Pacific Symposium on Biocomputing 13:52-63(2008)

COMPARING SEQUENCE AND EXPRESSION FOR PREDICTING microRNA TARGETS USING GenMiR3

J. C. HUANG[1], B. J. FREY[1,2] AND Q. D. MORRIS[2]

[1]*Probabilistic and Statistical Inference Group, University of Toronto,*
10 King's College Rd.,
Toronto, ON, M5S 3G4, Canada
E-mail: jim,frey@psi.toronto.edu

[2]*Banting and Best Department of Medical Research, University of Toronto,*
160 College Street,
Toronto, ON, M5S 1E3, Canada
E-mail: quaid.morris@utoronto.ca

We present a new model and learning algorithm, GenMiR3, which takes into account mRNA sequence features in addition to paired mRNA and miRNA expression profiles when scoring candidate miRNA-mRNA interactions. We evaluate three candidate sequence features for predicting miRNA targets by assessing the expression support for the predictions of each feature and the consistency of Gene Ontology Biological Process annotation of their target sets. We consider as sequence features the total energy of hybridization between the microRNA and target, conservation of the target site and the context score which is a composite of five individual sequence features. We demonstrate that only the total energy of hybridization is predictive of paired miRNA and mRNA expression data and Gene Ontology enrichment but this feature adds little to the total accuracy of GenMiR3 predictions using for expression features alone.

1. Introduction

Recent research into understanding gene regulation has shed light on the significant role of microRNAs (miRNAs). These small regulatory RNAs suppress protein synthesis[1] or promote the degradation[2] of specific transcripts that contain anti-sense target sequences to which the miRNAs can hybridize with complete or partial complementarity. The catalogue of putative microRNA-target interactions predicted on the basis of genomic sequence continues to grow[3,4,5], but the most accurate computational approaches rely on the presence of a highly conserved seed in the putative target, greatly reducing their sensitivity[6]. However, even these highly selective methods appear to have low specificity[3]. Expression profiling has been proposed as a complementary method for discovering miRNA targets[7], but this can become intractable and costly when multiple miRNAs and their effects across multiple tissues must be considered.

We have recently described a probabilistic method, GenMiR++ (**Gen**erative model for **miR**NA regulation)[8,9], which incorporates miRNA and mRNA expression data with a set of candidate miRNA-target interactions to greatly improve the precision in predicting functional miRNA-target interactions. While our method was shown to be robust[8] and to improve predictive accuracy[9] according to several independent measures, it does not consider sequence-specific features of miRNA target sites beyond the presence of a highly conserved miRNA seed. Recently it has been reported that many sequence features such as secondary structure[10] or the relative positioning of sites within the target mRNA's 3'UTR[11] may play a crucial role in miRNA target recognition. We therefore set out to evaluate whether such sequence features could increase the predictive power of our model for miRNA regulation.

In this paper, we present GenMiR3, a generative model of miRNA regulation which uses sequences features to establish a prior probability of a miRNA-target interaction being functional and then uses paired expression data for miRNAs and mRNAs to compute the likelihood of a putative miRNA-target interactions. By combining these two sources of information together to compute a posterior probability of a miRNA-target relationship being functional, we score candidate miRNA-target interactions in terms of both expression support and sequence features. We evaluate several candidate sequence features by computing their predictions with the expression data and by comparing the Gene Ontology enrichment of target sets obtained using sequence and/or expression features. We then determine whether these features could be used in tandem with expression data to improve the accuracy of our miRNA target predictions.

2. The GenMiR3 model and learning algorithm

GenMiR3 makes two significant improvements over our previous model GenMiR++ [8,9]: we use sequence features to establish a prior on whether a given miRNA will bind to a target site in the 3'UTR and we use a different prior on many model parameters to give more flexibility in our posterior probability estimates. We first describe the changes to our generative model of mRNA expression and then describe how we propose to integrate sequence features

2.1. *A Bayesian model for gene and microRNA expression*

GenMiR3 is a generative model of mRNA expression levels that computes the expression support for a putative miRNA-mRNA by evaluating the degree to which the miRNA expression levels could explain the observed mRNA expression levels given of all other predicted regulators for that mRNA. Given two expression data sets profiling G mRNA transcripts and K miRNAs across T tissues, we denote by $\mathbf{x}_g = (x_{g1}x_{g2}\cdots x_{gT})^{\mathsf{T}}$ and $\mathbf{z}_k = (z_{k1}z_{k2}\cdots z_{kT})^{\mathsf{T}}$ the expression profiles over the T tissues for mRNA transcript g and miRNA k respectively. Here x_{gt} refers to the expression of the g^{th} transcript in the t^{th} tissue and z_{kt} refers to the expression of the k^{th} miRNA in the same tissue.

Our model also takes as input a set of candidate miRNA-target interactions in the form of a binary matrix $\mathbf{C}$, where $c_{gk} = 1$ if transcript g is a candidate target of miRNA k and $c_{gk} = 0$ otherwise. For each (g, k) pair for which $c_{gk} = 1$, we also introduce an indicator variable s_{gk}. In our model, $s_{gk} = 1$ indicates that the candidate interaction between (g, k) is truly functional. Thus, the problem of scoring putative miRNA-target interactions can be formulated as calculating a posterior probability of $s_{gk} = 1$ given $c_{gk} = 1$.

To complete the formulation of our generative model, we introduce a set of nuisance parameters $\mathbf{\Lambda} = \{\lambda_k\}$ that each scale the regulatory effect of a given miRNA and $\mathbf{\Gamma} = \mathrm{diag}(\gamma_1, \cdots, \gamma_T)$ to account for normalization differences between the miRNA and mRNA expression levels in tissue t. We assign prior distributions $P(\mathbf{\Lambda}|\boldsymbol{\alpha})$ and $P(\mathbf{\Gamma}|\boldsymbol{\alpha})$ and we integrate over these distributions when making predictions. Having defined the above parameters and variables, we can write the probabilities of the mRNA expression profiles $\mathbf{X} = \{\mathbf{x}_g\}$ conditioned on the expression profiles of miRNAs $\mathbf{Z} = \{\mathbf{z}_k\}$, and a set of functional miRNA-target interactions, $\mathbf{S} = \{s_{gk}\}$, as

$$P(\mathbf{x}_g|\mathbf{Z}, \mathbf{S}, \mathbf{\Gamma}, \mathbf{\Lambda}, \mathbf{\Theta}) = N(\mathbf{x}_g; \boldsymbol{\mu} - \sum_k \lambda_k s_{gk} \mathbf{\Gamma} \mathbf{z}_k, \mathbf{\Sigma}) \tag{1}$$

$$P(\mathbf{\Gamma}|\boldsymbol{\alpha}) = \prod_{t=1}^{T} P(\gamma_t|m, n) = \prod_{t=1}^{T} Gamma(\gamma_t; m, n) \tag{2}$$

$$P(\mathbf{\Lambda}|\boldsymbol{\alpha}) = \prod_{k=1}^{K} P(\lambda_k|a, b) = \prod_{k=1}^{K} Gamma(\lambda_k; a, b) \tag{3}$$

where $\boldsymbol{\mu}$ is a background transcriptional rate vector and $\mathbf{\Sigma}$ is a data noise covariance matrix. Note that in the above model, we use a point-estimate of $\mathbf{\Theta} = \{\boldsymbol{\mu}, \mathbf{\Sigma}\}$. The set $\boldsymbol{\alpha} = \{a, b, m, n\}$ corresponds to fixed hyperparameters which characterize the prior distributions on the parameters $\mathbf{\Gamma}, \mathbf{\Lambda}$. In the above model, we represent the expression profile of a given mRNA transcript g as being negatively regulated by all candidate miRNAs for which $s_{gk} = 1$.

2.2. *Incorporating sequence features*

To include sequence features of the miRNA target site in the model, we introduce an N-dimensional vector $\mathbf{f}_{gk} = (f^1_{gk} f^2_{gk} \cdots f^N_{gk})$ containing a description of N sequence features associated with the miRNA-mRNA pair (g, k). We denote by $\pi_{gk} = P(s_{gk} = 1 | c_{gk} = 1, \mathbf{f}_{gk}, \mathbf{w})$ the prior probability that indicator variable $s_{gk} = 1$ given the sequence features. As a simplifying assumption, we will assume that each of the N sequence features independently contribute to π_{gk} with weight equal to $w_n, n = 1, \cdots, N$. We will also assume that the s_{gk} variables are *a priori* independent of one another. This yields

$$\begin{aligned} P(\mathbf{S}|\mathbf{C}, \mathbf{F}, \mathbf{w}) &= \prod_{(g,k)} p(s_{gk}|\mathbf{C}, \mathbf{F}, \mathbf{w}) \\ &= \prod_{(g,k)|c_{gk}=0} [s_{gk} = 0] \prod_{(g,k)|c_{gk}=1} \pi_{gk}^{s_{gk}} (1 - \pi_{gk})^{1-s_{gk}} \qquad (4) \\ \pi_{gk} = P(s_{gk} = 1 | c_{gk} = 1, \mathbf{f}_{gk}, \mathbf{w}) &= \frac{1}{1 + \exp(-\mathbf{w}^{\mathsf{T}} \mathbf{f}_{gk})} \qquad (5) \end{aligned}$$

where $[H] = 1$ if H is true, otherwise $[H] = 0$.

Given the above, we can write the probabilities in our model, conditioned on the expression of miRNAs and a set of candidate miRNA targets, as

$$\begin{aligned} P(\mathbf{X}, \mathbf{S}, \mathbf{\Gamma}, \mathbf{\Lambda}|\mathbf{C}, \mathbf{F}, \mathbf{Z}, \mathbf{\Theta}, \mathbf{w}, \boldsymbol{\alpha}) = &P(\mathbf{S}|\mathbf{C}, \mathbf{F}, \mathbf{w}) P(\mathbf{\Gamma}|\boldsymbol{\alpha}) P(\mathbf{\Lambda}|\boldsymbol{\alpha}) \\ &\prod_g P(\mathbf{x}_g|\mathbf{Z}, \mathbf{S}, \mathbf{\Gamma}, \mathbf{\Lambda}, \mathbf{\Theta}) \qquad (6) \end{aligned}$$

Because we have formulated our model in a Bayesian framework, we can marginalize out our nuisance parameters when calculating the likelihood of the mRNA expression data or when calculating the posterior probabilities of $s_{gk} = 1$, *e.g.*,

$$P(\mathbf{X}|\mathbf{C}, \mathbf{F}, \mathbf{Z}, \mathbf{\Theta}, \mathbf{w}, \boldsymbol{\alpha}) = \sum_{\mathbf{S}} \int_{\mathbf{\Gamma}} \int_{\mathbf{\Lambda}} P(\mathbf{X}, \mathbf{S}, \mathbf{\Gamma}, \mathbf{\Lambda}|\mathbf{C}, \mathbf{F}, \mathbf{Z}, \mathbf{\Theta}, \mathbf{w}, \boldsymbol{\alpha}) \, d\mathbf{\Lambda} \, d\mathbf{\Gamma} \qquad (7)$$

Figure 1 shows the Bayesian network for our model of miRNA regulation. Under our model, each transcript g in the network is associated with a

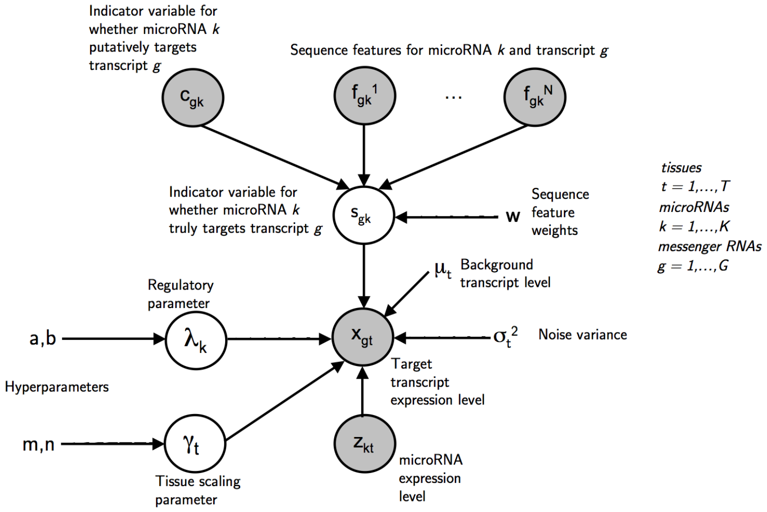

Indicator variable for whether microRNA *k* putatively targets transcript *g*
Sequence features for microRNA *k* and transcript *g*
c_{gk}
f_{gk}^1
...
f_{gk}^N
Indicator variable for whether microRNA *k* truly targets transcript *g*
s_{gk}
w
Sequence feature weights
tissues
t = 1,...,T
microRNAs
k = 1,...,K
messenger RNAs
g = 1,...,G
μ_t
Background transcript level
Regulatory parameter
a,b
λ_k
x_{gt}
σ_t^2
Noise variance
Target transcript expression level
Hyperparameters
m,n
γ_t
z_{kt}
Tissue scaling parameter
microRNA expression level

Figure 1. Bayesian network used for modelling microRNA regulation using both sequence and expression features. Nodes correspond to observed and unobserved variables as well as model parameters, with directed edges between nodes representing conditional dependencies encoded by our probability model. Each variable node and all incoming/outgoing edges associated with that node are replicated a number of times according to the number of such variables in the model. Shaded nodes correspond to observed variables and unshaded ones are unobserved. Model parameters which are estimated in a pointwise fashion are shown without nodes.

set of indicator variables $\{s_{gk'}\}, k' \in \{k|c_{gk} = 1\}$ which indicate which of its candidate miRNA regulators affect its expression level. The posterior probabilities over these variables are the predictions of the model: these posteriors are determined by combining priors over s_{gk} which are determined by examining the sequence of transcript g and miRNA k in addition to support from the expression data through our inference and learning procedure. We describe our learning method in the next section.

2.3. *Learning the model of gene and microRNA expression*

Exact Bayesian learning of our model is intractable, so we use a variational method[12,13] to derive a tractable approximation. Our learning procedure is similar to that for GenMiR++[8,9]. Here we will describe only the changes and refer the reader to our previous work[8] for the rest of the derivation. In particular, we specify the Q-distribution via a mean-field factorization with $Q(\mathbf{S}, \mathbf{\Lambda}, \mathbf{\Gamma}|\mathbf{C}) = Q(\mathbf{S}|\mathbf{C})Q(\mathbf{\Lambda})Q(\mathbf{\Gamma})$ such that

$$\begin{aligned} Q(\mathbf{S}|\mathbf{C}) &= \prod_{g,k} Q(s_{gk}|\mathbf{C}) = \prod_{(g,k)|c_{gk}=0} [s_{gk} = 0] \prod_{(g,k)|c_{gk=1}} \beta_{gk}^{s_{gk}} (1-\beta_{gk})^{1-s_{gk}} \\ Q(\mathbf{\Lambda}) &= \prod_{k=1}^{K} Q(\lambda_k) = \prod_{k=1}^{K} Gamma(\lambda_k; a_k, b_k) \\ Q(\mathbf{\Gamma}) &= \prod_{t=1}^{T} Q(\gamma_t) = \prod_{t=1}^{T} Gamma(\gamma_t; m_t, n_t) \end{aligned} \tag{8}$$

where β_{gk} is the approximate posterior probability that a given miRNA-target pair (g, k) is functional given the data. Using this Q-distribution, we iteratively minimize the upper bound $L(Q)$ on the negative data likelihood with respect to the distribution over unobserved variables $Q(\mathbf{S}|\mathbf{C})$ (variational Bayes E-step), the distribution over model parameters $Q(\mathbf{\Gamma})Q(\mathbf{\Lambda})$ (variational Bayes M-step) and with respect to the regular model parameters.

2.4. *Setting the sequence-based priors using the posteriors from the gene and microRNA expression model*

The prior probability $\pi_{gk} = p(s_{gk} = 1|c_{gk} = 1, \mathbf{f}_{gk}, \mathbf{w})$ is parametrized by the weight vector $\mathbf{w}$. We estimate this weight vector by maximizing the expected log-likelihood $E_Q[\log p(\mathbf{S}|\mathbf{C}, \mathbf{F}, \mathbf{w})]$ of the s_{gk} variables. This then reduces to a standard logistic regression problem, with each output label set to β_{gk}, or the expected value of s_{gk} under $Q(\mathbf{S})$. We can perform the required optimization via a conjugate-gradient method, with the gradient $\nabla_{\mathbf{w}} E_Q[\log p(\mathbf{S}|\mathbf{C}, \mathbf{F}, \mathbf{w})]$ and the Hessian $\nabla\nabla_{\mathbf{w}} E_Q[\log p(\mathbf{S}|\mathbf{C}, \mathbf{F}, \mathbf{w})]$ given by

$$\nabla_{\mathbf{w}} E_Q[\log p(\mathbf{S}|\mathbf{C}, \mathbf{F}, \mathbf{w})] = \sum_{(g,k)|c_{gk}=1} \mathbf{f}_{gk}(\beta_{gk} - \pi_{gk}) \tag{9}$$

$$\nabla\nabla_{\mathbf{w}} E_Q[\log p(\mathbf{S}|\mathbf{C}, \mathbf{F}, \mathbf{w})] = \sum_{(g,k)|c_{gk}=1} \mathbf{f}_{gk}\mathbf{f}_{gk}^{\mathsf{T}} \pi_{gk}(1 - \pi_{gk}) \tag{10}$$

We iteratively run the variational Bayes algorithm to estimate the approximate posterior probabilities β_{gk} and then update the weight vector $\mathbf{w}$ until convergence to a minimum of $L(Q)$. We can then assign a score to each candidate miRNA-target interaction using the log posterior odds $\log \frac{\beta_{gk}}{1-\beta_{gk}}$ so that a higher score reflects a higher posterior probability of a miRNA-target pair (g, k) being functional.

3. Results

To assess the impact of including sequence features, we downloaded human miRNA and mRNA expression data generated by [14,15] in addition to the set of TargetScanS [3] candidate human miRNA target-interactions from the UCSC Genome Browser[16] (build hg17/NCBI35) and mapped these interactions to the expression data. This yielded 6,387 candidate miRNA-target interactions between 114 human miRNAs and 890 mRNA transcripts, with patterns of expression across 88 human tissue samples. We then learned the GenMiR3 model without the sequence prior and once the algorithm converged, we selected the 100 highest and 100 lowest-scoring miRNA-target interactions and we downloaded the corresponding 3'UTR genomic sequences for each of the corresponding targeted mRNAs from the UCSC Genome Browser. The score assigned by GenMiR3 in the absence of sequence features predicts whether a given candidate miRNA-target interaction is functional based on joint patterns of expression of miRNAs and their target mRNAs across multiple tissues/cell types. We have previously shown that a similar score can distinguish functional and non-functional candidate miRNA/mRNA target pairs[9]. Here we use this "expression-only" GenMiR3 score to compare predictions made using both sequence and expression features with those made based solely on expression data. Once we evaluate the sequence features alone, we use a Gene Ontology enrichment test to evaluate the effect of combining these features with expression data using the full GenMiR3 model.

3.1. *Evaluating sequence features using cross-validation*

We evaluate three different sequence features: the *total hybridization energy*[10], a measure of the free energy of binding of the miRNA to its candidate target site that also considers any RNA secondary structure that the target site may participate in; the *context score*[11], an aggregate score combining the AU content with ± 30 bp of each miRNA target site, proximity to residues pairing to sites for coexpressed miRNAs, proximity to residues pairing to miRNA nucleotides 13-16, positioning of sites within the 3'UTR at least 15 nt from the stop codon and positioning away of sites from the center of the 3'UTR; and the *PhastCons score*, which is a measure of the conservation of the whole target site basefd on the PhastCons algorithm [17].

We calculated the total hybridization energy ΔG_{total} using a procedure related to [10]. Briefly, we set $\Delta G_{total} = \Delta G_{hybrid} - \Delta G_{disrupt}$, where ΔG_{hybrid} is the the total hybridization energy between a miRNA and its target mRNA computed by aligning the miRNA and target sequences and evaluating the total energy of hybridization using standard energy parameters. The expected disruption energy $< \Delta G_{disrupt} >$ was then obtained by using first calculating the probability that each base in the target site was paired with another base in the 3'UTR using RNAfold[18] and then using these base pair probabilities to calculate the expected hybridization energy of the target site in absence of the miRNA. If there was more than one possible site for a given miRNA in the 3'UTR, we summed ΔG_{total} over all sites.

We then downloaded the context scores from the TargetScan 4.0 website [11] and we calculated the PhastCons score by summing all of log-probabilities of conservation (obtained from the UCSC Genome Browser) over all base positions of all sites with seed matches to the mature miRNA in the target mRNA's 3'UTR. We then normalized each of these three features to be zero mean and unit variance.

We randomly split the above set of 200 high/low-scoring miRNA-target interactions under the expression-only GenMiR3 model into 1000 training and test sets of size 150/50 respectively. For each sequence feature, we trained two logistic regression models for each of the training sets: one with the feature included and a null model with the feature excluded. We evaluated the test likelihood given the learned weights and computed the average likelihood ratio between the test likelihood $L_{feature}$ for each feature and the likelihood of the null model L_{null} with no features. The median and standard deviations of the test likelihood ratios over the 1000 training/test splits are shown in Figure 2. The ΔG_{total} score is most predictive of the three queried features, as including it in the model tends to increase the median test likelihood with respect to the null model. Neither the Phast-

Feature	$Median\left(\frac{L_{feature}}{L_{null}}\right)$	s.d.
ΔG_{total}[10]	2.3597	4.0859
PhastCons[17]	1.1262	1.9774
Context[11]	1.2129	2.0497

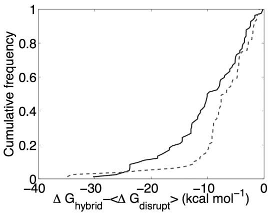


Figure 2. Sequence features and median test likelihood ratios computed over 1000 test/train splits; the total hybridization energy ΔG_{total} between a miRNA and its target mRNA transcript is shown for high GenMiR3-scoring targets (solid) and low GenMiR3-scoring targets (dashed)

Cons score nor the context score increased the median test likelihood with respect to the null model. We also found that the individual features used to compute the context score (such as AU content around the target site) did not increase the test likelihood with respect to the null model, nor was there a significant difference in these median feature values between high- and low-scoring GenMiR3 targets (data not shown). For the ΔG_{total} score however, we found that the high-scoring GenMiR3 miRNA-target interactions indeed have a lower median ΔG_{total} score than low-scoring GenMiR3 candidates (p = 0.0138, Wilcoxon-Mann-Whitney (WMW) test; Figure 2).

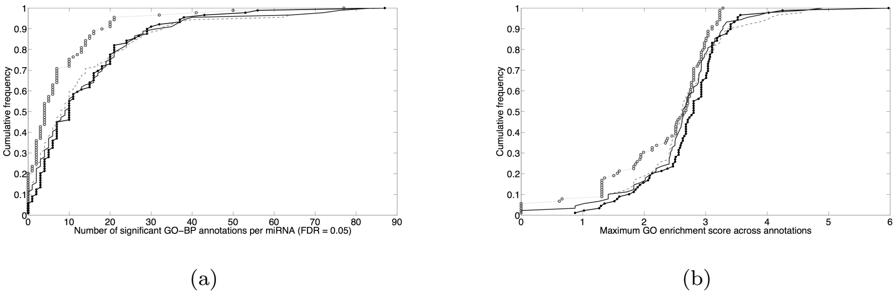
3.2. *Evaluating sequence features using functional enrichment analysis*

We have also previously shown that predicted target sets of many microRNAs are enriched for Gene Ontology Biological Process (GO-BP) categories [9]. As such, we reasoned that more accurate target predictions should show higher levels of GO-BP enrichment and we used GO-BP enrichment to assess target prediction accuracy. To calculate the different sequence features, we downloaded 3'UTR sequences for each of the mRNAs putatively targeted by a miRNA and filtered out all 3'UTR's with length greater than 5,000 bp and those without a published context score. This process yielded 410 candidate miRNA-target interactions between 89 human miRNAs and 150 mRNA transcripts. We then computed ΔG_{total} for each of these 410 candidate miRNA-target interactions and trained GenMiR3 on the expression data and ΔG_{total} as a sequence feature.

To compute GO-BP enrichment, we downloaded human GO-BP annotations from BioMart[19]. After up-propagation, we had a total of 13,003 functional annotations of which we removed annotations which were associated with less than 5 annotated Ensembl genes, leaving us with 2,021 GO-BP annotations. To establish the target sets, we selected the top 25% of candidate miRNA-target interactions for each miRNA under four scoring schemes:

(1) GenMiR3 score obtained from expression features alone
(2) GenMiR3 score obtained from both ΔG_{total} and expression features
(3) ΔG_{total} alone
(4) Context score

We computed enrichment by using Fisher‘s exact test to measure the statistical significance of the overlap between each GO-BP category and predicted target set of each of the 89 miRNAs in our data set (for a total of 179,869 enrichment scores). For each miRNA, we used these p-values to compute the number of significantly enriched categories (FDR < 0.05, linear step-up[20]), shown in Figure 3(a) and the maximum $-\log_{10}$ p-value across the GO-BP categories, shown in Figure 3(b). As can be seen, selecting miRNA targets on the basis of either expression alone, ΔG_{total} alone, or both, yields a higher number of enriched GO categories than selecting on the basis of the context score alone ($p = 8.2016 \times 10^{-4}, p = 2.7903 \times 10^{-5}, p = 0.0049$, respectively, Wilcoxon-Mann-Whitney). Our results also indicate, however, that adding the ΔG_{total} sequence feature to the model for expression does not significantly improve the GO enrichment GenMiR3 target sets. We will discuss possible reasons for this in the last section.



(a)

(b)

Figure 3. Cumulative frequency plots of a) Number of significant GO categories per miRNA at FDR= 0.05 and b) maximum GO enrichment scores per miRNA obtained from using the GenMiR3 score obtained from expression features alone (solid), using the GenMiR3 score obtained from both ΔG_{total} and expression features (dashed), ΔG_{total} alone (star) and the context score (circle).

4. Discussion and conclusion

In this paper we have proposed the GenMiR3 probabilistic model for miRNA regulation using both sequence and expression features. We examined three sequence features: the total energy of hybridization ΔG_{total} between the microRNA and target, conservation of the target site and the context score, which itself is an aggregate score based on five sequence features. Using cross-validation, we found that the ΔG_{total} sequence feature was the best predictor of GenMiR3 score computed from expression features alone. Using a functional enrichment analysis, we found that selecting miRNA targets based on GenMiR3 score (with and without ΔG_{total}) and the ΔG_{total} score alone yielded a significantly higher number of enriched GO categories than selecting on the basis of the context score.

The relative performance of the context score[11] compared to the total hybridization score[10] was particularly surprising. Many of the features included in the context score should be predictive of whether or not the target site is likely to be single-stranded or double-stranded prior to miRNA binding, whereas the total hybridization score is a more direct indicator of this state. The results of our tests therefore suggest that single-strandedness of the miRNA target site is the most accurate sequence feature for predicting binding.

There are a number of possible explanations for the fact that adding the ΔG_{total} sequence feature to the model for expression does not improve the enrichment of GenMiR3 target sets. It is unlikely that the expression features are redundant with ΔG_{total}, as ΔG_{total} and expression-only GenMiR3 scores cease to be correlated outside of the 100 highest and lowest scoring interactions under GenMiR3 ($\rho = -0.0696, p = 0.1595$, Spearman correlation), suggesting ΔG_{total} and the expression data are making different predictions about miRNA targets. It is unclear whether ΔG_{total} or GenMiR3 are making better predictions, as we may have reached the limitation of the power of the GO analysis and require a more sensitive test. The expression signal does appear to be quite strong though because, when added to the GenMiR3 model, ΔG_{total} does not change GenMiR3 predictions: the Spearman correlation is 0.99 between the expression-only GenMiR3 posteriors and the posteriors in the GenMiR3 which also accounts for sequence data. This suggests that when expression data is limited or unavailable, the ΔG_{total} sequence prior will be a very useful addition to the GenMiR3 model, in addition to being predictive of functionality in its own right.

References

1. Ambros, V. (2004) The functions of animal microRNAs. *Nature* **431**, 350-355.
2. Bagga, S. *et al* (2005) Regulation by let-7 and lin-4 miRNAs results in target mRNA degradation. *Cell* **122**, 553-63.
3. Lewis, B.P., Burge, C.B., and Bartel, D.P. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15-20.
4. Krek, A. *et al* (2005) Combinatorial microRNA target predictions. *Nat. Gen.* **37**, 495-500.
5. Huynh, T. *et al* (2006) A pattern-based method for the identification of microRNA-target sites and their corresponding RNA/RNA complexes. *Cell* **126**, 1203-1217.
6. Sood, P. *et al* (2006) Cell-type-specific signatures of microRNAs on target mRNA expression. *Proceedings of the National Academy of Sciences (PNAS)* **103**, 2746-2751.

7. Lim, L.P. *et al* (2005) Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* **433**, 769-773.
8. Huang, J.C., Morris, Q.D., and Frey, B.J. (2007) Bayesian learning of microRNA targets using sequence and expression data. *J. Comp. Bio.* **14**(5), 550-563.
9. Huang, J.C., *et al.* (2007) Using expression profiling to identify human microRNA targets. *In press.*
10. Long, D. *et al.* (2007) Potent effect of target structure on microRNA function. *Nat. Struct. Mol. Bio.* **14**, 287-294.
11. Grimson, A. it et al (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell* **27**, 91-105.
12. Attias, H. (1999) Inferring parameters and structure of latent variable models by variational Bayes. *Proceedings of the 15th Conference on Uncertainty in Artifical Intelligence*, 21-30.
13. Neal, R.M., and Hinton, G.E. (1998) A view of the EM algorithm that justifies incremental, sparse, and other variants, 355-368. *In* Jordan, M.I., ed., *Learning in Graphical Models*, MIT Press.
14. Lu, J. *et al.* (2005) MicroRNA expression profiles classify human cancers. *Nature* **435**, 834-8.
15. Ramaswamy, S. *et al.* (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences (PNAS)* **98**, 15149-15154.
16. Karolchik, D. *et al* (2003) The UCSC Genome Browser Database. *Nucl. Acids Res.* **31**(1), 51-54.
17. Siepel, A., *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034-1050.
18. Hofacker, I. (2003) Vienna RNA secondary structure server. *Nucl. Acids Res.* **31**(13), 3429-3431.
19. Durinck, S. (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**(16), 3439-40.
20. Benjamini, Y., and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol* **57**, 289-300.