Computational prediction of microRNA targets in mouse using proteomics data

Jingjing Li[1,4], Renqiang Min[2], Anthony Bonner[2], Ruth Isserlin[3,4], Andrew Emili[1,3,4] and Zhaolei Zhang[1,3,4]

Departments of Medical Genetics & Microbiology[1], Computer Scieice[2], Banting & Best Department of Medical Research[3], Donnelly Centre for Cellular & Biomolecular Research[4], University of Toronto

Introduction

Sequence complementarity is crucial for the recognition between microRNAs and their target mRNAs, and thus most algorithms that predict microRNA targets are based on sequence complementarily [1,2]. To eliminate false positives, in microRNA-mRNA predictions[3], mRNA and microRNA expression profiles were combined with sequence-based predictions, to filter out reliable microRNA-mRNA interactions; the rationale being that the highly expressed microRNAs may repress the expression of target mRNAs. However, recent experiments have shown that the expression of target mRNAs can vary greatly, due to different mechanisms of post-transcriptional regulation by microRNAs, i.e. either by degradation or translational repression [4]. Thus, using microRNA and mRNA expression data alone is only suitable for identifying targets that are targeted for degradation, while the mRNA targets regulated via translational repression without degradation will not be detected.

Since microRNAs usually suppress protein synthesis of their targeted mRNAs, regardless of the regulatory mechanisms, the *protein abundance* of their targets should always be low, with the mRNA expression possibly indicating how the targets are regulated by the microRNAs. Here, we describe a model to search for reliable microRNA-mRNA interactions, in which highly expressed microRNAs are coupled with gene targets whose protein abundance is low. After identifying a set of high confident interactions, we then trace back to the mRNA expression of each target to investigate the mechanisms of their post-transcriptional regulation (Figure 1).

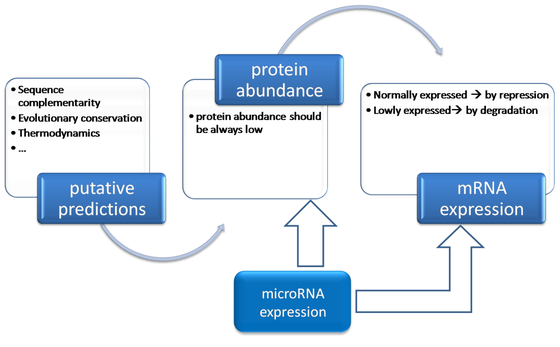

Figure 1. Flowchart depicting the steps in the model

Data compilation

The microRNA expression data were from [5], mRNA expression data were from [6], and proteomics data were from [7]. Among 4,768 mouse proteins, 1,677 were mapped to the non-redundant transcripts in [7]. We started with a set of 21,712 putative microRNA-gene predictions from TargetScan [2], invloving 1,404 transcripts/proteins and 75 microRNAs. Our goal is to select a subset of highly confident predictions with low protein abundance and high microRNA expression. Since we have little prior knowledge about the true microRNA-gene targets, we built a probabilistic model to rank the putative microRNA-gene predictions.

Model Construction

To identify reliable microRNA-mRNA interactions, we used the negative binomial distribution to model the proteomics data, and used binary latent indicator variables $\delta_{i,j}$ to represent whether or not microRNA i regulates mRNA j, and then we used Bayesian negative binomial regression to associate mRNA protein abundance with microRNA expression.

Having first identified microRNA-mRNA interactions with the highest confidence, we then further investigated their probability of being regulated via degradation. Conditioned on $\delta_{i,j}=1$, we further introduced a new binary latent variable $b_{i,j}$ to indicate whether or not microRNA i regulates mRNA j through degradation. Then we used Bayesian logistic regression to regress the mRNA low expression probabilities based on their regulating microRNA expressions. So that $b_{i,j}$'s between the mRNAs with low expression and the microRNAs with high expression are more likely to be set to 1, implying degradation; and $b_{i,j}$'s between the mRNAs with high expression and microRNAs with high expression are more likely to be set to 0, implying translational repression (Figure 2). We used Gibbs sampling to make inference from the posterior distributions after 15,000 iterations.

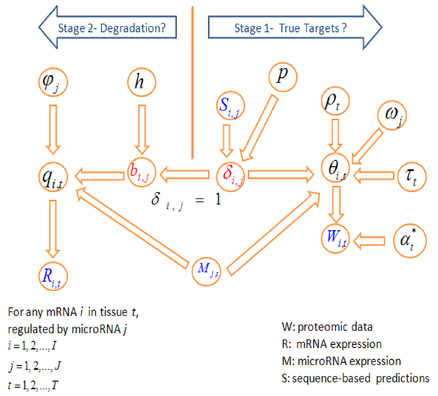

Figure 2. Graphic representation of the model. Blue nodes represent observed data, red nodes represent latent variables we are interested in, and black nodes are the parameters of the model.

Blind Tests

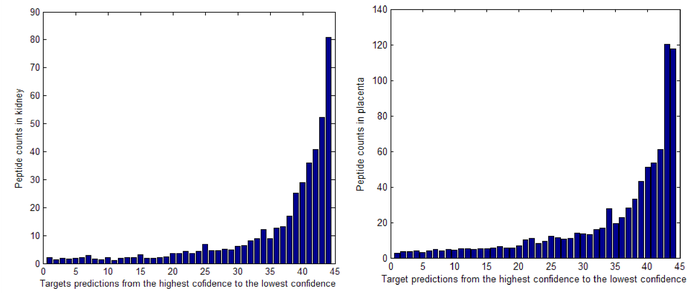

Figure 4. Blind tests on proteomics data in kidney and placenta

We applied our predictions to the proteomics data in kidney and placenta, which are not used for training the model, similarly, top ranked predictions also have the lowest protein abundance

Model Checking

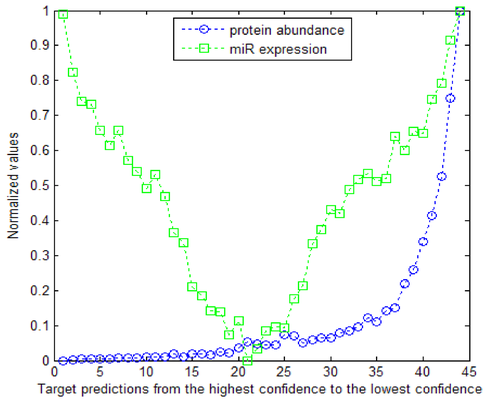

Figure 3. As a general trend, the microRNA-gene predictions with high confidence usually have low protein abundance and high miR expression

For all the 21,712 putative microRNA-gene predictions, our model assigned confidence scores (Figure 3). Since the confidence scores only have comparative meaning, we then grouped the predictions into 44 bins, and each bin consisting of 500 ranked interactions. A decrease in microRNA-gene prediction confidence had a corresponding increase in protein abundance, which demonstrates the effectiveness of our model.

The top ranked predictions have higher posterior probability to be true interactions as they all have low protein abundance and high microRNA expression.. The intermediate confident predictions have relatively low protein abundance and low microRNA expression, and the least confident predictions have high protein abundance and high microRNA expression.

Degradation or Repression ?

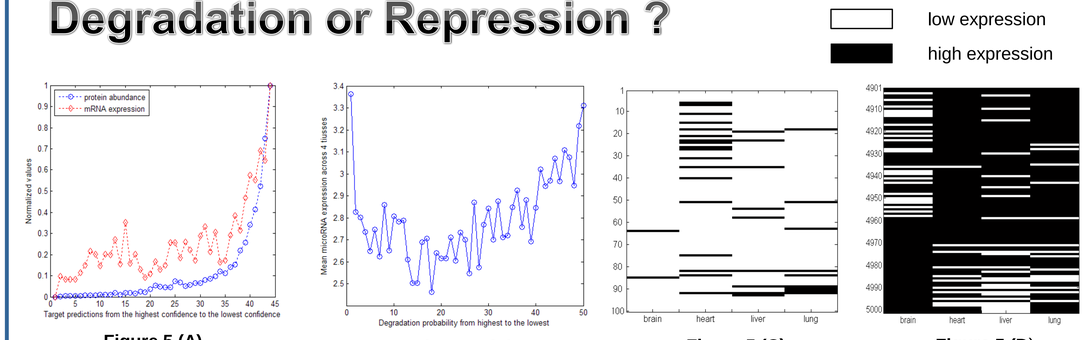

Figure 5 (A) **Figure 5 (B)** **Figure 5 (C)** **Figure 5 (D)**

- As shown in Figure 5(A), the mRNA expression for the top ranked predictions fluctuates greatly, implying microRNAs may regulate their targets through different mechanisms.
- We selected the top ranked 5,000 microRNA-gene interactions as reliable predictions, and then traced back their mRNA expression profiles.
- After Bayesian learning, we assigned the degradation confidence score to each microRNA-mRNA interaction, and grouped the ranked predictions into 50 bins.
- As shown in Figure 5(B,C), the top ranked interactions have the highest microRNA expression and the lowest mRNA expression, suggesting those microRNAs be more likely to regulate the mRNAs through degradation.
- As shown in Figure 5(B,D), the microRNA expression of the bottom ranked interactions is high and their targets are also highly expressed, suggesting those microRNAs be more likely to regulate the mRNAs through translational repression.

References

1. John, B. et al. PLoS Biol. 2, e363 (2004).
2. Lewis, B.P., Shih, I.H., Jones-Rhoades, M.W., Bartel, D.P. and Burge, C.B. Cell 115, 787-798 (2003)
3. Huang, J.C., Morris, Q.D., and Frey, B.J. J. Comp. Biol. 14(5), 550-563 (2007)
4. He, L. and Hannon, G. Nature Rev. Genet. 5, 522−531
5. Babak, T. et al. RNA 10,1813-9.(2004)
6. Zhang, W. et al. J. Biol 3, 21(2004)
7. Kislinger, T. et al. Cell 125(1), 173-86 (2006)