Discovery of miR-mRNA interactions via simultaneous Bayesian inference of gene networks and clusters using sequence-based predictions and expression data

Brian Godsey[1,*]

[1]Department of Statistics and Probability Theory, Vienna University of Technology, 1040 Vienna, Austria

Summary

MicroRNAs (miRs) are known to interfere with mRNA expression, and much work has been put into predicting and inferring miR-mRNA interactions. Both sequence-based interaction predictions as well as interaction inference based on expression data have been proven somewhat successful; furthermore, models that combine the two methods have had even more success. In this paper, I further refine and enrich the methods of miR-mRNA interaction discovery by integrating a Bayesian clustering algorithm into a model of prediction-enhanced miR-mRNA target inference, creating an algorithm called *PEACOAT*, which is written in the *R* language. I show that *PEACOAT* improves the inference of miR-mRNA target interactions using both simulated data and a data set of microarrays from samples of multiple myeloma patients. In simulated networks of 25 miRs and mRNAs, our methods using clustering can improve inference in roughly two-thirds of cases, and in the multiple myeloma data set, KEGG pathway enrichment was found to be more significant with clustering than without. Our findings are consistent with previous work in clustering of non-miR genetic networks and indicate that there could be a significant advantage to clustering of miR and mRNA expression data as a part of interaction inference.

1 Introduction

It is known that microRNAs (miRs) can interfere with mRNA expression, potentially regulating a large number of critical biological processes; for comprehensive information and specific examples of ways that miRs affect biological processes, visit `mirbase.org` [1, 2, 3]. Though many miR-mRNA interactions have been investigated and validated, the vast majority of such interactions is yet undiscovered [4]. Therefore, considering the potential importance and lack of knowledge about miR interference, much effort has been put into both predicting interactions based on the short ($\approx$22nt) sequence length of miRs as well as inferring interactions from expression data; an overview can be found in [5]. Though the former approach addresses issues particular to short sequences, the latter approach is closely related to other interaction models, particularly genetic interaction models, for which there exists even more research and literature than for miR-mRNA interaction models. A recent review enumerates many state-of-the-art methods for genetic interaction modeling: [6]. The primary difference between miR-mRNA interaction models and the canonical genetic interaction models is that, in genetic models, the set of potential regulators and the set of potential regulatees are one and the same, whereas in

*To whom correspondence should be addressed. Email: briangodsey@gmail.com

miR-mRNA models it is assumed that some miRs regulate some mRNAs, and no other interactions exist. Despite this notable difference, when designing models of miR-mRNA interactions, there is a lot to be learned from genetic interaction models of varying types.

Two prominent miR-mRNA interaction models, *TaLasso* [7] and *GenMir* [8], utilize Lasso regression and Bayesian networks, respectively, both of which are also used in genetic interaction models [9, 10, 11, 12]. One popular method of inferring genetic interactions, the Dynamic Bayesian Network (DBN), applies only to time-series expression data, but has proven to be quite useful in inferring interactions [9, 10]. It has been shown in [13] that aspects of a DBN can be applied to miR-mRNA interaction models if paired expression data (i.e. miR and mRNA expression data from the same biological samples or groups) are available, given a partial ordering of samples or groups. The requirement of a partial ordering is much more flexible than in a traditional DBN, which requires a total ordering, and in addition often requires the size of each step (i.e. time elapsed) between stages to be equal. The model from [13] requires neither total ordering nor equal step size, a flexibility which is enabled by the mutual exclusivity of the regulator set from the regulatee set.

Another notable notion from genetic interaction models that has yet to be fully utilized in miR-mRNA interaction models is the idea of a regulatory cluster, and likewise a regulated cluster. Though miR clusters and clusters of miR-regulated mRNAs have indeed been investigated [14, 15], the idea has not been incorporated into probabilistic models using paired expression data.

I have recently shown that, in genetic DBN models, clustering of genes by their expression profiles improves interaction inference by reducing the interaction parameter space as well as the uncertainty arising from highly correlated potential regulators [16]. To summarize, it is difficult to determine the true regulator if two (or more) potential regulators are highly correlated, and this high inferential uncertainty can cause both potential regulators (together with their common regulatee) to fall far down the list of top-ranked inferred interactions; thus, it is better to group highly correlated regulators together and allow all members of the group to maintain a top ranking than to allow competition to diminish the inferred contribution of all members. In this paper, I expand upon this idea and adapt it for miR-mRNA interaction models. More specifically, clustering components are included in a new, updated version of the sequential miR-mRNA interaction model presented in [13]. The resulting model and algorithm are called *PEACOAT*: a Prediction-Enhanced Algorithm for Clustered, Ordered Assessment of Targeting in miR-mRNA interactions. *PEACOAT* not only offers the ability to infer miR-mRNA interactions as well as cluster miRs and/or mRNAs, but can also incorporate arbitrarily many miR-mRNA target predictions and prediction scores, and it allows the user to place an arbitrarily high or low amount of weight on such prediction information.

PEACOAT is tested on simulated networks of different sizes, and both with and without the use of prediction information. It is shown that clustering is frequently advantageous in the inference of true miR-mRNA interactions, and that the inclusion of helpful prediction information is likewise advantageous. The paper analyzes when and how to use clustering and predictions, and then applies what is learned to a data set of miR and mRNA expression from multiple myeloma patients, comparing the results to those of the sequential model from [13] and *TaLasso*.

2 Methods

This paper describes a new Bayesian model of miR-mRNA interaction that adds substantial capabilities to a simpler model presented in [13]. That model was specifically designed to infer miR-mRNA interactions in partially ordered expression data, while making use of target prediction databases/algorithms. Here, an improved version of this model is presented and, in addition, it is combined it with a Bayesian clustering model, allowing us to reduce the dimension of the interaction space without reducing the data set in a manner that has been shown in genetic interaction models to infer interactions more reliably when correlation between interacting elements might be high [16].

Therefore, while this model requires paired miR-mRNA expression data and a partial ordering of samples or groups of samples, the partial ordering can be made to resemble a reference design, or indeed any other design of choice, as long as every sample is designated to precede or follow at least one other sample in the partial ordering. However, this model's strengths lie in inferring interactions in inherently partially ordered data (e.g. development stages, time series, etc.) where correlation between regulatory actors (i.e. miRs) is often high.

2.1 Model specifications

In short, the model is formulated as in [13], with two notable modifications as well as the incorporation of clustering. The first exception is the specified partial ordering of sample stages only to the mRNA and not to the miR expression values. Instead of prior distributions based on a stage's specified parent stage(s), it is assumed that the miR expression values in each stage are normally distributed from the same prior distribution. This gives a lower probabilistic penalty to large changes in expression between consecutive stages and better enables proper fitting of miR parameters across all stages. The second notable change from the model in [13] is that the two model precisions within the interaction parameters are tied together so that we may *a priori* specify the level of influence the target predictions have on the inferred interactions, when compared to the influence of the expression data. The model in [13] typically, through optimization of parameters and priors, placed high emphasis on the target predictions, but this is not always desireable and the new model includes a fixed parameter allowing the adjustment of this influence; more details can be found below.

2.1.1 miR and mRNA expression parameters

For a set of partially-ordered stages such that each stage s_0 has a set of parent stages $\rho_{s0} = \{s \in S : s < s_0\}$ according to the partial ordering, and given that there are N_{miR} total miRs, N_m total mRNAs, K_{miR} miR clusters, and K_m mRNA clusters, each miR cluster expression value $F^{miR}_{k,s}$, for cluster k and stage s, follows the distribution

$$F^{miR}_{k,s} \sim \mathcal{N}(0, \lambda_F) \tag{1}$$

where λ_F is a scalar precision. By not considering the partial ordering among the stages for miR expression, both the development and trend parameters included in [13] cannot be included, but we are left with a slightly simpler, more focused model.

The cluster expression values $F^m_{k,s}$ for mRNA cluster k and stage s are modeled in the same manner as individual mRNA expression was in [13]. That is, given a scalar development parameter $\Delta_{\rho,s}$ (intuitively a distance) from a parent stage ρ to its child s, the $F^m_{k,s}$ are assumed to be normally distributed with mean

$$\mu_{k,s} = \frac{\sum\limits_{\rho\in\rho_s} \frac{1}{\Delta_{\rho,s}} \left(F^m_{k,\rho} + \boldsymbol{\eta_k} \bullet (\boldsymbol{F^{miR}_s} - \boldsymbol{F^{miR}_\rho}) \right)}{\sum\limits_{\rho\in\rho_s} \frac{1}{\Delta_{\rho,s}}} \tag{2}$$

and precision

$$\lambda_{k,s} = \frac{\sum\limits_{\rho\in\rho_s} \left(\frac{1}{\Delta_{\rho,s}} \right)^2}{\sum\limits_{\rho\in\rho_s} \frac{1}{\Delta_{\rho,s}}} \Lambda_k \tag{3}$$

where the vector $\boldsymbol{\eta_k}$ contains interaction coefficients between mRNA cluster k and all miR clusters, the vector $\boldsymbol{F^{miR}_s}$ contains the values $F^{miR}_{s,l}$ for all miR clusters l, and Λ_k is a precision parameter of the expression of cluster k.

For each miR or mRNA i, it is assumed that i belongs to exactly one cluster k (either a miR cluster or an mRNA cluster), and that its expression $\omega^\Xi_{i,s}$—where Ξ is either "miR" or "m" (short for mRNA)—is normally distributed about the cluster expression $F^\Xi_{k,s}$, as in

$$\omega^\Xi_{i,s} \sim \mathcal{N}(F^\Xi_{k,s}, \gamma_{k,s}) \tag{4}$$

It is also assumed that the mean expression $\varepsilon^\Xi_{i,s,m}$ of miR or mRNA i from a given expression set m (e.g. the mean of repeated spots on microarray m) in stage s is normally distributed as

$$\varepsilon^\Xi_{i,s,m} \sim \mathcal{N}(\omega^\Xi_{i,s}, \kappa_\Xi) \tag{5}$$

Furthermore, within each expression set m, it is assumed that the expression data $x^\Xi_{i,s,m,n}$ for within-set replicate n (e.g. the n^{th} replicate spot on a microarray) and stage s are normally distributed as

$$x^\Xi_{i,s,m,n} \sim \mathcal{N}(\varepsilon^\Xi_{i,s,m}, \kappa'_\Xi) \tag{6}$$

where again Ξ is either "miR" or "m" ("mRNA").

2.1.2 Interaction parameters

For miR cluster l and mRNA cluster k, each element $\eta_{k,l}$ of the vector of interaction coefficients $\boldsymbol{\eta_k}$ mentioned above is normally distributed according to

$$\eta_{k,l} \sim \mathcal{N}\left(\sum_{i\in l, j\in k} \Theta_{i,j},\ \varphi \right) \tag{7}$$

for miRs i and mRNAs j, so that the cluster interaction parameter $\eta_{k,l}$ is assumed to be the sum of all individual miR-mRNA interaction parameters $\Theta_{i,j}$ such that miR i is in miR cluster l and mRNA j is in mRNA cluster k. These individual miR-mRNA interaction parameters $\Theta_{i,j}$ are assumed to be distributed as in [13], but with the addition of the prediction weight value $\Upsilon \in (0, \infty)$, which is multiplied by the same precision parameter φ as above, as in

$$\Theta_{i,j} \sim \mathcal{N}(\boldsymbol{\beta} \bullet \boldsymbol{P}_{i,j}, \Upsilon\varphi) \tag{8}$$

allowing arbitrary weight to be placed on the targeting predictions relative to the $\eta_{k,l}$. The vector $\boldsymbol{P}_{i,j}$ contains fixed parameters concerning miR i and mRNA j from target prediction algorithms, and $\boldsymbol{\beta}$ is a vector of estimated coefficients.

2.1.3 Prior distributions

Conjugate prior distributions are chosen, where possible. Thus, normal prior distributions on the parameters $F_{i,s}^{miR}$ and $F_{i,0}^{m}$ were chosen, as well as gamma prior distributions on Λ_j, $\gamma_{k,s}$, κ_{miR}, κ_{mRNA}, κ'_{miR}, κ'_{mRNA}, and φ. The prior distribution for $\boldsymbol{\beta}$ is the equivalent multivariate normal with zero mean and precision matrix $10^{-10}\mathbf{I}$, where $\mathbf{I}$ is the identity matrix of the appropriate size. The model fitting begins with vaguely informative priors and then iteratively updates the prior distributions to maximize the marginal likelihood, as in [13] and [9]. The development parameters $\Delta_{\rho,s}$ are again treated as fixed, though as in [13] the parameters are typically updated (unless stated otherwise) to maximize the marginal likelihood.

2.2 Fitting the model using variational Bayes methods

Variational Bayesian methods are used to estimate the parameters of our model, as in [13]. Our methods of model fitting (though not necessarily the model itself) are very similar to those of [8, 17, 18] in discovering miR-mRNA target pairs as well as other analyses of gene expression data in [9] and [19]. For a thorough explanation of variational Bayesian methods, see [20] or [21]. This algorithm was coded in the *R*[22] statistical programming language.

2.3 Target prediction algorithms

This paper analyzes miR-mRNA target prediction data from both *TargetScan* [23, 24, 25, 26] and *MiRanda* [27, 28]. For each of these, a table of predicted miR-mRNA interactions and the targeting scores calculated by the respective algorithms was downloaded. *TargetScan* includes a *context+* score and *MiRanda* includes a *mirSVR* score [29].

If one of the included algorithms predicts that miR i targets mRNA j, the vector $\langle 1, \alpha_{i,j} \rangle$ is used, where $\alpha_{i,j}$ is the prediction score from the algorithm. The vectors from multiple prediction algorithms are concatenated such that, if a pair $\{i, j\}$ is predicted by more than one algorithm, the prediction parameter vector $\boldsymbol{P}_{i,j}$ is of the form $\langle 1, \alpha_{i,j}, 1, \alpha'_{i,j} \rangle$.

2.4 Simulations

Data were simulated in order to evaluate the performance of the model under different conditions. Since there is no gold standard data set including all true-positive and true-negative miR-mRNA interactions, the simulated data sets are used as a substitute in which it is known whether a true interaction exists in each case.

The simulated networks consist of miR and mRNA expression data for each of eight ordered stages/samples. For each of the eight stages, there are three replicates for each mRNA type. To do this it is assumed that, for any given miR-mRNA pair, there is a 10% chance that the pair is predicted by our fictional targeting database. Then, it is assumed that the predicted pairs have a higher probability of being true target pairs: each predicted pair has a 50% chance of being a true target pair while non-predicted pairs have a 10% chance. The choice of each pair as true, according to these probabilities, is independent, and each true target pair is assigned a random negative interaction coefficient by taking the negation of a random draw from a gamma distribution with shape=rate=1. These are treated as the true interactions.

The miR expression data are then simulated through a random walk process starting at $t = 0$ and ending at $t = 8$, with each subsequent step, starting at position 0, chosen by the standard normal distribution. The data from points $t = 1$ through $t = 8$ are then used as simulated miR data. From these simulated miR expression values, the true interaction matrix is used to generate mRNA expression data for the same eight time points. Finally, the three "technical" replicates of each of the eight time points are created and independent Gaussian noise, with variance 0.1, is added to each data point; this is the final data set.

2.5 Multiple myeloma data

The model is demonstated using the multiple myeloma data set from [30], which can be downloaded from *GEO* [31], accession number GSE17498. In this data set, there are both miR and mRNA expression values for samples from 36 patients, 34 of whom have been diagnosed with multiple myeloma (MM) and 2 of whom have been diagnosed with plasma cell leukemia (PCL). Unfortunately, from healthy donors there is only miR expression data and no mRNA data, so healthy samples cannot be included in our study. However, the diseased samples can be arranged into four Durie-Salmon stages: IA, IIA, IIIA, and IIIB, which gives an obvious ordering for our non-PCL samples.

As in [13], prior to the main analysis, quantile normalization was performed across all arrays of the data set using the *limma* package for *R* [32, 22]. I then performed a probewise ANOVA to test for differential expression across the stages IA, IIA, IIIA, IIIB, and PCL (using individuals within a stage as replicates) and removed those probes/probesets (for both miR and mRNA) whose (unadjusted) p-value from the ANOVA F-test was greater than or equal to 0.05, as well as those miRs and mRNAs not involved in any predicted targeting interactions, leaving 28 miRs and 367 mRNAs as possible candidates for targeting interaction. Lastly, the data were re-scaled so that each probe/set—across all samples—had a mean of zero and a standard deviation of one.

3 Results

Below, details are given for the performace of *PEACOAT* on simulated data sets as well as a microarray data set of multiple myeloma samples.

3.1 Simulations

Five simulated networks are evaluated, with 10 miRs and 10 mRNAs, as well as 5 networks with 25 miRs and 25 mRNAs. When inferring interactions for the simulations, all development parameters $\Delta_{\rho,s}$ are left at a fixed at a value of 1, and they are not updated. This saves a considerable amount of time without materially affecting the integrity of the simulation results. The prediction weight parameters are set to $\Upsilon = 1$. For each simulated network and model configuration mentioned, a range of are tried, and for each of these 5 models were fit, each starting from different random parameter values.

In order to evaluate the performance of the model in inferring predictions, the AUROC statistic (area under the receiver operator characteristic) was used. Given a ranked list of interactions by statistical significance and a cut-off point on that list, The receiver operating characteristic (ROC) is the true positive rate (proportion of true interactions appearing above the cut-off) divided by the false positive rate (proportion of false interactions appearing above the cut-off). The AUROC is the area under the curve generated by calculating the ROC for all possible cut-off points. In general, it is a good measure of whether or not the items at or near the top of the list are true or not, and thus provides a general idea of the verifiability in practice of miR-mRNA interactions with the highest statistical significance.

Table 1 show the AUROC scores for *PEACOAT* over a range of cluster numbers on the 5 simulated networks with 10 miRs and 10 genes. For each network and number of clusters the AUROC shown is that of the inferred model with the highest likelihood score among the five model fits with random parameter initializations. For all networks, the highest AUROC is obtained by having fewer than 10 clusters, leading us to believe that clustering aids in interaction inference; however, the best number of clusters differs between networks, and, as shown in the table, setting the number of clusters to 7 is the only choice that out-performs 10 clusters (*i.e.* no clustering) in most cases.

The simulated networks with 25 miRs and genes, on the other hand, are best inferred using 11-15 clusters. As shown in Table 2, using a number of clusters in the range of 11-15 out-performed the no-clustering model 60-80% of the time, depending on the exact number of clusters. Furthermore, a randomly selected model within this range (instead of choosing the most likely of the 5 randomly initialized models for each cluster number), has a 68% chance of out-performing the no-clustering model. That implies that, when given a choice between fitting a model without clustering and fitting one with clustering, it is advantageous, more often than not, to choose to cluster, given the manner by which our simulated data were generated.

3.2 Multiple myeloma data

Given that the multiple myeloma data set, after pre-processing, contains 28 miRs, *PEACOAT* was fit to the data using a number of clusters from the optimal range from the simulated net-

Table 1: ***Size 10 network AUROC scores.*** **For each of 5 simulated data sets, the table shows the AUROC from the highest-likelihood inferred network (of 5) from each of a range of cluster numbers. The best score for each network is shown in bold. The bottom two rows give (1) the percentage of data sets for which the best model for the particular number of clusters out-performed the best no-clustering (*i.e.* 10 cluster) model, and (2) the percentage of the all models fit for the particular number of clusters that out-performed the best no-clustering model.**

Number of clusters:	3	4	5	6	7	8	9	10
Network 1:	0.677	0.637	0.664	0.667	0.711	**0.732**	0.671	0.712
Network 2:	0.487	0.545	0.482	**0.608**	0.558	0.515	0.507	0.532
Network 3:	0.416	0.450	0.486	0.498	**0.581**	0.531	0.550	0.537
Network 4:	0.634	0.518	0.651	0.647	**0.655**	0.592	0.580	0.652
Network 5:	0.588	0.682	**0.825**	**0.825**	0.723	0.819	0.779	0.732
% better than no clustering:								
each best model (of 5):	0%	20%	20%	40%	60%	40%	40%	0%
all models (5 models $\times$ 5 data sets):	12%	36%	20%	44%	56%	32%	40%	0%

Table 2: ***Size 25 network AUROC scores.*** **For each of 5 simulated data sets, the table shows the AUROC from the highest-likelihood inferred network (of 5) from each of a range of cluster numbers. The best score for each network is shown in bold. The bottom two rows give (1) the percentage of data sets for which the best model for the particular number of clusters out-performed the best no-clustering (*i.e.* 25 cluster) model, and (2) the percentage of the all models fit for the particular number of clusters that out-performed the best no-clustering model.**

Number of clusters:	11	13	15	17	19	21	23	25
Network 1:	**0.709**	0.674	0.628	0.565	0.590	0.596	0.598	0.609
Network 2:	0.530	0.614	0.593	0.524	0.603	0.603	**0.639**	0.601
Network 3:	0.591	0.606	**0.639**	0.557	0.624	0.546	0.570	0.588
Network 4:	0.538	0.506	0.521	0.559	0.611	0.593	0.612	**0.617**
Network 5:	0.632	0.634	**0.641**	0.632	0.603	0.568	0.574	0.520
% better than no clustering:								
each best model (of 5):	60%	80%	60%	20%	60%	40%	40%	0%
all models (5 models $\times$ 5 data sets):	68%	68%	68%	40%	44%	28%	28%	0%

works of size 25. Since 28 is slighly larger than 25, 15 clusters was chosen as an appropriate number. The greatly increased computational time on this larger data set prevents us from trying more model fits and cluster numbers. The results using 15 clusters were compared to the results from the unclustered individual-ordered (*I-O*; where samples are ordered by developmental stage and samples from the same stage are not treated as replicates) model from [13] in order to show that *PEACOAT* with clustering can improve the accuracy of predicted interactions.

Table 3 shows the KEGG pathway terms [33] that are enriched within the set of unique genes appearing in the top 100 ranked interactions. Enrichment was calculated using the *GeneCoDis* tool [34, 35].

Table 3: ***Enriched KEGG pathways among genes in the top 100 interactions*** **The top row gives the number of unique genes present in the top 100 miR-mRNA interactions according to each model; the remaining rows give, per column, the number of these genes annotated by *KEGG* pathway terms with significant enrichment (FDR corrected $p < 0.05$) for each of the models. The [uncorrected] hypergeometric p-value is given in parentheses. A blank entry indicates that the particular pathway was not significantly enriched in the model. Cancer-related *KEGG* pathways are shown at the top of the table, with other enriched pathways below.**

	I-O model	TaLasso	*PEACOAT*
Number of unique genes in the top 100 interactions	54	84	25
05215 :Prostate cancer	3 (0.000511)		2 (0.002343)
05214 :Glioma	2 (0.005557)		2 (0.001211)
05218 :Melanoma	2 (0.006818)		2 (0.001498)
05219 :Bladder cancer	2 (0.002509)		
05016 :Huntington's disease	4 (0.000296)		
04115 :p53 signaling pathway	3 (0.000239)		2 (0.001408)
05010 :Alzheimer's disease	3 (0.003057)		
05014 :Amyotrophic lateral sclerosis (ALS)	2 (0.003820)		
04622 :RIG-I-like receptor signaling pathway	2 (0.007008)		
04976 :Bile secretion	2 (0.007008)		
04210 :Apoptosis	2 (0.010362)		
04120 :Ubiquitin mediated proteolysis		4 (0.0005292)	
04960 :Aldosterone-regulated sodium reabsorption			2 (0.000539)
04722 :Neurotrophin signaling pathway			2 (0.004585)
05160 :Hepatitis C			2 (0.005331)

Though fewer KEGG pathways were found to be significantly enriched by *PEACOAT* (7 pathways) than by the *I-O* model (11 pathways), the main concern is cancer-related pathways, of which the two models indicated 3 and 4 pathways, respectively. Of note is that *PEACOAT* found 3 pathways to be significant among only 25 genes. More importantly, the enrichment p-values for these three pathways are more significant than all but one of the 4 pathways found significant by the *I-O* model, indicating that a researcher is more likely to find a gene related to one of these pathways by independently verifying some genes from the list generated by *PEACOAT* than by the non-clustering model. This is true admittedly only by a narrow margin, but nonetheless provides further evidence of the usefulness of clustering within a miR-mRNA targeting model.

4 Discussion

The results given in this paper have shown that there are significant advantages to using clustering within a miR-mRNA interaction model. Our algorithm, which expands upon a previous algorithm incorporating both expression data and sequence-based predictions, performs best when clustering is enabled, as shown with both simulated and biological data. Since the algorithm fully integrates a version of a dynamic Bayesian network with a clustering model, the clusters and miR-mRNA interactions can be inferred simultaneously via an interative variational Bayesian algorithm.

First, the results showed that, given our simulated data with 10 miRs and 10 mRNAs, the optimal number of clusters was typically 7, but varied from 5 to 8. For the 10-miR networks, choosing 7 clusters gave better results than no-clustering in the majority of cases. One might expect that choosing 8 or 9 clusters would also give better results, since 7 clusters was better than 10 (i.e. no clustering), but this was not the case. Most likely, this is the result of two factors: (1) there is some randomness in the results, particularly when there are only five networks that were tested, and (2) choosing 10 clusters by design disables the clustering algorithm and removes an entire layer of inference that could potentially cause a lot of uncertainty in results. That is, when clustering is disabled, the cluster membership variables are fixed to exact values, and this absolute certainty can make the interaction inference algorithm converge much more quickly. This is something that could be tested rigorously in the future.

For the simulated networks of 25 miRs and 25 mRNAs, clustering proved much more valuable. The optimal number of clusters, depending on specific network, ranged from 11 all the way to 25, but choosing a number of clusters between 11 and 15 gave better results than a no-clustering model 68% of the time. This is strong evidence that clustering improves inference, particularly when few samples are present (in this case, 8 time points).

Lastly, when applied to a multiple myeloma data set, *PEACOAT* inferred 3 cancer-related KEGG pathways that were significantly enriched among the top 100 interactions. The level of significance of these enriched pathways was better than the top three cancer-related pathways from the non-clustering model upon which *PEACOAT* was based. If the simulated data proved that *PEACOAT* could infer true interactions more strongly with clustering than without, then these results demonstrate that the algorithm is a step forward in miR-mRNA target inference from real expression data.

It is commonly acknowledged that some genes work together as groups or modules within certain processes to accomplish a task, and that these genes are often co-expressed or at least highly correlated under certain circumstances. There is no reason to believe that miRs behave otherwise. Admitting that miRs sometimes form functional or de-facto clusters gives extra statistical power to network inference tasks. In a linear model such as *PEACOAT* as well as the most commonly used network inference tools, causality cannot be determined if two potential regulators are highly correlated, and only clustering can avoid the adverse effects that such uncertainty causes.

The results have shown using simulated data that algorithms with clustering can outperform algorithms with no clustering with regards to the task of miR-mRNA network inference, particularly when more miRs are included as potential regulators. On biological data, *PEACOAT* found more highly enriched gene function than non-clustering algorithms. Collectively, these

results suggest, first of all, that the clustering of miRs is potentially a critical part of faithfully inferring miR-mRNA network interactions, and second of all that *PEACOAT* is a valuable and significant step in the right direction.

PEACOAT represents the state of the art in miR-mRNA interaction inference, incorporating both expression data and sequence-based target prediction, plus an integrated clustering algorithm that infers cluster members concurrently with the interactions themselves.

Source code for *PEACOAT*, in the *R* language, is freely available from the author upon request.

References

[1] A. Kozomara and S. Griffiths-Jones. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Research*, 39(suppl 1):D152–D157, 2011.

[2] S. Griffiths-Jones, H. K. Saini, S. van Dongen and A. J. Enright. miRBase: tools for microRNA genomics. *Nucleic Acids Research*, 36(suppl 1):D154–D158, 2008.

[3] S. Griffiths-Jones, R. J. Grocock, S. van Dongen, A. Bateman and A. J. Enright. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research*, 34(suppl 1):D140–D144, 2006.

[4] J. Gäken, A. M. Mohamedali, J. Jiang et al. A functional assay for microRNA target identification and validation. *Nucleic Acids Research*, 40(10):e75, 2012.

[5] A. Muniategui, J. Pey, F. Planes and A. Rubio. Joint analysis of miRNA and mRNA expression data. *Briefings in Bioinformatics*, 14(3):263–278, 2013.

[6] C. A. Penfold and D. L. Wild. How to infer gene networks from expression profiles, revisited. *Interface Focus*, 1(6):857–870, 2011.

[7] A. Muniategui, R. Nogales-Cadenas, M. Vázquez, Aranguren, X. Agirre, A. Luttun, F. Prosper, A. Pascual-Montano and A. Rubio. Quantification of miRNA-mRNA Interactions. *PLoS ONE*, 7(2):e30766, 2012.

[8] J. C. Huang, Q. D. Morris and B. J. Frey. Bayesian Inference of MicroRNA Targets from Sequence and Expression Data. *Journal of Computational Biology*, 14(5):550–563, 2007.

[9] M. J. Beal, F. Falciani, Z. Ghahramani, C. Rangel and D. L. Wild. A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics*, 21(3):349–356, 2005.

[10] S. Lèbre. Inferring Dynamic Genetic Networks with Low Order Independencies. *Statistical Applications in Genetics and Molecular Biology*, 8(1):1–38, 2009.

[11] M. Gustafsson, M. Hornquist and A. Lombardi. Constructing and Analyzing a Large-Scale Gene-to-Gene Regulatory Network-Lasso-Constrained Inference and Biological Validation. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 2(3):254–261, 2005.

[12] A. Fujita, J. Sato, H. G. Malpartida, R. Yamaguchi, S. Miyano, M. Sogayar and C. Ferreira. Modeling gene expression regulatory networks with the sparse vector autoregressive model. *BMC Systems Biology*, 1(1):e39, 2007.

[13] B. Godsey, D. Heiser and C. Civin. Inferring MicroRNA Regulation of mRNA with Partially Ordered Samples of Paired Expression Data and Exogenous Prediction Algorithms. *PLoS ONE*, 7(12):e51480, 2012.

[14] A. Tanzer and P. F. Stadler. Molecular Evolution of a MicroRNA Cluster. *Journal of Molecular Biology*, 339(2):327–335, 2004.

[15] L. Wang, A. L. Oberg, Y. W. Asmann et al. Genome-Wide Transcriptional Profiling Reveals MicroRNA-Correlated Genes and Biological Processes in Human Lymphoblastoid Cell Lines. *PLoS ONE*, 4(6):e5878, 2009.

[16] B. Godsey. Improved inference of genetic regulatory networks through integrated Bayesian clustering and dynamic modeling of time-course expression data. *PLoS ONE*, In press, 2013.

[17] J. C. Huang, T. Babak, T. W. Corson, G. Chua, S. Khan, B. L. Gallie, T. R. Hughes, B. J. Blencowe, B. J. Frey and Q. D. Morris. Using expression profiling data to identify human microRNA targets. *Nature Methods*, 4(12):1045–1049, 2007.

[18] J. C. Huang, B. J. Frey and Q. D. Morris. Comparing sequence and expression for predicting microRNA targets using GenMiR3. *Pacific Symposium on Biocomputing*, pages 52–63, 2008.

[19] A. E. Teschendorff, Y. Wang, N. L. Barbosa-Morais, J. D. Brenton and C. Caldas. A variational Bayesian mixture modelling framework for cluster analysis of gene-expression data. *Bioinformatics*, 21(13):3025–3033, 2005.

[20] J. M. Winn. *Variational Message Passing and its Applications*. Ph.D. thesis, St Johns College, Cambridge, Cambridge, England, 2003.

[21] M. J. Beal. *Variational algorithms for approximate Bayesian inference*. Ph.D. thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

[22] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009.

[23] 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.

[24] R. C. Friedman, K. K. Farh, C. B. Burge and D. P. Bartel. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, 19(1):92–105, 2009.

[25] A. Grimson, K. K. Farh, W. K. Johnston, P. Garrett-Engele, L. P. Lim and D. P. Bartel. MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing. *Molecular Cell*, 27(1):91–105, 2007.

[26] D. M. Garcia, D. Baek, C. Shin, G. W. Bell, A. Grimson and D. P. Bartel. Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. *Nature Structural & Molecular Biology*, 18(10):1139–1146, 2011.

[27] B. John, A. J. Enright, A. Aravin, T. Tuschl, C. Sander and D. S. Marks. Human MicroRNA targets. *PLoS Biology*, 2(11):e363, 2004.

[28] A. Enright, B. John, U. Gaul, T. Tuschl, C. Sander and D. Marks. MicroRNA targets in Drosophila. *Genome Biology*, 5(1):R1, 2003.

[29] D. Betel, A. Koppal, P. Agius, C. Sander and C. Leslie. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biology*, 11(8):R90, 2010.

[30] M. Lionetti, M. Biasiolo, L. Agnelli et al. Identification of microRNA expression patterns and definition of a microRNA/mRNA regulatory network in distinct molecular groups of multiple myeloma. *Blood*, 114(25):e20–e26, 2009.

[31] T. Barrett, D. B. Troup, S. E. Wilhite et al. NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Research*, 37(suppl 1):D885–D890, 2009.

[32] G. Smyth and T. Speed. Normalization of cDNA microarray data. *Methods*, 31:265–273, 2003.

[33] M. Kanehisa, M. Araki, S. Goto et al. KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, 36(suppl 1):D480–D484, 2008.

[34] P. Carmona-Saez, M. Chagoyen, F. Tirado, J. M. Carazo and A. Pascual-Montano. GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. *Genome Biology*, 8(1):R3, 2007.

[35] R. Nogales-Cadenas, P. Carmona-Saez, M. Vazquez, C. Vicente, X. Yang, F. Tirado, J. M. M. Carazo and A. Pascual-Montano. GeneCodis: interpreting gene lists through enrichment analysis and integration of diverse biological information. *Nucleic Acids Research*, 37(suppl 2):W317–W322, 2009.