Genome wide predictions of miRNA regulation by transcription factors

Matthew Ruffalo and Ziv Bar-Joseph*

Department of Computational Biology, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA 15213

*To whom correspondence should be addressed.

Abstract

Motivation: Reconstructing regulatory networks from expression and interaction data is a major goal of systems biology. While much work has focused on trying to experimentally and computationally determine the set of transcription-factors (TFs) and microRNAs (miRNAs) that regulate genes in these networks, relatively little work has focused on inferring the regulation of miRNAs by TFs. Such regulation can play an important role in several biological processes including development and disease. The main challenge for predicting such interactions is the very small positive training set currently available. Another challenge is the fact that a large fraction of miRNAs are encoded within genes making it hard to determine the specific way in which they are regulated.

Results: To enable genome wide predictions of TF–miRNA interactions, we extended semi-supervised machine-learning approaches to integrate a large set of different types of data including sequence, expression, ChIP-seq and epigenetic data. As we show, the methods we develop achieve good performance on both a labeled test set, and when analyzing general co-expression networks. We next analyze mRNA and miRNA cancer expression data, demonstrating the advantage of using the predicted set of interactions for identifying more coherent and relevant modules, genes, and miRNAs. The complete set of predictions is available on the supporting website and can be used by any method that combines miRNAs, genes, and TFs.

Availability and Implementation: Code and full set of predictions are available from the supporting website: http://cs.cmu.edu/~mruffalo/tf-mirna/

Contact: zivbj@cs.cmu.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Transcription factors regulate gene expression by binding to genome sequences in their promoters or enhancer regions (Latchman, 1997a; Karin, 1990; Roeder, 1996). MicroRNAs (or miRNAs) are another type of regulatory molecules which post-transcriptionally inhibit target genes by binding to complementary sequences of mRNAs and preventing their translation (Bartel, 2004). While several methods have been developed for integrating various high throughput datasets to predict gene targets of transcription factors (TFs) (Boulesteix and Strimmer, 2005; Bulyk et al., 2004; Kaplan et al., 2005; Pique-Regi et al., 2011) and to predict gene targets of miRNAs (Agarwal et al., 2015; Coronello and Benos, 2013; Lewis et al., 2003), much less work has focused on predicting the regulation of miRNAs by TFs. However, models that integrate TF, genes and miRNA require such information to provide a comprehensive overview of gene regulation in the condition being studied (Marson et al., 2008; Murali et al., 2011).

One major problem that may have prevented the development of methods for predicting such interactions is the lack of large scale, verified, TF–miRNA regulation databases. While several databases exist for recording validated TF–gene interactions (Han et al., 2015; Kummerfeld and Teichmann, 2006; Matys et al., 2003; Sandelin et al., 2004), to the best of our knowledge, the largest TF–miRNA interaction database, TransmiR (Wang et al., 2010) contains only 441 validated interactions for human TFs and miRNAs. Given the large number of potential interactions (there are more than 2000 human TFs and more than 2000 identified human miRNAs) such a small number of validated interactions is likely less than a small fraction of the total number of interactions. Indeed, of the current ~2800 human miRNAs listed in miRBase (Griffiths-Jones et al., 2006) only 126 have any validated TF regulator in TransmiR.

Another problem that is unique to miRNAs (when compared with genes) is that many of them are encoded within a gene coding region (roughly 75.5% of identified human miRNAs are within such
coding regions). For these miRNAs, it is not immediately clear if regulation occurs upstream of the miRNA genomic location (i.e. within the gene open reading frame or ORF) or only upstream of the gene. If the former that miRNA can be expressed independent of the gene whereas if the latter holds then the miRNA is only expressed as part of the gene itself. Since almost all TF–gene prediction methods rely on the analysis of upstream regions (integrating various information sources about TF activity in these regions, Ernst et al., 2010; Raj et al., 2015; Zhong et al., 2013) any method for predicting TF–miRNA interactions would need to determine whether it focuses on the upstream region of the miRNA or the gene in which it is encoded.

To address these problems and to generate a comprehensive set of predicted TF–miRNA interactions for use in systems biology applications, we developed and tested semi-supervised learning approaches (Zhu, 2011) which integrate sequence, conservation, expression and ChIP-Seq data. The main advantage of semi-supervised learning for this problem is its ability to utilize both labeled (known interactions) and unlabeled (interactions that have not been tested) when learning model parameters. For the semi-supervised learning, we tested both co-training (Blum and Mitchell, 1998), a standard and very successful method and a new label propagation method that utilizes graphs to reassign labels to unobserved potential interactions. As we show, while both methods worked well in a cross-validation setup, the co-training led to results that did not generalize well while the label propagation method performed much better on an independent dataset. In addition, our analysis identified that performance is much improved when using the gene upstream regions (as opposed to the miRNA upstream regions) for predicting interactions for those miRNAs that are encoded within genes.

2 Methods

We first discuss the types of data we integrated in order to create features for the prediction task. Next we discuss the classifiers we used for predicting these interactions, including classifiers that were shown in the past to work well for predicting interactions in genomic data and a novel label propagation approach which we specifically tailored to the (usually small number of) positive training points we had.

2.1 Datasets used for the prediction task

Previous work (Banovich et al., 2014; Ernst et al., 2010; Raj et al., 2015; Zhong et al., 2013) has shown the value of integrating multiple data types to predict TF binding (for gene targets). Based on these observations, our approach also combines several different types of sequence and TF binding data to predict TF–miRNA interactions. Specifically, we used.

2.1.1 Transcription factor motifs

Transcription factors bind to specific sequences of DNA (Latchman, 1997b), and the set of binding sites is often represented as a position frequency matrix (Stormo, 2000). We use motifs for 205 transcription factors, obtained from the JASPAR database (Sandelin et al., 2004).

2.1.2 PhastCons score

Although several transcription factor binding sites can diverge across species (Borneman et al., 2007; Nitta et al., 2015), such sites are often much more conserved than other intergenic regions and conservation of binding sites has been shown to result in higher gene expression (Hemberg and Kreiman, 2011). PhastCons scores (Siepel et al., 2005) quantify the degree to which each genomic locus is conserved between species; we obtained this data for the human genome from the UCSC Genome Browser (Hinrichs et al., 2006).

2.1.3 DNase I hypersensitivity data

Chromatin accessibility, measured via sensitivity to the DNase I enzyme, quantifies how open the chromatin is at specific genomic location. Recent work (John et al., 2011; Li et al., 2011; Thurman et al., 2012; Zhong et al., 2013) has shown that DNase I cleavage corresponds with TF binding. We obtain DNase I hypersensitivity data from ENCODE (Consortium et al., 2004).

2.1.4 ChIP-seq binding score

Chromatin immunoprecipitation (ChIP) followed by next-generation sequencing (ChIP-seq) is a popular technique for identifying the binding sites of transcription factors. We obtained ChIP-seq data for 202 transcription factors from the ENCODE (Consortium et al., 2004) database.

2.1.5 Genome locations of known human miRNAs

From miRBase (Griffiths-Jones et al., 2006), we obtained the chromosomal coordinates of 2813 known human miRNAs.

2.1.6 Genome locations of known human genes

We obtained chromosomal coordinates of known human protein coding sequences from UCSC (Hinrichs et al., 2006).

2.1.7 miRNA expression measurements

From ENCODE (Consortium et al., 2004), we downloaded miRNA expression measurements for 134 tissues, across 2734 of the 2813 miRNAs present in miRBase.

2.1.8 Known TF–miRNA interactions

The TransmiR database (Wang et al., 2010) contains data for 745 manually curated regulatory interactions between TFs and miRNAs; of these, 441 were identified in humans. Of these 441, 266 match miRNAs present in the miRBase data.

2.1.9 Human genome

We obtained chromosomal sequences from the hg38 build of the human genome, from the UCSC Genome Browser (Hinrichs et al., 2006).

2.1.10 Paired mRNA and miRNA expression from cancer samples

From TCGA (Brennan et al., 2013), we downloaded mRNA and miRNA expression data for 560 glioblastoma (GBM) samples, across 17 814 genes and 534 miRNAs.

2.1.11 Gene targets of human miRNAs

The TargetScan v7.0 (Agarwal et al., 2015) database contains 389 339 mRNA → gene targets, from 219 distinct miRNAs to 12 781 distinct genes.

2.1.12 Gene targets of human TFs

From the TTRUST database (Han et al., 2015), we obtained 8215 TF → gene interactions, between 748 distinct TFs and 2374 distinct target genes.
2.2 Extracting features from genomic datasets

We examined the intersection between the coordinates of the 2813 miRNAs from miRBase (Griffiths-Jones et al., 2006) with the chromosomal coordinates for known human genes from the UCSC Genome Browser (Hinrichs et al., 2006). We marked the 2124 miRNAs (75.5%) overlapping known genes as ‘genic’ and the rest as ‘intergenic’. Following Ernst et al. (2010) and Zhong et al. (2013), we extracted DNA sequences from the corresponding miRNA coordinates of the human genome: 5kb upstream and 1kb downstream of each miRNA, and for genic miRNAs, an additional 5kb upstream of the gene containing that miRNA.

With these genomic sequences from putative regulatory regions for each miRNA, we next used motifs for each human transcription factor in the JASPAR database (Sandelin et al., 2004) to search for PWM hits in these genomic regions, using the BioPython library (Cock et al., 2009). We then combined data for each PWM hit with additional data sources described above, producing the following features for each TF–miRNA pair: DNase hypersensitivity data (from ENCODE, Consortium et al., 2004) and PhastCons scores (from UCSC, Hinrichs et al., 2006) for the corresponding genome region and ChIP-seq data (from ENCODE, Consortium et al., 2004) for the appropriate transcription factor.

With the data described above, we extract the following feature vector for each TF–miRNA pair:

- Maximum PWM score.
- Number of PWM hits with scores above predefined thresholds: {0, 5, 10, 15, 20, 25}.
- Mean score of all hits with score higher than the above thresholds.
- Maximum PhastCons score for the motif location.
- Maximum DNase I hypersensitivity score for the motif.
- Maximum ChIP-seq score for the corresponding TF in that genomic region.

Note that the PhastCons and DNase scores are defined for each genomic locus, while the ChIP-seq score and motif PWM scores are specific to a given transcription factor binding to a specific genomic coordinate.

With the 16 features defined above (four individual features and two features for each of the 6 PWM score thresholds), we calculate a feature vector that represents each TF–miRNA pair. In the following analysis, we represent each potential TF–miRNA interaction by its feature vector.

2.3 Baseline classification

We assign the TransmiR labels \( L(i) \) for each (TF, miRNA) pair \( i \) as follows, with the set of all TransmiR TFS \( T \), all TransmiR miRNAs \( M \), label 1 denoting positive, 0 denoting negative, and \( \emptyset \) as an unlabeled interaction:

\[
L(i) = \begin{cases} 
1 & \text{if } (\text{TF}, \text{miRNA}) \in \text{TransmiR pairs} \\
0 & \text{if } (\text{TF}, i) \in T \lor (\text{miRNA}, i) \in M \\
\emptyset & \text{otherwise} 
\end{cases}
\]

In other words, while no ‘negatives’ are reported in the TransmiR database (as is usually the case for interaction databases), we assign negative labels to TF–miRNA pairs where either the TF or the miRNA (but not both together) appear in one of the positive pairs. While most prior methods use a random set as negatives (since the vast majority of random pairs are not interacting, Qi et al., 2006), our negative labels reflect the fact that both the TF and the miRNA have been studied and since no interaction is reported for them they are more likely than a random pair to represent negative interaction. However, as we discuss below we do not treat this set as a strong negative and allow the classifiers we learn to flip some of these labels if their features indicate that they are interacting. Under this scheme, 266 interacting pairs are assigned positive labels, 188 923 are assigned negative labels, and the remaining 683 277 interacting pairs are not assigned a label.

Using the features derived from the genomic data sources, we tested a number of different classification and semi-supervised learning methods. We extracted the 266 validated interacting pairs from TransmiR and obtained the negative pairs for training by selecting a random subset of 798 (3 × 266) negatively-labeled interacting pairs.

Using these labels, we perform 5-fold cross-validation experiments for each of the classifiers we consider.

2.4 Co-training

In addition to using a standard classifier (Random Forest which was shown to be the best at dealing with skewed genomic interaction prediction tasks, Qi et al., 2006), we also used a well-known semi-supervised learning method, Co-Training (Blum and Mitchell, 1998) which was shown to be useful for predicting TF–gene interactions (Ernst et al., 2008). As mentioned in the Introduction, since the vast majority of the data is unlabeled, it may be beneficial to use both labeled and unlabeled data when learning a classifier. Co-training works by dividing the feature vector for each pair into two, preferably independent, subsets of features and training separate classifiers for each of the two subsets. Next, both classifiers are used to calculate predictions for each unlabeled sample. New labels are assigned to samples when both classifiers’ predictions agree.

We perform co-training in an iterative fashion, using random forests as the classification method used to train on each independent subset of features. After fitting both classifiers in each iteration, we obtain the prediction score for each unlabeled sample from the pair of classifiers, and use the minimum of the two scores to select the top 1000 positive predictions and bottom 1000 negative predictions. We assign new positive and negative labels to these interactions, and then start another iteration of the co-training process with the expanded set of labels (Note that as the co-training algorithm begins to converge, fewer than 1000 samples are assigned positive labels in successive iterations – we assign up to 1000 positive labels among those interactions with score above cutoff 0.5, and are always able to assign 1000 new negative labels due to the disparity in scores output by the paired classifiers.). We repeat this process until an iteration assigns no additional positive interactions. Note that after this co-training process is complete, many interactions still lack labels. As such, we learn the final random forest classifier (which uses all features) only on the set of labels assigned by the final iteration of the co-training procedure.

2.5 Network smoothing for semi-supervised label assignment

In addition to the random forest and co-training method, we developed a network propagation method following the approach of Vanunu et al. (2010), adjusted for semi-supervised label imputation. As described by Vanunu et al. (2010), this method operates on a network \( G = (V, E, w) \) and prior knowledge vector \( Y : V \rightarrow [0, 1] \), computing a function \( F(v) \forall v \in V \) that is both smooth over the network and accounts for the prior knowledge about each node. Although this method is typically used with prior knowledge vectors
having values in the range \([0, 1]\), this is not a requirement; the method converges with any vector \(Y : V \rightarrow C\) as an input.

In our setting, we treat the feature vector for each potential TF–miRNA interaction (defined in Section 2.2) as a point in \(n\)-dimensional space, and build a \(k\)-nearest neighbor network from these data points. In this network construction, we treat each TF–miRNA pair as a node and add edges between each node and the \(k\) nodes nearest to it, using Euclidean distance as the distance metric. The value of \(k\) is user defined though lower values of \(k\) produce networks with more disconnected components which is problematic in cases such as ours where only a small fraction of the data is positively labeled. We thus use \(k = 25\) in this paper, leading to a network with three connected components. Previous work (Shnaps et al., 2015; Ruffalo et al., 2015; Vanunu et al., 2010) has shown that similar network propagation methodology can be used with protein interaction networks to predict drug targets and genes associated with disease; here, we extend this method to a broader task of label imputation in the context of semi-supervised learning algorithms.

Given this \(k\)-NN network over the TF–miRNA interactions, we use Laplacian normalization to produce the normalized network edge weight \(w'\). Briefly, we construct a \(|V| \times |V|\) matrix \(W\) from the edge weights \(w\), and construct a diagonal matrix \(\Delta\) with \(\Delta[i, i] = \sum_{j} W[i, j]\). The normalized weight matrix is computed as \(W' = \Delta^{-1/2} W \Delta^{-1/2}\). This normalization procedure is especially valuable in networks with non-uniform degree distributions (e.g. scale-free networks such as protein interaction networks), since it provides some compensation for degree bias. Although our \(k\)-NN network is constructed to have uniform degree distribution, Laplacian normalization is still necessary to ensure that \(W'\) is similar to a stochastic matrix—namely, that the maximum eigenvalue of \(W'\) is bounded by \(|\lambda|_{\text{max}} \leq 1\).

With the normalized weight matrix \(W'\), we use the iterative procedure described by Zhou et al. (2004) to compute \(F\). Namely, starting with \(F^{(0)} = Y\), we update \(F\) at iteration \(t\) as follows:

\[
F^{(t)} = 2 W' F^{(t-1)} + (1 - 2) Y
\]

This procedure is repeated iteratively until convergence: until \(|F^{(t)} - F^{(t-1)}|_2 < \epsilon\) or \(|F^{(t)} - F^{(t-1)}|_1 < \epsilon\).

For each positive label \(L[i, f]\), we assign a positive score to \(F[i, f]\). For each negative label \(L[i, f]\), we assign a negative score to \(F[i, f]\). We then apply \(F\) to the network to produce networks with more disconnected components which is problematic in cases such as ours where only a small fraction of the data is positively labeled. We thus use \(k = 25\) in this paper, leading to a network with three connected components. Previous work (Shnaps et al., 2015; Ruffalo et al., 2015; Vanunu et al., 2010) has shown that similar network propagation methodology can be used with protein interaction networks to predict drug targets and genes associated with disease; here, we extend this method to a broader task of label imputation in the context of semi-supervised learning algorithms.

Figure 1, demonstrating the network smoothing process over simulated data.

3 Results

3.1 Baseline classification

Using the genomic data we collected, we assigned a feature vector to each TF–miRNA pair. We first trained several standard classifiers (Logistic Regression, Naive Bayes and Random Forest) using this feature vector and the labels extracted from TransmiR and performed cross-validation analysis to determine the accuracy of each of the classifiers. Supplementary Figure S1 shows 5-fold cross-validation results for this analysis. We see that logistic regression and Gaussian Naive Bayes show poor cross-validation performance, suggesting local behavior and statistical interdependence between features. In contrast, and as seen for other types of genomic interaction prediction tasks, Random Forest classifiers perform well, with cross-validation ROC AUC ranging from 0.88 to 0.97. As such, we focus on random forest classifiers as a baseline in successive experiments.

3.1.1 Semi-supervised methods

In addition to the standard methods, we have also tested two semi-supervised classification algorithms (Section 2). Figure 2 shows the
2009; Segal often regulated by the same TFs (as is the case for genes, Li et al., 2003; Zhang et al., 2005). Thus, a good prediction method is expected to lead to high correlation in the set of TFs controlling intergenic and genic miRNAs belonging to the same cluster.

We use the QuBiC biclustering method (Li et al., 2009) with default parameters to group miRNAs by expression, and for each bicluster, we determine the TFs binding to each miRNA using the aforementioned classification systems: RF with ‘baseline’ TransmiR labels, RF with co-trained labels, and propagated labels. We then count the occurrences of each TF binding to miRNAs in that bicluster, separately tracking the intergenic, genic-upstream-gene, and genic-upstream-miRNA binding sites. These binding count vectors are indexed by TF, and two example vectors are shown in Table 1.

Finally, we examine the correlation between intergenic versus genic-upstream-gene, and intergenic versus genic-upstream-miRNA, to determine which binding profile more closely matches the ‘complete’ intergenic signal.

For each bicluster, we compute the correlation between the TF vector counts for the genic and intergenic miRNA targets. As a control, we also perform permutation tests to evaluate the significance of the correlation values obtained from each scoring system. For each bicluster, we shuffle its miRNAs 100 times, preserving the proportion of the miRNAs that are genic. We then calculate new TF count vectors for each shuffled miRNA set, and repeat the correlation analysis described above. The results are presented in Figure 3. Biclusters on the x-axis are sorted by increasing P-value obtained from the QuBiC algorithm (such that the left most clusters are the most significant). As can be seen, both semi-supervised methods greatly outperform the baseline RF method. In all biclusters, the ‘baseline’ RF classifier described in Section 2.3 produces a very sparse signal, predicting few TFs to regulate each miRNA. In addition, for the most significant biclusters, the propagated labels defined in Section 2.5 produce the most regulatory consistency and results obtained using these predictions are distinctively better than the random assignment results. Specifically, for 8 of the top 10 biclusters, the propagated labels produce the highest score for prediction consistency.

We further compared the TF binding correlations between predictions based on the upstream of the gene and the upstream of the miRNA for genic miRNAs (Supplementary Fig. S3). As can be seen in Supplementary Figure S3, only predictions based on the upstream of the gene display the expected trend of decreasing correlation with less biclustering significance. The other setting (where features are extracted from the upstream of the miRNA itself) are very similar to random results. These results indicate that most genic miRNAs are likely regulated by as part of the gene they are encoded in.

### 3.3 Reconstructing cancer networks using predicted interactions

To test if the new set of predicted interactions can aid in the analysis and modeling of biological system regulation, we combine mRNA and miRNA expression data from TCGA glioblastoma multiforme data (GBM) (Brennan et al., 2013). GBM is the most common and most lethal form of brain cancer (Parsons et al., 2008), and analysis of associated omic data shows promise in producing novel treatments (Huang et al., 2015). We combined mRNA and miRNA expression data for 560 glioblastoma samples, across 17 814 genes and 534 miRNAs. We then filtered the gene selection, preserving the 3000 with the highest variance between samples. We use this combined expression data as an input to the cMonkey2 algorithm (Reiss et al., 2015). Solid lines show mean ROC, shaded areas show one standard deviation around this mean, and translucent lines show individual ROC curves for each validation run.
et al., 2015). cMonkey2 is a biclustering method that can automate many types of genomic analysis; notably, it can use interaction networks to inform the scoring of gene clusters. As such, we use cMonkey2 to investigate the clustering performance with multiple regulatory networks, to determine the additional insight gained with the full set of TF–miRNA interactions that we compute. For reconstructing the networks, in addition to using the TF–miRNA interactions predicted by our method we have also used TF–gene interactions from TRRUST (Han et al., 2015) and predicted miRNA–gene interactions from TargetScan v7.0 (Agarwal et al., 2015).

We next created four different types of interaction data-sets. The first, ‘NO’, did not contain any of the interaction data and only used the mRNA and miRNA expression data. The second ‘Gene Only’ contained, in addition to the expression data, TF–gene and miRNA–gene interactions (similar to the type of data used in Schulz et al., 2013). The third, ‘TransmiR’ contained all interactions contained in the second plus TF–miRNA interactions from TransmiR. Finally, the last ‘full’ contained all interactions from the third plus our new predicted TF–miRNA interactions.

We invoke cMonkey2 four times, once for each input type described above, producing four sets of gene/TF/miRNA clusters. We next performed Gene Ontology (GO) analysis on the clusters generated by using each of the data types. The results listing the top categories (in terms of $P$ value) are presented on the supporting website.

Given the large number of clusters produced for this data (over 300 for all input types), we have focused on the top-enriched categories, since these are the ones that researchers often follow up on (Huang et al., 2009). We thus compared the top 20 enriched categories (based on $P$ value enrichment) identified for each of the input types to determine their relevance to the cancer data being analyzed. While several relevant categories were enriched for all input types, we saw major differences in the $P$ values assigned to specific top categories when comparing the results of the different inputs. Specifically, as can be seen on the Supporting Website, immune response related categories (e.g. ‘immune process’ and ‘immune response’, $P$ values $10^{-12}$ and $10^{-14}$, respectively) were much more significant in the ‘full’ network when compared to all other types of input data. Other categories, including ‘cell death’ ($P$ value $10^{-10}$) were also most enriched in the ‘full’ results. Immune response has recently received much attention in cancer analysis and treatment (Havel and Chan, 2015; Tseng et al., 2015; Varn et al., 2016) and the ability to identify a more accurate set of genes, miRNAs and modules for such process is an important advantage of using the predicted interactions. The effects of such immune response can also explain the significance associated with cell death, another hallmark of cellular response to cancer (Clarkson et al., 2004; Lu et al., 2008; Portet et al., 2011). Several other categories related to cell cycle and cell division were also enriched in the clusters identified for ‘full’ (although many were also highly enriched when using the other types of inputs). Figure 4 presents four of the modules identified when using the complete set of predicted interactions. All these networks contain both genes and miRNAs and are associated with GO terms related to glioblastoma. Figure 4a shows a subnetwork enriched for cell proliferation, Figure 4b shows a subnetwork enriched for tissue morphogenesis, Figure 4c shows a subnetwork enriched for regulation of cell death and Figure 4d shows a subnetwork enriched for serine binding. These subnetworks contain several genes known to be involved in cancer progression. These include TGFx (Fig. 4a), which has been shown to be active in several cancers (Bates et al., 1988; Moskal et al., 1995; Wilding et al., 1989), HOXB7 which has been used to predict survival in pancreatic adenocarcinoma (Chile et al., 2013; Nguyen Kovichich et al., 2013), and FOXF2 whose downregulation has been linked to poor prognosis in prostate and breast cancers (Hirata et al., 2013; Kong et al., 2012). Other genes have been linked specifically to brain function and tumors, including NEUROD1 (Fig. 4b) which is known to promote tumor formation in neuroblastoma (Huang et al., 2011), and INSMI which is involved in neuroendocrine differentiation (Lan and Breslin, 2009) and NEUROG2 (Fig. 4d) which is involved in neural tube and spinal cord development (Henke et al., 2009; Ichi et al., 2010). EGR2 (Fig. 4c) plays a significant role in PTEN-induced apoptosis (Unoki and Nakamura, 2003) and its downregulation by miR-130 promotes gastric cancer proliferation (Wu et al., 2010). Several of the miRNAs contained in these networks have also been implicated in cancers. Of these, miR-490 (Fig. 4d) has been identified as a regulator of epithelial to mesenchymal transition (EMT) (Zhang et al., 2013), which is a key process in cancer metastasis, and miR-506 (Fig. 4d) specifically regulates EMT in breast cancer (Avora et al., 2013). Hypomethylation and overexpression of miR-200a (Fig. 4d) have been linked to pancreatic cancer development (Li et al., 2010), and miR-187 (Fig. 4d) has been used as a prognostic biomarker in both breast (Mulrane et al., 2012) and prostate (Casanova-Salas et al., 2014) cancers. The methylation status of miR-9 (Fig. 4d) has known associations with renal cell
carcinoma development and metastasis (Hildebrandt et al., 2010) and breast cancer development (Lehmann et al., 2008). miR-551b (Fig. 4a) has been identified as part of a signaling pathway that contributes to acquired resistance to apoptosis and chemotherapy (Xu et al., 2014).

4 Discussion

Identification of regulatory relationships is a key component in understanding complex biological processes such as tissue development and disease progression. In this work, we have used semi-supervised learning techniques to expand the high-confidence TransmiR TF–miRNA interaction. We show that a network smoothing-based approach achieves good performance on the (very small) set of labeled data as well as on a more general co-expression dataset. Using this method, we created a genome-wide set of predicted TF–miRNA interactions, encompassing 3782 interactions between 191 distinct TFs and 1800 distinct miRNAs. We have used the predicted interactions by combining them with other genomic and interactions datasets to analyze cancer data. Using the full set of predicted interactions will prove beneficial for systems biology studies that are aiming to reconstruct regulatory networks for biological processes.

The full list of predicted interactions is available from the supporting website. Based on our results, we believe that using the full set of predicted interactions will prove beneficial for systems biology studies that are aiming to reconstruct regulatory networks for biological processes.

Funding

This work was supported in part by the National Science Foundation (Grant no. DBI-1356505 to Z.B.J.) and by the Pennsylvania Department of Health (Health Research Nonformula Grant (CURE) Awards to Z.B.J.).

This project is funded, in part, under a grant (#4100070287) with the Pennsylvania Department of Health. The Department specifically disclaims responsibility for any analyses, interpretations or conclusions.

Conflict of Interest: none declared.

References

Agarwal, V. et al. (2015) Predicting effective miRNA target sites in mammalian mRNAs. Elife, 4, e05005.

Arora, H. et al. (2013) mir-506 regulates epithelial mesenchymal transition in breast cancer cell lines. PLoS One, 8, e64273.

Banovich, N.E. et al. (2014) Methylation qtls are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. PLoS Genet., 10, e1004663.

Bartel, D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. Cell, 116, 281–297.

Bates, S.E. et al. (1988) Expression of transforming growth factor α and its messenger ribonucleic acid in human breast cancer: its regulation by estrogen and its possible functional significance. Mol. Endocrinol., 2, 543–555.

Blum, A. and Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In: Proceedings of the Eleventh Annual Conference on Computational Learning Theory. ACM, New York, pp. 92–100.

Bornsman, A.R. et al. (2007) Divergence of transcription factor binding sites across related yeast species. Science, 317, 815–819.

Boulesteix, A.L. and Strimmer, K. (2005) Predicting transcription factor activities from combined analysis of microarray and chip data: a partial least squares approach. Theor. Biol. Med. Model., 2, 23.

Brennan, C.W. et al. (2013) The somatic genomic landscape of glioblastoma. Cell, 155, 462–477. 2412042[pmid].

Bulyk, M.L. et al. (2004) Computational prediction of transcription-factor binding site locations. Genome Biol., 5, 201–201.
Casanova-Salas,J. et al. (2014) Identification of mir-187 and mir-182 as biomarkers of early diagnosis and prognosis in patients with prostate cancer treated with radical prostatectomy. J. Urol., 192, 252–259.

Chile,T. et al. (2013) HOXB7 mRNA is overexpressed in pancreatic ductal adenocarcinomas and its knockdown induces cell cycle arrest and apoptosis. BMC Cancer, 13, 451.

Clarkson,R. et al. (2004) Gene expression profiling of mammary gland development reveals putative roles for death receptors and immune mediators in post-lactational regression. Breast Cancer Res., 6, R92–R109.

Cock,P.J. et al. (2009) Biopython: freely available python tools for computational molecular biology and bioinformatics. Bioinformatics, 25, 1422–1423.

Consortium,E.P. et al. (2004) The encode (encyclopedia of dna elements) project. Science, 306, 636–640.

Coronnello,C. and Benos,P.V. (2013) Comir: combinatorial microrna target prediction tool. Nucleic Acids Res., 41, W159–W164.

Ernst,J. et al. (2008) A semi-supervised method for predicting transcription factor–gene interactions in escherichia coli. PLoS Comput. Biol., 4, e1000044.

Ernst,J. et al. (2010) Integrating multiple evidence sources to predict transcription factor binding in the human genome. Genome Res., 20, 526–536.

Griffiths-Jones,S. et al. (2006) mirbase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res., 34, D140–D144.

Han,H. et al. (2015) Trust: a reference database of human transcriptional regulatory interactions. Sci. Rep., 5, 11432.

Havel,J.J. and Chan,T.A. (2015) High-resolution genomic analysis: the tumor–immune interface comes into focus. Genome Biol., 16, 1–3.

Hemberg,M. and Kreiman,G. (2011) Conservation of transcription factor binding events predicts gene expression across species. Nucleic Acids Res., 39, 7092–7102.

Henke,R.M. et al. (2009) Neurog2 is a direct downstream target of the ptf1a-rbpj transcription complex in dorsal spinal cord. Development, 136, 2945–2954.

Hildebrandt,M. et al. (2010) Hsa-mir-9 methylation status is associated with cancer development and metastatic recurrence in patients with clear cell renal cell carcinoma. Oncogene, 29, 5724–5728.

Hinrichs,A.S. et al. (2006) The ucsc genome browser database: update 2006. Nucleic Acids Res., 34, D590–D598.

Hirata,H. et al. (2013) Microrna-182-5p promotes cell invasion and proliferation by down regulating fox2, reek and mst1 genes in human prostate cancer. PloS One, 8, e55502.

Huang,D.W. et al. (2009) Bioinformatic enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res., 37, 1–13.

Huang,P. et al. (2011) The neuronal differentiatation factor neuron1 downregulates the neuronal repellent factor slat2 expression and promotes cell motility and tumor formation of neuroblastoma. Cancer Res., 71, 2938–2948.

Huang,X. et al. (2015) Targeted multi-epitope therapeutic vaccine for the treatment of invasive glioblastoma multiforme. J. Immunother. Cancer, 3, P436.

Ichis,S. et al. (2010) Folic acid remodels chromatin on hes1 and neurog2 promoters during caudal neural tube development. J. Biol. Chem., 285, 36922–36932.

John,S. et al. (2011) Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. Nat. Genet., 43, 264–268.

Kaplan,T. et al. (2003) Ab initio prediction of transcription factor targets using structural knowledge. PLoS Comput. Biol., 1, e1.

Karin,M. (1990) Too many transcription factors: positive and negative interactions. New Biol., 2, 126–131.

Kong,P.Z. et al. (2012) Decreased fox2 mRNA expression indicates early-onset metastasis and poor prognosis for breast cancer patients with histological grade ii tumors. PloS One, 8, e61591–e61591.

Kummerfeld,S.K. and Teichmann,S.A. (2006) Dbs: a transcription factor prediction database. Nucleic Acids Res., 34, D74–D81.

Lan,M.S. and Breslin,M.B. (2009) Structure, expression, and biological function of ins1 transcription factor in neuroendocrine differentiation. FASEB J., 23, 2024–2033.
Thurman, R.E. et al. (2012) The accessible chromatin landscape of the human genome. *Nature*, 489, 75–82.

Tseng, W.W. et al. (2015) Analysis of the intratumoral adaptive immune response in well differentiated and dedifferentiated retroperitoneal liposarcoma. *Sarcoma*, 2015, 547460–25705114. [pmid].

Unoki, M. and Nakamura, Y. (2003) Egr2 induces apoptosis in various cancer cell lines by direct transactivation of bnp3l and bak. *Oncogene*, 22, 2172–2185.

Vanunu, O. et al. (2010) Associating genes and protein complexes with disease via network propagation. *PLoS Comput. Biol.*, 6, e1000641.

Varm, F.S. et al. (2016) Integrative analysis of breast cancer reveals prognostic haematopoietic activity and patient-specific immune response profiles. *Nat. Commun.*, 7, Article.

Wilding, G. et al. (1989) Role of transforming growth factor-α in human prostate cancer cell growth. *Prostate*, 15, 1–12.

Wu, Q. et al. (2010) Mir-150 promotes gastric cancer proliferation by negatively regulating the pro-apoptotic gene egr2. *Biochem. Biophys. Res. Commun.*, 392, 340–345.

Xu, X. et al. (2014) A signaling pathway consisting of mir-551b, catalase and muc1 contributes to acquired apoptosis resistance and chemoresistance. *Carcinogenesis*, 35, 2457–2466.

Zhang, L. et al. (2013) mir-490-3p modulates cell growth and epithelial to mesenchymal transition of hepatocellular carcinoma cells by targeting endoplasmic reticulum–golgi intermediate compartment protein 3 (ergic3). *J. Biol. Chem.*, 288, 4035–4047.

Zhao, Y. et al. (2005). A time-series biclustering algorithm for revealing co-regulated genes. In *International Conference on Information Technology: Coding and Computing, 2005. itcc 2005*, vol. 1, IEEE, New York, pp. 32–37.

Zhong, S. et al. (2013) Predicting tissue specific transcription factor binding sites. *BMC Genomics*, 14, 796.

Zhou, D. et al. (2004) Learning with local and global consistency. *Adv. Neural Inf. Process. Syst.*, 16, 321–328.

Zhu, X. (2011). Semi-supervised learning. In: *Encyclopedia of Machine Learning*. Springer, New York, pp. 892–897.