Multitask regression for condition-specific prioritization of miRNA targets in transcripts

Azim Dehghani Amirabad[1,2,3] **and Marcel H. Schulz**[1,2]

[1] **Cluster of Excellence for Multimodal Computing and Interaction, Saarland University, Saarbrücken**

[2] **Max Planck Institute for Informatics, Saarbrücken**

[3] **International Max Planck Research School for Computer Science, Saarbrücken**

ABSTRACT

Deregulation of miRNAs is implicated in many diseases in particular cancer, where miRNAs can act as tumour suppressors or oncogenes. As sequence-based miRNA target predictions do not provide condition-specific information, many algorithms combine expression data for miRNAs and genes for prioritization of miRNA targets. However, common strategies prioritize miRNA-gene associations, although a miRNA may only target a subset of the alternative transcripts produced by a gene. Thus, current approaches are suboptimal. Here we address the problem of transcript and not gene based miRNA target prioritization. We show how to leverage methods that were developed for gene expression based miRNA-target prioritization for transcripts. In addition, we introduce a new multitasking based learning (MTL) method that uses structured-sparsity inducing regularization to improve accuracy of the learning. The new MTL approach performs especially favorable in small sample size settings, for genes with many transcripts and with noisy transcript expression level estimates as shown with simulated data. In an analysis of real liver cancer RNA-seq data we show that the MTL approach better predicts transcript expression and outperforms simpler approaches for miRNA-target prediction.

Keywords: Multi-task regression, RNA-seq, microRNA target prediction, Regulatory networks

1 INTRODUCTION

microRNAs (miRNAs) are an abundant class of small regulatory RNAs that are involved in the post-transcriptional regulation of the majority of human genes (Friedman *et. al*, 2009). Deregulation of miRNAs is implicated in many diseases in particular cancer (Esquela-Kerscher and Slack, 2006), where miRNAs can act as tumour suppressors or oncogenes. Therefore understanding the relevant targets of (deregulated) miRNAs has become an active field of research.

Initial methods for miRNA-target predictions based on seed-sequence complementarity, sequence conservation, and thermodynamic considerations (Friedman *et. al*, 2009; Krek *et. al*, 2005; John *et. al*, 2004) lead to *static* and not *condition-specific* predictions.

A common strategy for prioritization of miRNA targets, in different cancers for example, is to augment static sequence-based predictions with condition-specific RNA expression data for miRNAs and genes (Huang et al., 2007; Lu *et. al*, 2011; Engelmann and Spang, 2012; Muniategui et al., 2013; Schulz *et. al*, 2013; Jacobsen *et. al*, 2013). Most of these prioritization approaches allow to rank miRNA-target interactions according to some measure of condition-specific regulatory strength, which is a useful information for designing downstream validiation experiments. While successful, previous methods have used gene expression and not transcript expression estimates for this prediction. Gene expression denotes the summed activity of all transcripts of a gene, and a gene may have a number of transcripts due to alternative splicing or alternative polyadenylation. However, miRNAs bind to the mRNA molecules (transcripts) and therefore transcript annotation should be taken into account.

As Fig. 1 illustrates, aggregation at the gene level may lead to a loss of information and blur the effect of miRNA regulation, with possible false positive (FP) or false negative (FN) miRNA associations.

In a pioneering study by Deng *et. al* (Deng *et. al*, 2011), who used Mutu I cells that overex-

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2377v1 | CC BY 4.0 Open Access | rec: 19 Aug 2016, publ: 19 Aug 2016

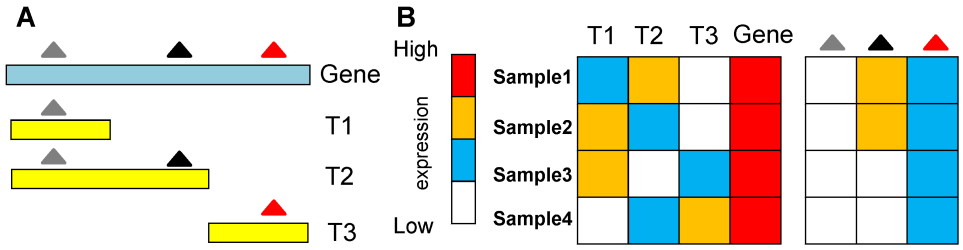

Figure 1. A) Example of three transcripts (T1,T2,T3) of the same gene, that differ in their 3'UTR sequence (yellow boxes) and are bound by different miRNAs (coloured triangles) in a condition-specific manner. Common approaches project miRNA binding sites in transcripts to gene level (top row). B) Hypothetical expression values for transcripts and gene (left matrix) and the miRNAs (right matrix) for four samples (rows). Common miRNA prioritization approaches search for negative association of miRNA expression with gene expression. Gene expression, obtained by summing expression of all transcripts, removes variability of transcript expression. For example, expression of the red miRNA is negatively correlated with gene expression, although the expression of T3, the target of the red miRNA does not show strong condition-specific negative correlation.

pressed miR-155, it was shown that a simple transcript expression cutoff and seed enrichment strategy revealed more true miR-155 targets compared to using a cutoff on gene expression only. They had used RNA-seq data of Mutu I cells to estimate transcript expression levels using an Expectation Maximization (EM) algorithm. Over the last years a number of EM-related approaches were introduced that allow the quantification of transcript expression levels from RNA-seq data (Richard et al., 2010; Li and Dewey, 2011; Patro et al., 2014). Therefore, we believe there is an opportunity to develop methods that allow to prioritize miRNA interactions on the level of individual transcripts, in particular in the light of a large number of available paired miRNA and mRNA-seq cancer expression data sets, *e.g.* from The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013).

We show how to leverage methods that were developed for gene expression based miRNA-target prioritization for transcripts. In addition, we introduce a new multitasking based learning (MTL) method that uses structured-sparsity induced regularization to improve learning accuracy. As miRNA binding sites can be shared between different transcripts of a gene (see Fig. 1A, grey triangle miRNA) our MTL method borrows information from several transcripts to significantly reduce estimation error. We investigate the difficulties of miRNA-target prioritization for transcripts using different simulation studies and analysis of liver cancer RNA-seq data. We show that the new MTL approach is able to outcompete simpler approaches and leads to the most accurate prediction of experimentally validated miRNA targets.

2 METHODS

2.1 A multi-task formulation for miRNA-transcript regression

Let us assume that data are collected for N RNA-seq samples of M miRNAs and T transcripts.

Let $\mathbf{X} \in \mathbb{R}^{NxM}$ be the expression matrix for M miRNA expression levels for all N samples and $\mathbf{Y} \in \mathbb{R}^{NxT}$ be the expression matrix of the T (output) transcripts. We assume a multivariate linear regression model for each transcript as follow:

$$\mathbf{y}_t = \mathbf{X}\beta_t + \varepsilon_t, \quad \forall t = 1, \ldots, T, \tag{1}$$

where $\beta_t = [\beta_{1t}, \ldots, \beta_{Mt}]$ is the regression coefficient vector of miRNAs for each transcript with ε_t Gaussian noise. We further assume that $X_{i,j}$'s are standardized and consider model without intercept.

Lasso and multi-task lasso with structured sparsity-inducing penalty

The least absolute shrinkage and selection operator (Lasso) is a sparse linear regression model for inferring the regression coefficients β and widely used in bioinformatics. The simplest approach would be to learn T separate sparse linear regression models by optimizing the lasso objective function for each transcript as follow:

$$\min_{\beta_t \in \mathbb{R}^{1xM}} f(\beta_t) \equiv \frac{1}{2}||y_t - X\beta_t||_F^2 + \lambda||\beta_t||_1 \,, \tag{2}$$

where $||\cdot||_F$ denotes the Frobenius norm, an Euclidean-based matrix norm, $||\cdot||_1$ is the L_1 norm. λ controls the sparsity in β. We call this the *disjoint* estimation method, because all miRNA associations are learned independently for each transcript.

Transcripts from the same gene may share miRNA binding sites because they have similar 3'-UTRs. In that case the lasso estimation (Eq. 2) selects relevant miRNAs for each transcript separately and does not encourage joint selection of common miRNAs. This may lead to suboptimal inference and instability in miRNA feature selection.

In order to capture the shared sparse pattern among multiple related tasks (here transcripts of a given gene), we formulate transcript expression level prediction as a multi-task regression problem (Chen et al., 2012).

Let $\mathbf{B} = [\beta_1, \ldots, \beta_T] \in \mathbb{R}^{MxT}$ denote the regression coefficient matrix for all transcripts.

We can solve the following optimization problem:

$$\min_{B \in \mathbb{R}^{MxT}} f(B) \equiv \frac{1}{2}||Y - XB||_F^2 + \lambda||B||_1 + \Omega(B) \,, \tag{3}$$

where $\Omega(B)$ is the structured sparsity inducing penalty over the input miRNA features. Different to Equation 2, we define a structured sparsity inducing penalty for miRNA binding sites that are shared between two or more transcripts of the same gene. This translates to the non-overlapping group lasso penalty, which in our setup is defined as :

$$\Omega(B) = \sum_{m=1}^{M} \gamma(||\beta_{mg}^{+}||_2) \,, \tag{4}$$

where γ and λ are the model hyper-parameters of Eq. 3. Denote as g_m^+ the set of transcripts of gene g that contain a binding site for miRNA m. In addition denote as $\beta_{mg}^{+} = \{\beta_{mt} | t \in g_m^+\}$ the coefficient vector of all transcripts of gene g that contain a binding site for miRNA m. The biological motivation to use the non-overlapping group lasso penalty in Eq. 4 is that we assume a similar regulatory effect of the miRNA for the transcripts of the same gene, because the miRNA binding site is identical and the local secondary structure of these transcripts, and hence target site accessibility, is likely to be similar. We used the CVX library (version 2.1 (Grant and Boyd, 2014)) for optimization. In order to optimize parameters, we partition the data three times: into training, validation and test data. The optimal hyper parameter values for γ and λ are first selected by minimizing the six-fold cross-validation error as estimated on the validation data using a fixed parameter grid applied to the training samples. Then a second local optimization around the best parameter pair from the first step is conducted. The final model performance is estimated on the test samples, that were not used for learning and parameter optimization.

2.2 Simulation setup

The main assumption for creating the synthetic data is that miRNA binding sites in different transcripts of the same gene, have the same regulatory strength.

We simulated miRNA expression levels for 50 miRNAs by sampling from a univariate Gaussian distribution for each miRNA. Mean and standard deviation of miRNA expression distribution is sampled from $\mu \sim U(0.1, 2)$ and $\sigma \sim U(0.01, 1)$, respectively. Then we sampled the miRNA regulatory effect vector (β) from $\beta \sim U(-0.5, 0.5)$. We created sparse miRNA transcript binding matrix such that transcripts of the gene share at least 50% of miRNA binding sites with each others. For the simulation for every task (transcript) we defined 6 true miRNA binding events (cf. Fig. 2A) and 42 miRNAs as false positive associations. Unless otherwise stated we simulated 35 samples in each setup.

We simulated transcript expression levels using the following linear model:

$$\mathbf{y}_t = \mathbf{X}\beta_t + \varepsilon_t, \quad \forall t = 1, \ldots, 16 \,, \tag{5}$$

where β_t is vector of coefficient for miRNAs that have binding sites at transcript t. For each sampled transcript expression level we add random noise $\varepsilon \sim N(0, 0.5)$. Both miRNA and transcript expression levels have been centered and we considered a model without intercept.

We simulated synthetic data for three different setups with the above setting

(I): We created 5 different simulation setups with 2, 6, 8, 10, and 16 transcripts using Eq. 5 with noise distribution $\varepsilon \sim N(0, 0.5)$.

(II): We created 6 simulations with different noise levels using 16 transcripts with Eq. 5 and noise

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2377v1 | CC BY 4.0 Open Access | rec: 19 Aug 2016, publ: 19 Aug 2016

levels with $\mu = 0$ and $\sigma = 0.15$, 0.5,0.75, 1.0, 1.5, and 2.0 respectively.
(III): We created 5 different simulation for each setup described in (I), so in total 25 simulations. We wanted to evaluate the performance of our algorithms in different settings of dimensionality, with 15, 25, 35, 45, and 50 training samples.

Given the simulated expression level $\mathbf{y}_t$ we measure the root mean square error (RMSE) between predicted and test-set expression levels $\hat{\mathbf{y}}_t$ as follows:

$$RMSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2\,, \tag{6}$$

where N is the number of test data samples (N=500).

Given true positive and true negative miRNA binding events from simulated data, we normalized the predicted coefficients to the interval [0,1]. Then we used R package pROC (version 1.8., Robin et al. (2011)) to compute Area Under the Curve (AUC) values for ROC curves of the models for each simulation setup.

2.3 Preprocessing miRNA input and gold standard

Potential miRNA-transcript interactions are retrieved from TargetScan (v. 7.0) (Agarwal *et. al*, 2015). We used miRTarBase release 6.0 (Chou *et. al*, 2015) to obtain experimentally validated miRNA target interactions. We downloaded RSEM-quantified gene and transcript expression levels (Li and Dewey, 2011) and quantified miRNA expression values (RPKM) from liver cancer RNA-seq data from TCGA (http://cancergenome.nih.gov/).

We filtered all miRNAs, transcripts and genes, if their expression was < 2 RPKM in 85% in all samples. After filtering, we were left with 461 mature expressed miRNAs. We used 150 paired miRNA and transcript or gene expression estimates for model training and additional data for testing. All RPKM values were log_2 transformed and standardized before learning. For comparison with real data only matching gold standard interactions of expressed miRNAs that have a TargetScan site were selected.

3 RESULTS

In this section we compare the performance of our new algorithms on simulated and liver cancer expression data to show that transcript based miRNA prioritization works well.

3.1 Evaluation with synthetic data

In order to assess the problem of transcript expression based inference of miRNA regulation we created a set of simulated datasets that allow us to explore the performance of the different approaches, see methods. In particular we are interested to test our hypothesis that the MTL approach can outcompete the disjoint model by borrowing information from several transcripts.

Effect of different number of transcripts on model performance

In the first setup we vary the number of transcripts (2,6,8,10,16) using the ground truth miRNA interaction matrix as shown in Fig. 2A, by taking different subsets of this matrix (columns from right to left). 8 miRNAs with different interaction strength are simulated, such that the interaction coefficient is the same among all regulated transcripts. 4 of the miRNAs are shared among all 16 transcripts and the other 4 are shared between 10 and 6 transcripts, respectively.

In Fig. 2B we show the normalized coefficient matrix for estimation with all 16 transcripts. Both the MTL and the disjoint model recover most of the true miRNA interactions, but select a number of false positive interactions. In this simulation $\sim 80\%$ of the interactions in the ground truth set are FP interactions, which makes it a challenging problem, resembling the amount of FP interactions when relying on static sequencing based prediction methods. We noticed that the MTL method shows overall more low intensity FP coefficients compared to the disjoint model, albeit at higher frequency. We believe that this is an advantage as these low intensity coefficients are easy to filter. However, the disjoint model has a number of non-zero interactions that are false positives, but have equally high values compared to TP interactions.

Figure 3A shows the RMSE (Eq. 6) on test samples for all setups and compares the disjoint with the MTL model. Generally, the higher the number of coupled transcripts per miRNA binding site, the smaller the error for the MTL method. The error for MTL converges at 10 transcripts with no further improvement with 16 transcripts. In contrast, the disjoint model shows increased error with more than 10 transcripts. This is due to our simulation setup, where the 10 and 16 transcript problems have more miRNAs with smaller true binding coefficients, that are harder to estimate correctly (cf. Fig.2A).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2377v1 | CC BY 4.0 Open Access | rec: 19 Aug 2016, publ: 19 Aug 2016

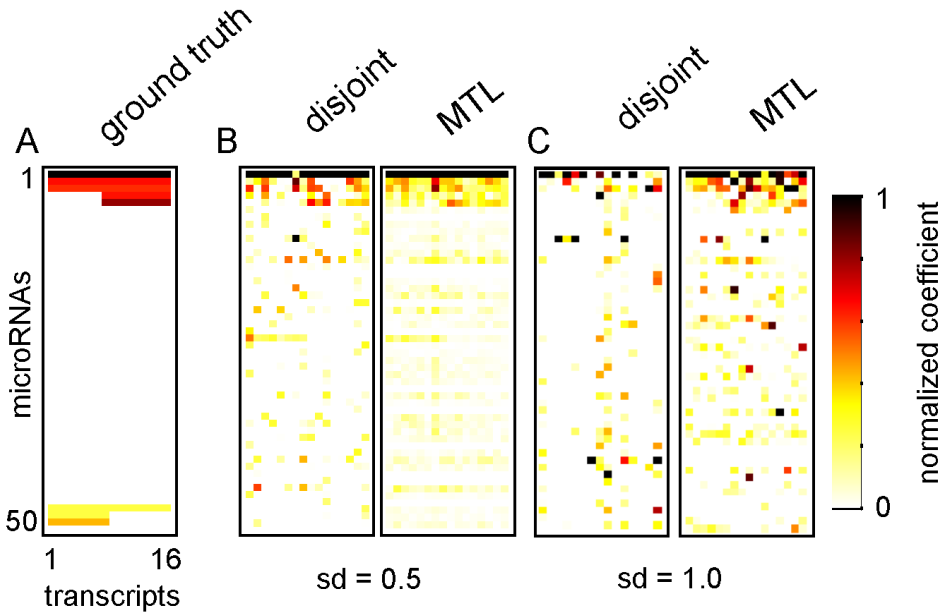

Figure 2. Comparison of coefficient heatmaps for the 16 transcript simulation. (A) Ground truth values of miRNA binding strength for the complete set of 16 transcripts. Out of 50 miRNAs (rows) each transcript (columns) has 6 miRNAs that are associated with it, visible as a colored spot in the heat map. The rows of the heat map are clustered using hierarchical clustering and group miRNAs that show similar associations in the simulation. (B) Estimated regression coefficients with the disjoint and MTL methods on the simulation dataset with $\sigma = 0.5$ for transcript expression levels. (C) Same as (B) with $\sigma = 1.0$. All absolute values of simulated and estimated coefficients are normalized to [0,1].

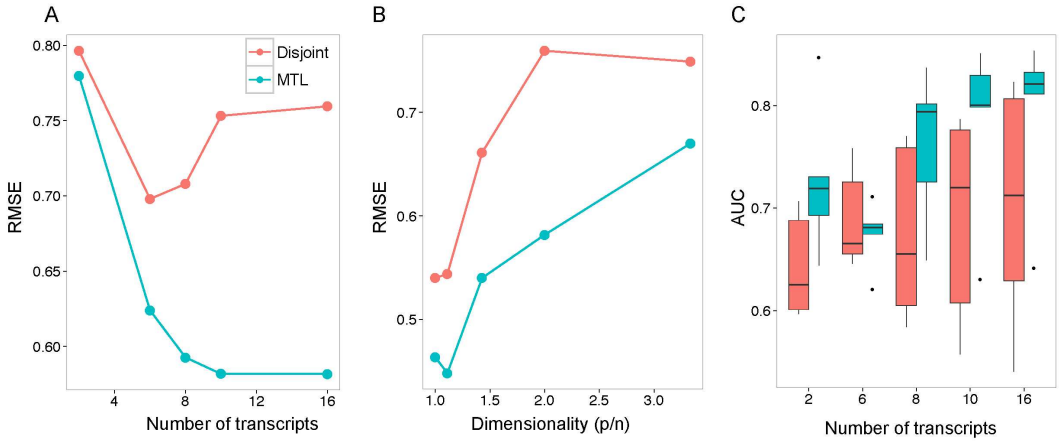

Figure 3. Comparison of the disjoint and MTL learning on different simulated datasets. Comparison of test RMSE (y-axis) obtained for an increasing number of transcripts per gene (A) or increasing dimensionality of the problem (ratio of features to samples) (B). (C) Distribution of AUC values for all simulations in setup (III), see Methods. The higher the number of transcripts that share a binding site, the more the MTL method outperforms the disjoint model.

Learning performance in a high dimensionality setting

Another important aspect of learning sparse models is the dimensionality (p/n) of the problem, where p denotes the number of parameters and n denotes the number of samples. Then if $p \gg n$

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2377v1 | CC BY 4.0 Open Access | rec: 19 Aug 2016, publ: 19 Aug 2016

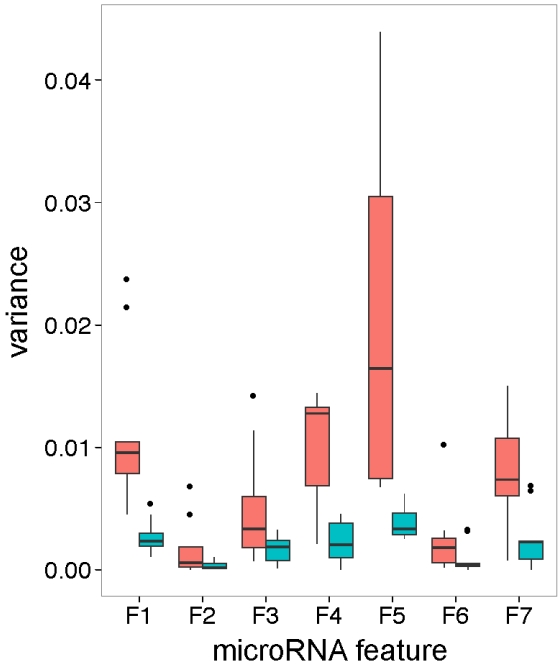

Figure 4. Variance in miRNA binding strength estimation. Variance of estimated feature coefficients compared to the ground truth value from simulations (y-axis) over all simulation setups and for each miRNA feature and method (x-axis), same colors as in Fig. 3. MTL shows consistently smaller variance among all simulation setups.

the problem is called high dimensional. We conducted additional simulations for learning with increased dimensionality (simulation setup III), as occurs often in practice when a limited number of paired miRNA and gene expression samples are available. In Fig. 3B we show the obtained error for the disjoint and MTL model. With higher dimensionality the error for both models increases as expected. For all values tested the MTL model shows smaller errors compared to the disjoint model. The disjoint model is more prone to overfitting in high dimensions.

While smaller RMSE values show, that the MTL model predicts true expression levels better, we used a ROC analysis in addition to evaluate the performance of true miRNA binding predictions (see Methods). In Figure 3C we show the distribution of AUC values for different ranges of problem dimensionality, stratified by the number of transcripts. Overall, we observed that for the MTL model, when the number of tasks increases the AUC values increases, with values up to 0.85 for 16 transcripts. Also the MTL model always shows higher AUC values and lower variance in the AUC compared to the disjoint model.

We further investigated the variance in the coefficient estimates in Fig. 4, studying all simulations (setup III). All 7 simulated miRNAs (features) that were consistently selected as non-zero by both models are shown. Note that the disjoint method failed to select one of the miRNA features most of the time and therefore we excluded it from this analysis. Over all feature values the variance of the MTL method is much smaller compared to the disjoint method. Features that are shared between all 16 transcripts show the largest reduction in variance.

Thus, if the assumption of shared miRNA binding strength for transcripts with the same site is true, then using MTL leads to better performance and can be trained with fewer samples.

Robustness to different noise levels

Multiple transcripts of a gene may share very similar exonic structures. This can potentially make it hard to discern the origin of the read and lead to noisy estimation of transcript expression levels from RNA-seq data (Teng et al., 2016). In order to prioritize miRNA target at transcript level, the algorithm should be robust to high noise levels in transcript expression data.

We evaluated performance of our algorithms on simulations (setup II) with different noise levels of transcript expression estimation. Figure 5A shows that the RMSE with both models increases with higher noise levels and that the MTL model outperforms the disjoint model, in particular with

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2377v1 | CC BY 4.0 Open Access | rec: 19 Aug 2016, publ: 19 Aug 2016

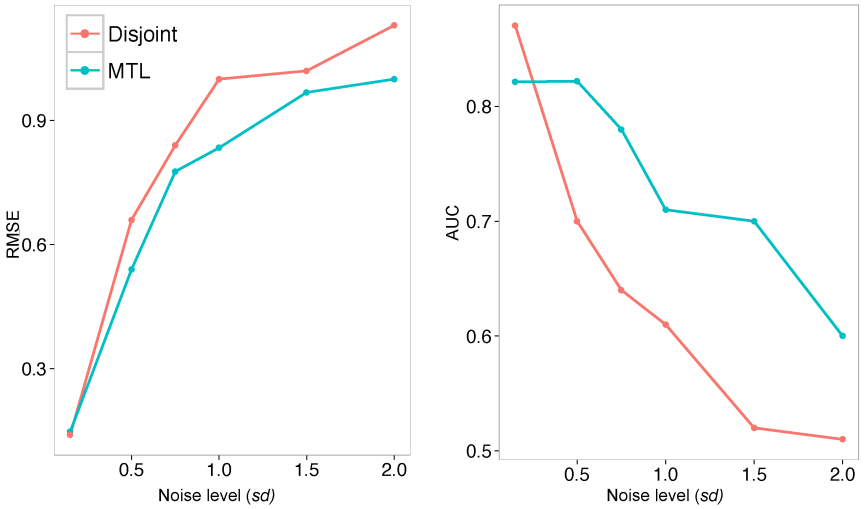

Figure 5. Impact of noise in transcript expression estimation on miRNA target prediction. Comparison of the MTL (blue) and disjoint (red) model in simulations with different noise for transcript expression levels (x-axis), see simulation setup (II). (A) shows obtained RMSE values (y-axis) and (B) AUC values (y-axis).

higher noise levels. Similarly, AUC values for high noise levels drop, but the MTL method tolerates noise much better than the disjoint method and leads to more accurate miRNA binding prediction (Fig. 5B). We conclude that accurate miRNA target predictions can be made for transcripts, even at high noise levels.

3.2 miRNA target prioritization in liver cancer

We used available TCGA liver cancer data to analyze the performance of our new formulations, see Methods. We retrieved all genes that had at least two expressed transcripts, such that the transcripts also had known miRNA-gene interactions in the miRTarBase database (Chou *et. al*, 2015) and had the same 3'-UTR. Using TargetScan 7.0 we build the feature matrix for each transcript, see Methods. We first analysed the MTL and disjoint method for prediction of miRNA-targets in transcripts and then compared to a standard approach that works on the gene and not the transcript level.

We compared the performance of MTL and the disjoint method for miRNA-target prediction in 114 transcripts, from 48 and 6 genes with 2 and 3 expressed transcripts, respectively. Figure 6A compares the error obtained on test samples with both approaches. We observed that the error for the disjoint model is higher for most of the transcripts, similar as in the simulated data examples. Similarly the observed correlation of predicted expression values for MTL was higher than for the disjoint approach (Fig. 6B) and the MTL models capture much more of the observed expression behavior of individual transcripts in liver cancer.

In order to test the ability of the methods to prioritize true miRNA interactions, we used Recall-Rank plots using experimentally validated miRNA interactions, to account for limited gold standard data. In the first analysis we compared the MTL and the disjoint method in their ability to rank validated miRNA interactions. After fitting models to transcripts, all predicted miRNA-transcript interactions are pooled together and ranked by their absolute regression coefficient. Then at each threshold, recall is computed as the fraction of retrieved experimentally validated interactions divided by all experimentally validated interactions for all involved miRNAs and transcripts (Chou *et. al*, 2015). As shown in Fig. 7A the MTL method is able to rank more true miRNA interactions in the top ranks and also detects a higher number of true interactions among all non-zero coefficients. This suggests that the improved expression prediction performance (cf. Fig. 6) is the result of more meaningful feature selection.

Finally, we compare transcript level methods against the current standard for analyzing RNA-seq data, where the expression of all transcripts of a gene is summed and one model is used for inference. For this gene-based learning we use lasso with gene expression values as response (Lu *et. al*, 2011) and with miRNA features annotated on the gene level as illustrated in Fig. 1. In order to compare the transcript models with the gene-based model we compute one gene level miRNA coefficient

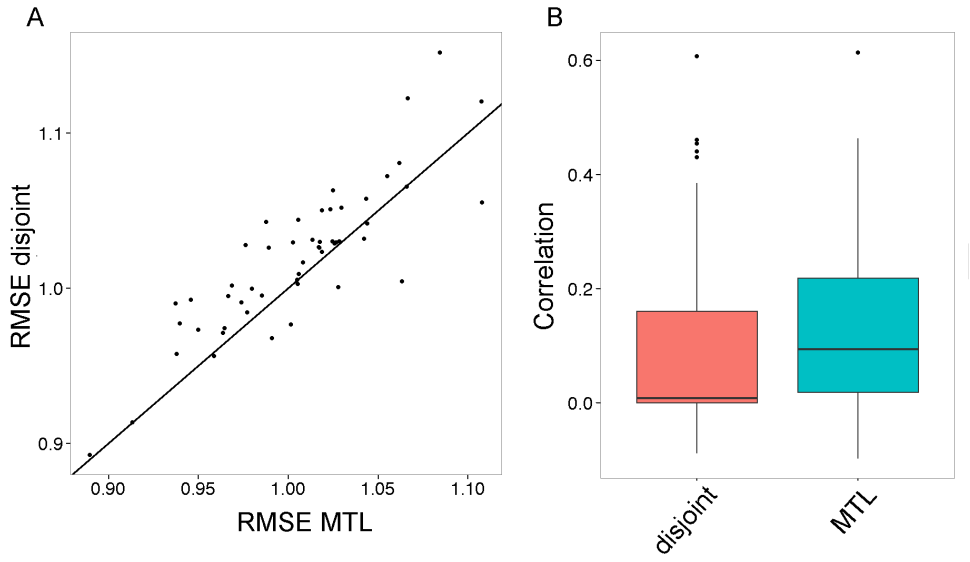

Figure 6. Comparison of the disjoint and MTL approach for prediction of miRNA-transcript target interactions in liver cancer RNA-seq data. A) Comparison of prediction error on test samples (RMSE) for disjoint (y-axis) and MTL (x-axis). B) Boxplots of correlation values comparing predicted and observed expression values for all 114 transcripts. With both metrics MTL provides better estimates.

by summing the absolute values of all miRNA transcript coefficients of that gene. These summed miRNA coefficients are used for ranking the MTL and disjoint models. In Fig. 7B it is shown that both transcript based methods outperform the gene based prediction in the top ranks and also find overall more true positive interactions.

4 DISCUSSION AND CONCLUSION

We have introduced a new paradigm for prioritizing miRNA target associations that works on the level of individual transcripts instead of the (virtual) gene. We have introduced two sparse learning setups that allow to infer miRNA-transcript interactions from paired miRNA and mRNA expression data. The MTL approach in particular makes full use of available transcript expression data and annotation to estimate miRNA interaction coefficients by borrowing from all transcripts of the same gene that share this binding site. If no binding sites are shared between transcripts of a gene, then the MTL approach reduces to the multi response lasso and therefore is no limitation in practice.

We conclude that the reduced variance on coefficients with the MTL approach leads to lower errors in simulated and real data. However, we note that our simulations are oversimplified, disregarding many other relevant aspects of miRNA target prediction, for example RNA secondary structure, and assuming that expression levels follow a Gaussian distribution. Further, one could explore more complex simulation setups with a varying number of miRNAs for instance.

In real data, although most of the genes we tested only had two expressed transcripts, the MTL approach showed a clear advantage, as sharing between more transcripts reduces the variance of coefficient estimates due to the L_2 group norm. Further, we note that the low recall values obtained in Fig. 7, are likely due to using a general set of gold-standard miRNA-target interactions obtained from miRTarBase. Many of these interactions may not be relevant in the liver cancer samples analyzed.

It has been reported that transcript expression estimation from RNA-seq data can be noisy. Therefore we evaluated robustness of our method on different noise levels on synthetic data and showed that the MTL formulation increases robustness to noisy transcript expression estimation. Studying how biases in transcript expression estimation impact miRNA prioritization further would be helpful to understand the limits of the suggested approach when applied to commonly available protocols for RNA-seq data sets (Griebel et al., 2012). However, more recent technologies like PacBio or Oxford Nanopore often allow for complete sequencing of the transcript, therefore mitigating the above mentioned biases in transcript expression estimation.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2377v1 | CC BY 4.0 Open Access | rec: 19 Aug 2016, publ: 19 Aug 2016

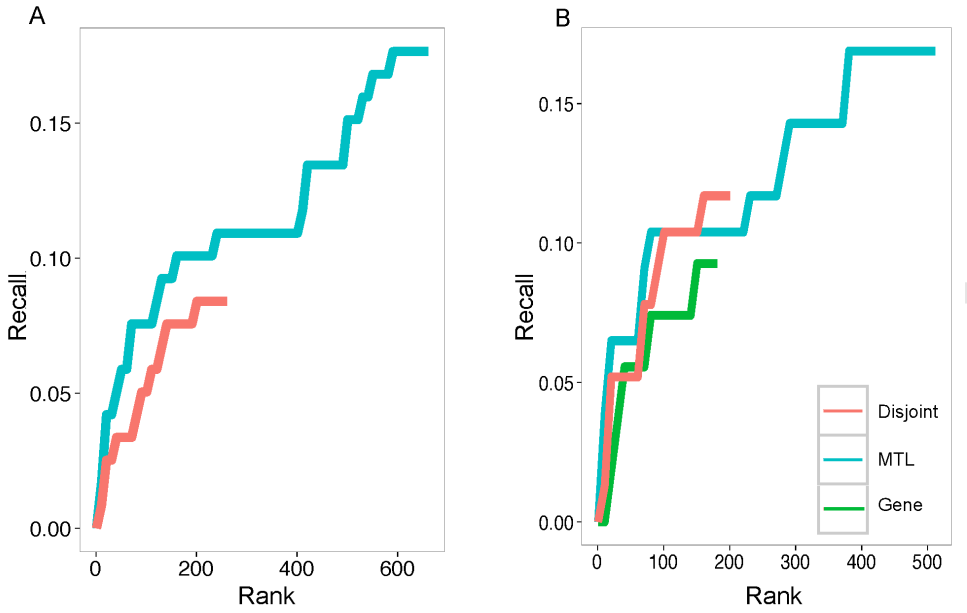

Figure 7. Analysis of experimentally validated miRNA binding sites from miRTarBase (Chou *et. al*, 2015). A) Recall (y-axis) of the MTL and disjoint method are shown as a function of ranked miRNA-transcript coefficients (x-axis) obtained on liver cancer RNA-seq data. B) Similar to (A) except that recall and ranking is done after projecting interactions to gene level, by summing all miRNA transcript coefficients. A model learnt on liver cancer gene expression data is compared to MTL and disjoint.

We observed that strong binding coefficients show reduced variance during learning, thus it could be interesting to use the adaptive multitask lasso to incorporate prior information of binding site strength for each coefficient, which may further improve estimation in practice.

The new formulations open the door for re-evaluation of miRNA-targeting in many cancer expression datasets, that previously only used aggregated gene expression measurements. Ultimately, predicting at the transcript level is not only likely to improve performance but also gives a direct link to the affected protein and thus to drug targets.

ACKNOWLEDGEMENTS

We thank Seyoung Kim for suggesting the multitasking framework for grouping binding sites in different transcripts. We thank all labs that contributed to the liver cancer data in TCGA.

REFERENCES

Agarwal *et. al*, V. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *eLife*, 4.

Chen, X., Lin, Q., Kim, S., Carbonell, J. G., Xing, E. P., et al. (2012). Smoothing proximal gradient method for general structured sparse regression. *The Annals of Applied Statistics*, 6(2):719–752.

Chou *et. al*, C.-H. (2015). miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic acids research*, page gkv1258.

Deng *et. al*, N. (2011). Isoform-level microRNA-155 target prediction using RNA-seq. *Nucleic Acids Research*, 39(9):e61–e61.

Engelmann, J. C. and Spang, R. (2012). A least angle regression model for the prediction of canonical and non-canonical mirna-mrna interactions. *PloS one*, 7(7):e40634.

Esquela-Kerscher, A. and Slack, F. J. (2006). Oncomirs - microRNAs with a role in cancer. *Nat Rev Cancer*, 6(4):259–269.

Friedman *et. al*, R. C. (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome research*, 19(1):92–105.

Grant, M. and Boyd, S. (2014). CVX: Matlab software for disciplined convex programming, version 2.1.

Griebel, T., Zacher, B., Ribeca, P., Raineri, E., Lacroix, V., Guigo, R., and Sammeth, M. (2012). Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Research*, 40(20):10073–10083.

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

Jacobsen *et. al*, A. (2013). Analysis of microRNA-target interactions across diverse cancer types. *Nat Struct Mol Biol*, 20(11):1325–1332.

John *et. al*, B. (2004). Human microRNA targets. *PLoS biology*, 2(11):e363.

Krek *et. al*, A. (2005). Combinatorial microRNA target predictions. *Nature genetics*, 37(5):495–500.

Li, B. and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics*, 12(1):323.

Lu *et. al*, Y. (2011). A Lasso regression model for the construction of microRNA-target regulatory networks. *Bioinformatics*, 27(17):2406–2413.

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

Patro, R., Mount, S. M., and Kingsford, C. (2014). Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat Biotech*, 32(5):462–464.

Richard, H., Schulz, M. H., Sultan, M., Nürnberger, A., Schrinner, S., Balzereit, D., Dagand, E., Rasche, A., Lehrach, H., Vingron, M., et al. (2010). Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments. *Nucleic Acids Research*, 38(10):e112–e112.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12(1):1–8.

Schulz *et. al*, M. H. (2013). Reconstructing dynamic microRNA-regulated interaction networks. *Proceedings of the National Academy of Sciences*, 110(39):15686–15691.

Teng, M., Love, M. I., Davis, C. A., Djebali, S., Dobin, A., Graveley, B. R., Li, S., Mason, C. E., Olson, S., Pervouchine, D., Sloan, C. A., Wei, X., Zhan, L., and Irizarry, R. A. (2016). A benchmark for RNA-seq quantification pipelines. *Genome Biology*, 17(1):1–12.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J. M., Network, T. C. G. A., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2377v1 | CC BY 4.0 Open Access | rec: 19 Aug 2016, publ: 19 Aug 2016