tissue was the highest amongst all tissues in the profile whereas a rank of 1 (white) denotes that expression in that tissue was lowest amongst all tissues. Targeting miRNA intensities are shown using a reverse colormap, with a rank of 17 (white) denoting that the expression was highest and a rank of 1 (black) denotes that expression in that tissue was lowest.

7 Discussion

In this paper we have presented evidence that miRNAs indeed regulate gene expression by degrading their target transcripts. Using this as a foundation, we have developed GenMiR, a novel probabilistic model and learning algorithm for detecting miRNA targets by combining candidate targets with a model for mRNA and miRNA expression. Our model accounts for both mRNA and miRNA expression microarray data given a set of candidate targets and learns the underlying set of biologically relevant miRNA targets. We have shown how to learn the model from expression data: the learned model has been shown to provide a good representation of miRNA regulation and can be used to accurately identify miRNA targets from expression data.

Our model is the first to explicitly use expression data and the combinatorial aspect of miRNA regulation to detect miRNA targets. Previous work done in [24] has focused on *de novo* finding of targets based on sequence and then associating miRNAs to their activity conditions through mRNA expression data alone. Our work differs from [24] in that we use observed miRNA expression to detect miRNA targets, whereas the model from [24] did not detect targets on the basis of miRNA expression data. In contrast to that method, we are explicitly modeling the generative process for mRNA expression given both miRNA expression and a set of candidate targets to perform detection: our model also explicitly takes into account the influence of multiple miRNAs on the expression of a single targeted mRNA transcript, an important feature which the model of [24] lacks.

We note that there are many sources of noise in our pipeline for detecting miRNA targets: these include the significant amounts of noise in the microarray

data sets and different hybridization conditions for the two microarray experiments. Additional noise is introduced by errors in the human genomic sequence data used to find the candidate targets, false positives within the set of candidate targets and the lossy mapping between the human and mouse genomes when mapping targets to our data. As a result, the fraction of candidate miRNA targets (34%) that we detect in our mouse expression data is surprisingly high. Given that we can accurately detect many miRNA targets in the presence of abundant noise using a relatively simple model, we can think of several ways in which we could extend the model to mitigate these sources of uncertainty.

We could learn from expression data given candidate miRNA targets from several target-finding programs and examine over-represented high-scoring targets. We could also relax the current assumption that the entire population of genes is generated from a single background expression profile: instead, we could model the background expression of co-expressed groups of genes. We expect that extending the model along these dimensions will greatly increase the accuracy with which we can identify biologically relevant miRNA targets from expression data and we are actively pursuing these ideas. In closing, our model provides a probabilistic framework for finding miRNA targets which uses microarray data. The model makes significant progress towards understanding the functional genomics of miRNAs, providing insight into a key mechanism of gene regulation.

8 Acknowledgment

We would like to thank Tomas Babak for many insightful discussions. JCH was supported by a NSERC Postgraduate Scholarship and a CIHR Net grant. QDM was supported by a NSERC Postdoctoral Fellowship. BJF was supported by a Premier's Research Excellence Award and a gift from Microsoft Corporation.

9 Appendix

A supplementary table containing all detected miRNA targets and corresponding human gene identifiers can be found at *http://www.psi.toronto.edu/~GenMiR*

References

1. Ambros V. The functions of animal microRNAs. *Nature* **431**, pp. 350-355, 2004.
2. Babak T, Zhang W, Morris Q, Blencowe BJ and Hughes, TR. Probing microRNAs with microarrays: Tissue specificity and functional inference. *RNA* **10**, pp. 1813-1819., 2004.
3. Bagga S *et al.* Regulation by let-7 and lin-4 miRNAs results in target mRNA degradation. *Cell* **122**, pp. 553-63, 2005.
4. Bartel DP. MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell* **116**, pp.281-297, 2004.
5. Bentwich I *et al.* Identification of hundreds of conserved and nonconserved human microRNAs. *Nature Genetics* **37**, pp. 766-770, 2005.
6. Berezikov E *et al.* Phylogenetic shadowing and computational identification of human microRNA genes. *Cell* **120(1)**, pp. 21-4, 2005.

7. Kent W.J. BLAT – The BLAST-Like Alignment Tool. *Genome Research* **4**, pp. 656-664, 2002.
8. Eddy S. Non-coding RNA genes and the modern RNA world. *Nature Reviews Genetics* **2**, pp. 919-929, 2001.
9. Frey BJ *et al.* Genome-wide analysis of mouse transcripts using exon-resolution microarrays and factor graphs. *Nature Genetics* **37**, pp. 991-996, 2005.
10. Hartemink A, Gifford D, Jaakkola T, and Young R. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Proceedings of the Pacific Symposium on Biocomputing 2001, World Scientific: New Jersey*, pp. 422-433.
11. Huang JC, Morris QD, Hughes TR and Frey BJ (2005). GenXHC: A probabilistic generative model for cross-hybridization compensation in high-density, genome-wide microarray data. *Proceedings of the Thirteenth Annual Conference on Intelligent Systems for Molecular Biology*, June 25-29 2005.
12. Hughes TR *et al.* Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nature Biotechnol.* **19**, pp. 342-347, 2001.
13. John B *et al.* Human MicroRNA targets. *PLoS Biol* **2(11)**, e363, 2004.
14. Jordan MI, Ghahramani Z, Jaakkola TS and Saul LK. An introduction to variational methods for graphical models. *Learning in Graphical Models*, Cambridge: MIT Press, 1999.
15. Kislinger T *et al.* Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling. Submitted to *Cell*, 2005.
16. Krek A *et al.* Combinatorial microRNA target predictions. *Nature Genetics* **37**, pp. 495-500, 2005.
17. Lewis BP, Burge CB and Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, pp.15-20, 2005.
18. Lewis BP *et al.* Prediction of mammalian microRNA targets. *Cell* **115**, pp.787-798, 2003.
19. Lim LP *et al.* The microRNAs of Caenorhabditis elegans. *Genes Dev.* **17**, pp. 991-1008, 2003.
20. Lim LP *et al.* Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* **433**, pp. 769-773, 2005.
21. Lockhart M *et al.* Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnol.* **14**, pp. 1675-1680, 1996.
22. Neal RM and Hinton GE. A view of the EM algorithm that justifies incremental, sparse, and other variants, *Learning in Graphical Models*, Kluwer Academic Publishers, 1998.
23. Xie X *et al.* Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* **434**, pp. 338-345, 2005.
24. Zilberstein CBZ, Ziv-Ukelson M, Pinter RY and Yakhini Z. A high-throughput approach for associating microRNAs with their activity conditions. *Proceedings of the Ninth Annual Conference on Research in Computational Molecular Biology*, May 14-18 2005.
25. Zhang W, Morris Q *et al.* The functional landscape of mouse gene expression. *J Biol* **3**, pp. 21-43, 2004.
26. Supplemental Data for Lewis *et al. Cell* **120**, pp. 15-20. http://web.wi.mit.edu/bartel/pub/Supplemental%20Material/Lewis%20et%20al%202005%20Supp/