Context-specific transcriptional regulatory network inference from global gene expression maps using double two-way *t*-tests

Jianlong Qi[1] Tom Michoel[1,2]

[1]Freiburg Institute for Advanced Studies
The University of Freiburg, Germany

[2]The Roslin Institute
The University of Edinburgh, Scotland, UK

MLSB, 2012

Outline

Outline

Regulation of biological processes in cells

- Transcriptional regulation
- Post-transcriptional regulation
- Post-translational regulation

Transcription factor (TF) in transcriptional regulation

- Transcription factors are proteins
- They regulate the expression of their target genes

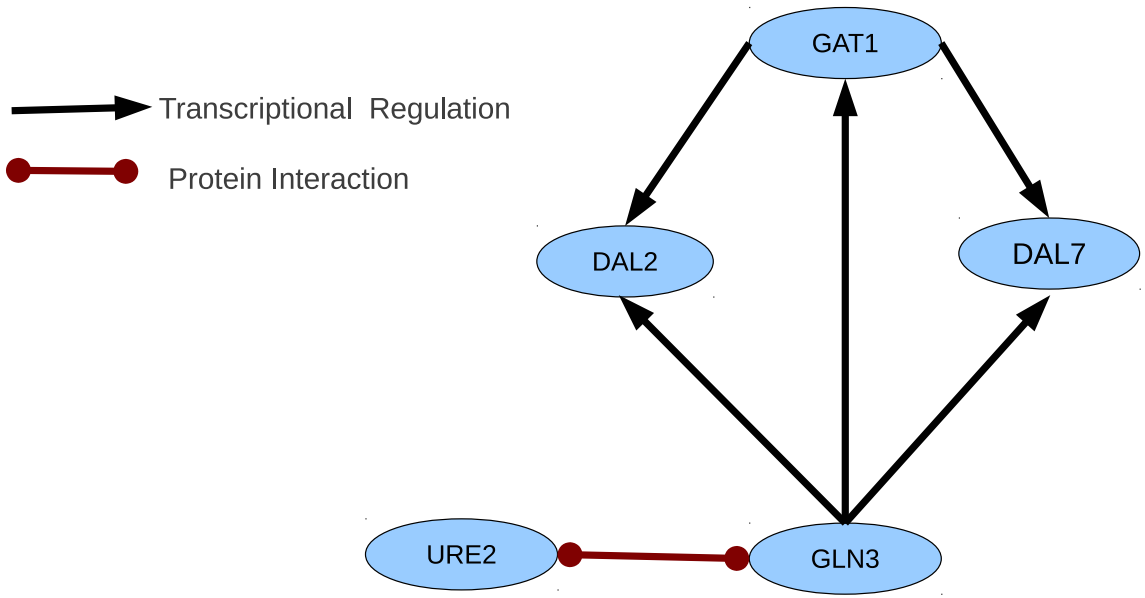


Figure: Regulation of nitrogen utilization in yeast.

Outline

- Gene expression data are often used to infer regulatory networks.
- Molecular interactions between transcription factors and their targets might lead to corresponding correlations between their expression values.

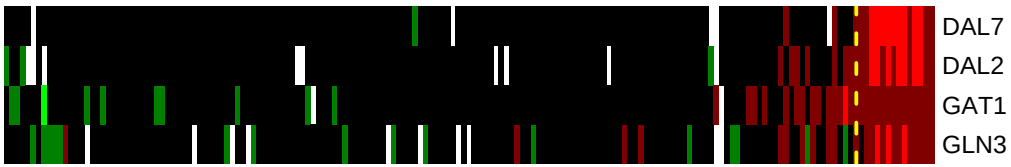


Figure: Heatmap showing the expression values of GLN3, GAT1, DAL2 and DAL7. Red - over-expressed, green - under-expressed and black - no change compared to wild-type expression levels.

Typical algorithms

- Bayesian network [Friedman *et al*., 2000]
- Mutual information [Faith *et al*., 2007]
- Linear regression [Bonneau *et al*., 2006]
- Random forest [Huynh-Thu *et al*., 2010]

Input

- A matrix of gene expression values
- A list of candidate transcription factors

Output and evaluation

- An ordered list of putative regulator-gene interactions
- Recall and Precision

$$\text{rec}(k) = \frac{\text{TP}(k)}{N_{\text{ref}}} \qquad \text{prec}(k) = \frac{\text{TP}(k)}{k},$$

where $\text{TP}(k)$ is the number of known interactions, among the first k predictions and N_{ref} is the total number of known interactions.

Benchmark dataset

- Yeast stress dataset [Segal *et al*., 2003] for 2355 genes under 173 conditions.
- *E. coli* dataset [Faith *et al*. 2007] for 4,345 genes under 189 conditions.

Performance

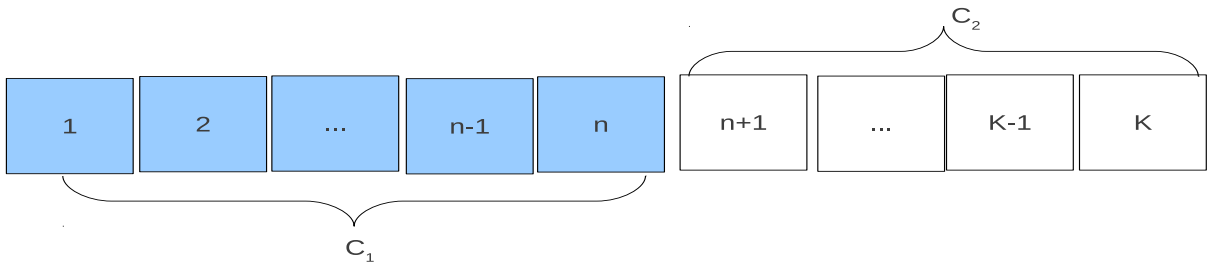
- Better performance at prokaryote than eukaryote
- Degraded performance at genes regulated by multiple transcription factors

Outline

- The differential expression of a gene g in a partition (C_1, C_2) of the set of samples in two distinct sets can be determined by the ordinary t-statistic,

$$t = \frac{|\mu_1 - \mu_2|}{\sqrt{\frac{(n_1-1)\sigma_1^2+(n_2-1)\sigma_2^2}{n_1+n_2-2}}\sqrt{\frac{n_1+n_2}{n_1 n_2}}}$$

- Given K samples in the dataset, the critical contrast of g can be determined by taking the maximum over all $K-1$ partitions.



- For each gene, sort expression levels and find critical contrast (2-way *t*-test)

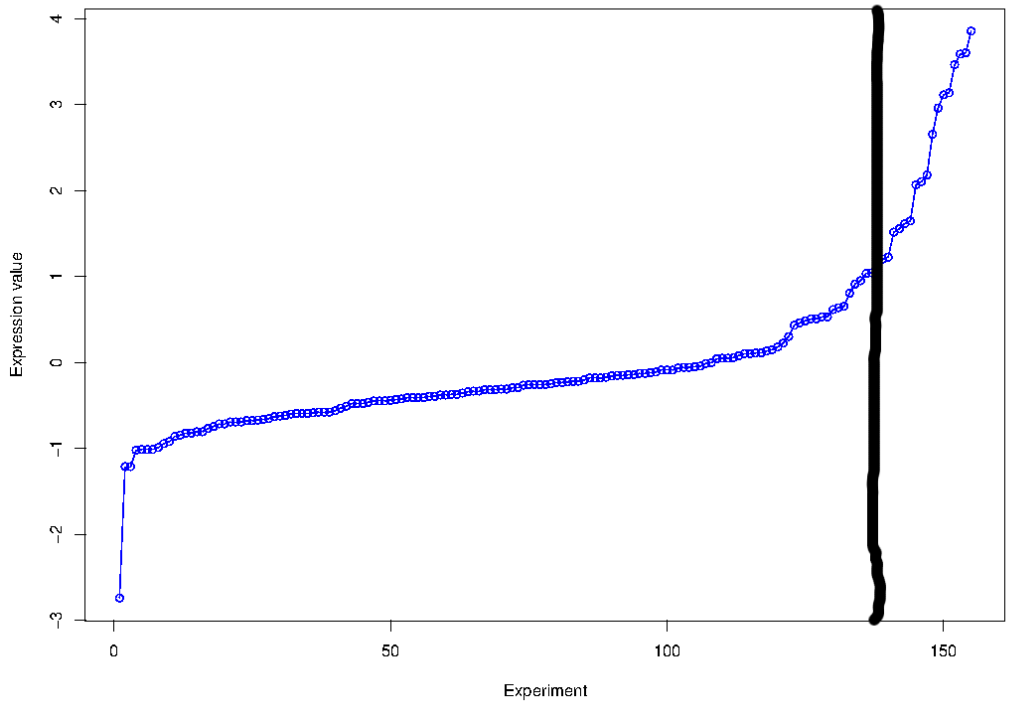


Figure: DAL2

Outline

- The interaction score $t_{f,g}$ between a TF f and g is determined by the t-statistic of f in the critical contrast of g.
- The higher $t_{f,g}$, the more confident we are about the predicted regulatory interaction $f \rightarrow g$.
- Background correction for $t_{f,g}$:

$$Z_{f,g} = \frac{t_{f,g} - \mu_g}{\sigma_g},$$

where μ_g and σ_g are the mean and standard deviation of $t_{f,g}$ over all TFs

- TFs and their targets are both differentially expressed in a gene-specific sample contrast.
- No assumption on any linear or non-linear relation between the expression profiles of TFs and their targets.
- Interactions found by the *t*-test procedure tend to only co-express locally.

Outline

Network inference methods

- TwixTrix: Two-way *t*-test
- CLR: Mutual information
- Inferelator: Linear regression
- LeMoNe: Two-way clustering
- GENIE3: Random forest
- Pearson correlation and Spearman correlation

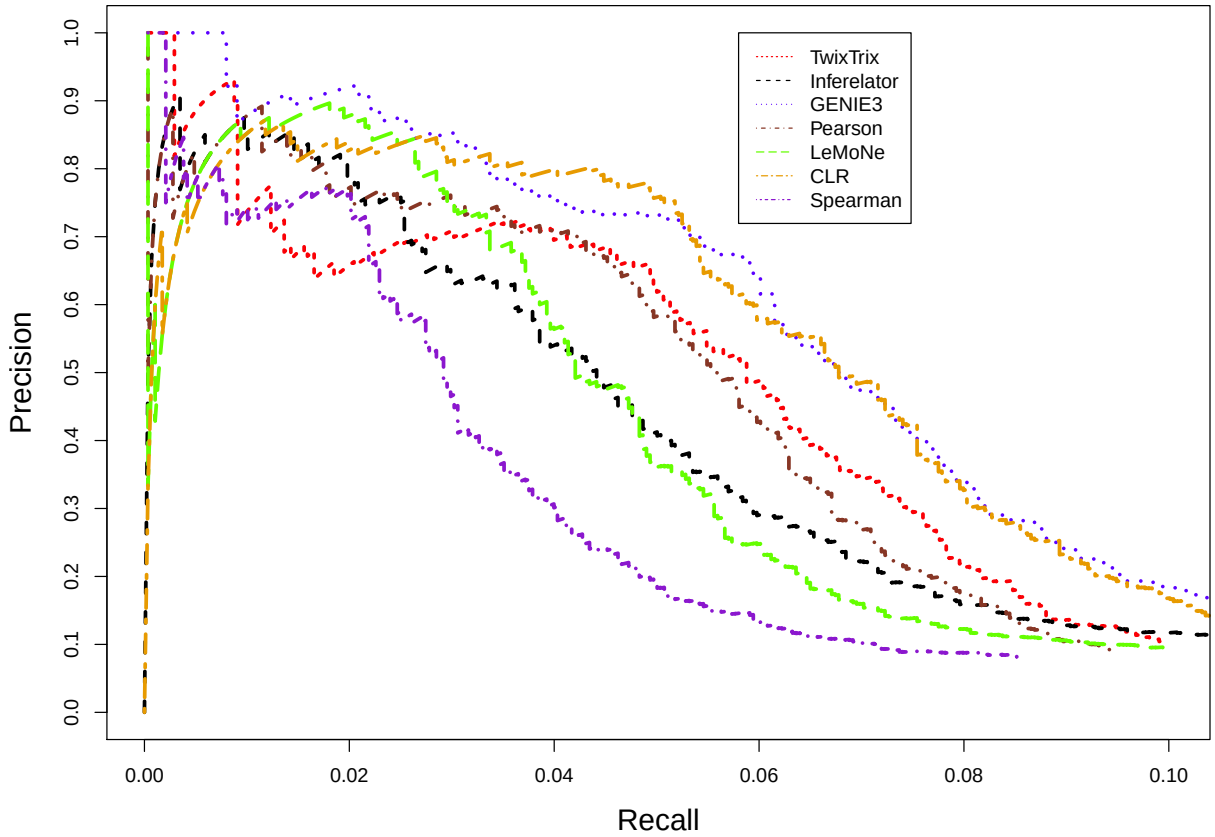


Figure: Recall-precision curves for seven transcriptional regulatory network inference algorithms in *E.coli*.

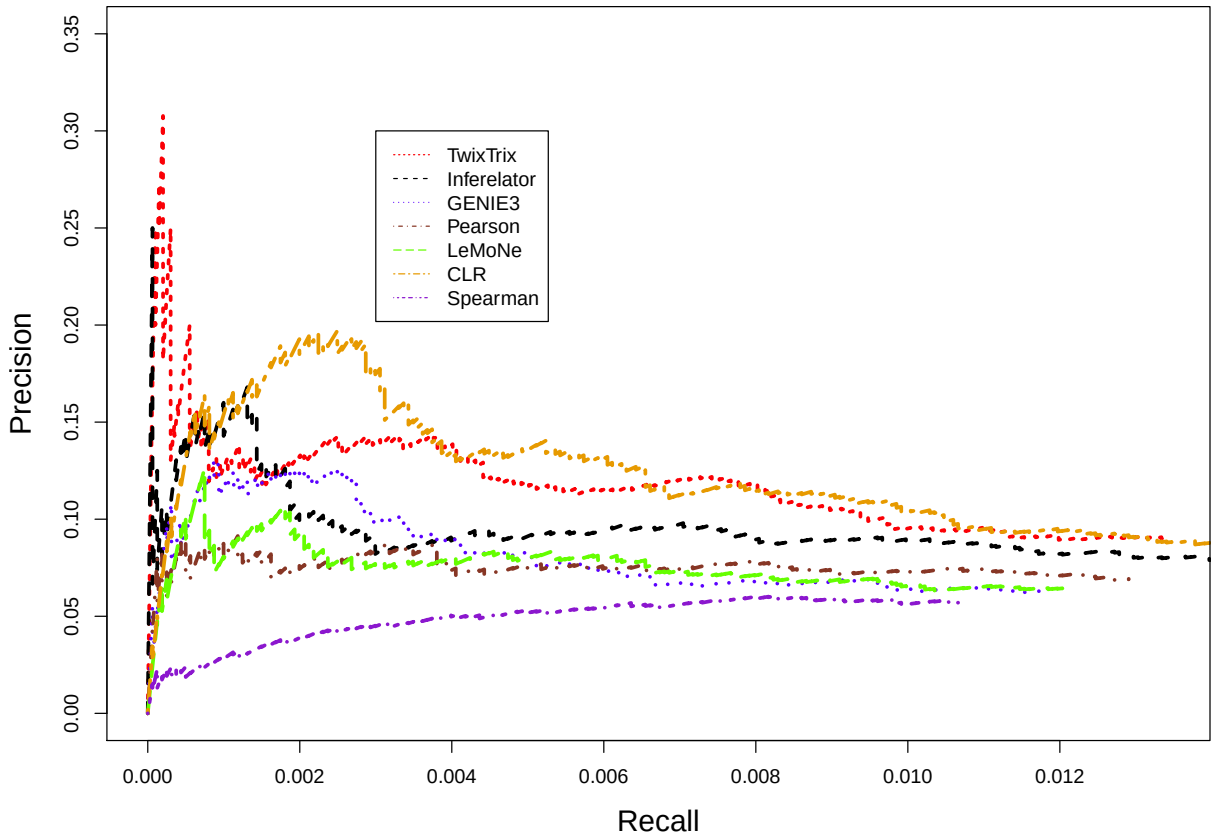


Figure: Recall-precision curves for seven transcriptional regulatory network inference algorithms in yeast.

	E. coli	Yeast
TwixTrix	0.05182	0.00157
Inferelator	0.04624	0.00140
GENIE3	**0.06767**	0.00097
LeMoNe	0.04415	0.00091
CLR	0.06269	**0.00190**
Pearson	0.05003	0.00097
Spearman	0.03157	0.00052

Table: Area under the recall-precision curve for each method in *E. coli* and yeast. The bold numbers indicate the highest value in each organism.

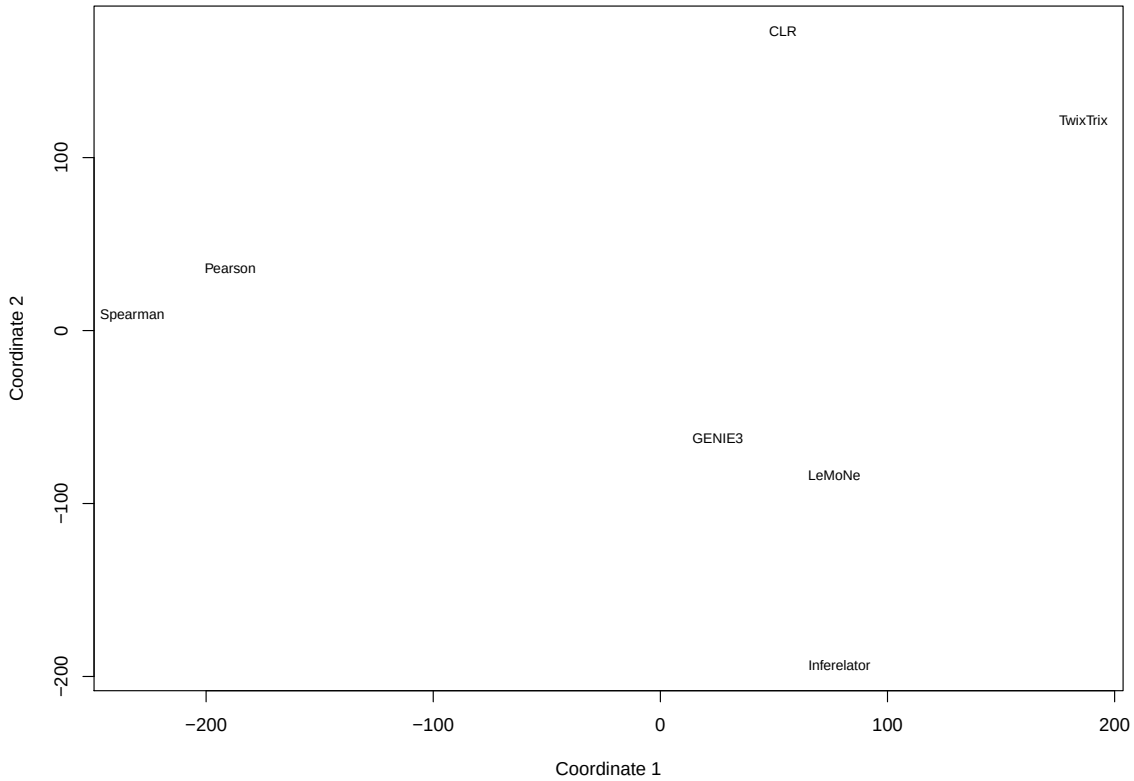


Figure: Multi-dimensional scaling plot, using the number of non-overlapping interactions among the top 500 predicted interactions as a distance measure between network inference methods.

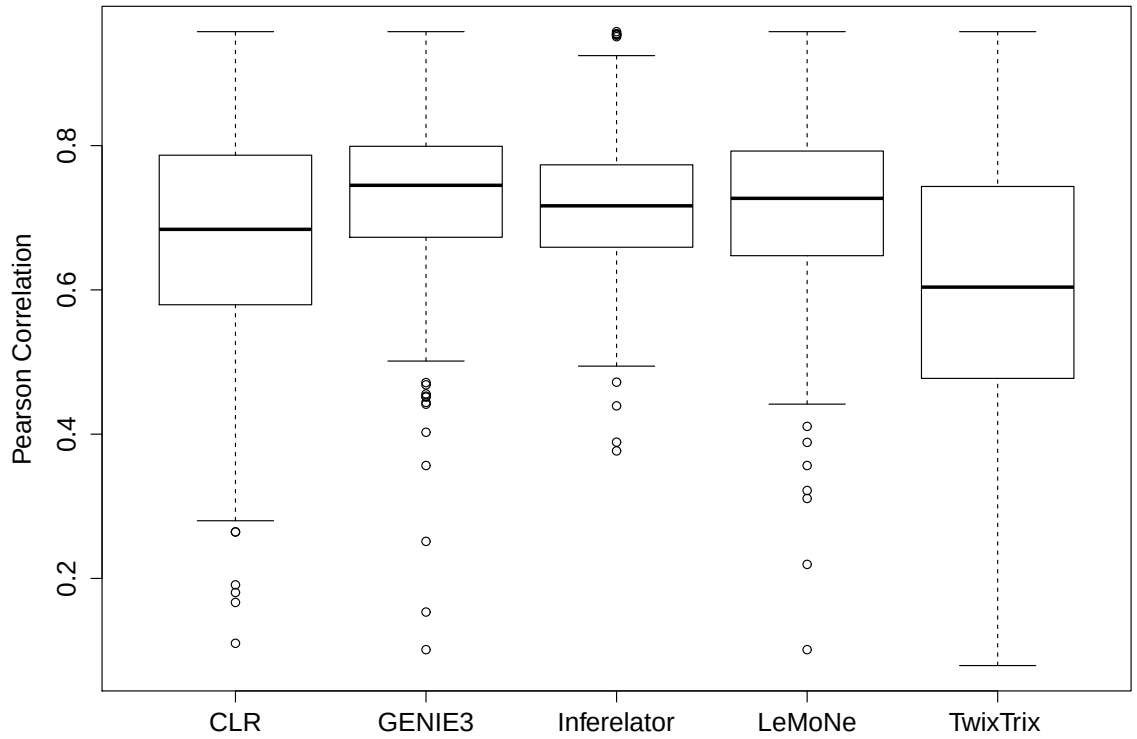


Figure: Distribution of Pearson correlations for the top 500 predicted TF-target interactions in yeast from five network inference methods.

Outline

Human dataset

- 12,568 genes, 1,033 samples from 64 tissue types
- Good for testing context-specific interactions and global interactions

Two-way *t*-test

- TBX5 $\rightarrow$ BMP10: TBX5 is a TF with a role in heart development.
- GCM1 $\rightarrow$ PAPP: GCM1 is the placental TF.

CLR

- BBX $\rightarrow$ TPR: BBX is a TF for cell cycle progression from G1 to S phase.
- ZNF24 $\rightarrow$ BPTF: ZNF24 is a TF involved in promoting the cell cycle.

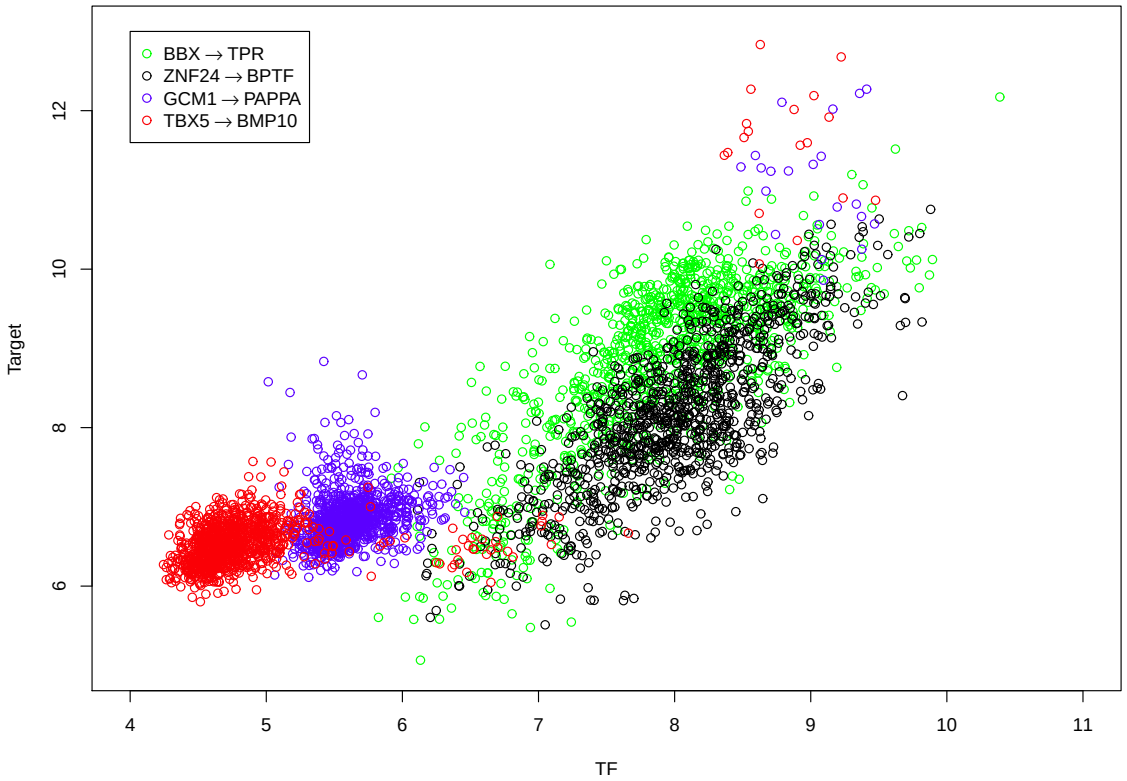


Figure: Scatter plot of expression levels for representative high-scoring TwixTrix (blue and red) and high-scoring CLR (green and black) predicted interactions.

Outline

Strength

- A simple method with performance on par with state-of-the-art methods.
- Sensitive to context-specific regulatory interactions.
- Very fast (e.g., less than a minute in the human dataset)

Weakness

- Assign less weight to globally co-expressed TF-target pairs.

Summary

- The two-way *t*-test method provides a useful addition to existing network inference methods.
- Integrating results from inference methods with different nature.

- In large expression compendia for multi-cellular organisms, it seems expression is highly tissue-specific and consistent with an off/on-model (which is what the proposed method detects). If we move to RNA-seq, will such an off/on-model still work or will we do better with a model where we assume a gene is off in most tissues, but with a more complicated relation in the other tissues?

- In yeast, our algorithm works as well as others which try to model the TF-gene interaction in a biophysically more accurate way. This probably means that microarray data is too noisy to reflect true biophysical expression relations. If we move to RNA-seq, will we get sufficient increase in resolution to model TF-gene interactions with biophysical models?

Thank you!