# Time-sensitive inference of gene regulatory networks

Pegah Tavakkolkhah, Ralf Zimmer, Robert Küffner

*Department of Informatics, Ludwig-Maximilians University, Amalienstr. 17, 80333 Munich, Germany*

*{pegah, zimmer, kueffner}@ifi.lmu.de*

**Introduction.** Many algorithms were devised to deduce gene regulatory networks (GRN) from mRNA expression data. Candidate transcription factor:target gene (TF:TG) relationships are assumed more likely if the expression of the TG depends on the expression of the TF. This dependency can for instance be evaluated by Pearsons linear correlation coefficient $\rho^2$ [2] or by $\eta^2$ [3], a non-parametric, non-linear correlation coefficient computed from an analysis of variance (ANOVA). In particular, $\eta^2$ performed significantly better than previously published methods in the recent DREAM5 competition [1].

Inference algorithms usually neglect to analyze whether expression changes in TFs precede expression changes in TGs. We present a simple but effective approach to extend standard algorithms (exemplified by $\rho^2$ and $\eta^2$) by an analysis of time shifted expression patterns from time series data and report the achieved performance improvements.

**Methods.** Usually, interactions are ranked by a correlation $c_1$ (here based on $\rho^2$ or $\eta^2$) that relates TF levels to the corresponding TG levels measured on the same chip. We propose to compute two more correlations, $c_{for}$ and $c_{rev}$. Therefore, a TF level at an earlier time point $t_1$ is also related to a TG level at a later time point $t_2$ yielding $c_{for}$ (and vice versa, yielding $c_{rev}$). Both measurements $t_1$ and $t_2$ are derived from different time points in the same time series and satisfy the following constraint: $T_1<t_1$ & $t_2<T_2$ & $t_2-t_1>T_3$ & $T_4>t_2-t_1$, where $T_1 \ldots T_4$ are time thresholds, to be determined empirically or from biological knowledge. A time band (Figure 1, left) is thus defined where meaningful expression changes are expected to occur. Using $c_2= c_{for}^2- c_{rev}^2$ we define a combined score $c=w*\text{rank}(c_1)+(1-w)*\text{rank}(c_2)$ with $w=0.9$. Candidate interactions are sorted for relevance according to $c$. We evaluate $c$ by a *directionality* test distinguishing known interactions TF:TG (true, 51%) from their reverse TG:TF (false, 49%) as well as by an *inference* test distinguishing known interactions (true, ≈1%) from all other possible interactions (false, ≈99%). In the directionality test, #true is larger than #false due to bidirectional interactions. Expression datasets, known TF-TG relationships and evaluation protocols, e.g. area under precision-recall curve (AUPR), were used as in the DREAM5 assessment [4].

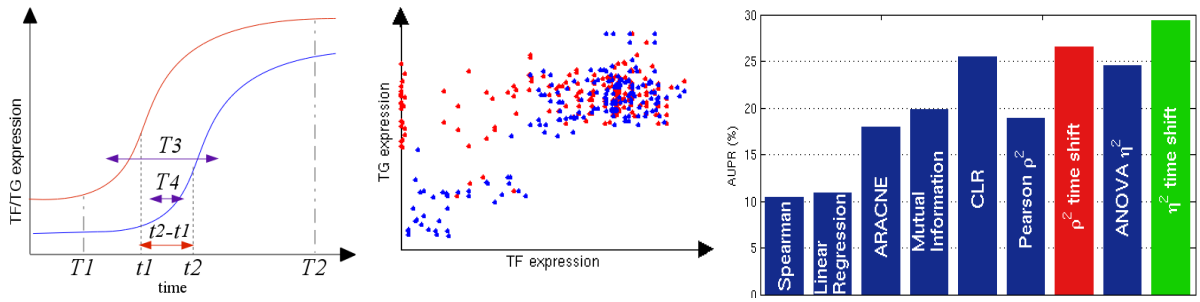


Figure 1. Expression of target genes (blue, left panel) lags behind the expression of their regulators (red). Thus, if earlier time points of the TF are correlated to later time points of the TG (blue, middle panel) the correlation between TF and TG will be higher than in the inverse case (red). Existing methods can be improved by analyzing time delays (red and green, right panel).

**Results and Discussion.** TF mRNA needs to be exported from the nucleus, translated into the TF protein, which has to be imported back into the nucleus before expression changes of a TF can become effective. This leads to considerable time delays between TF and TG expression changes. The majority of current inference methods neglect a dedicated analysis of time series but solely focus on correlation for the inference of causal dependencies. Testing for such time delays (i.e. TF expression changes preceding TG expression changes) should therefore improve the accuracy of the network inference algorithms. We determined that such temporal information can be extracted from expression data by a directionality test that resulted in AUPR of 81.9% on DREAM5 artificial data. We showed that different methods can be improved by integrating temporal dependencies, e.g. Pearsons correlation $\rho^2$ from 18.9 to 26.5% AUPR and ANOVA's $\eta^2$ from 24.5 to 29.3% AUPR. Other commonly used approaches, e.g. based on mutual information, should also benefit from time series analysis. The presented method is very simple and we expect additional gains in performance by using more involved analyses of time series data.

**References**

[1] R. Küffner, T.Petri, P. Tavakkolkhah, and L. Windhager, *Inferring Gene Regulatory Networks by ANOVA,* submitted to Bioinformatics, 2011.

[2] A. Butte, and I. Kohane, *Unsupervised knowledge discovery in medical databases using relevance networks,.* In Proceedings of the AMIA Symposium, American Medical Informatics Association, 1999, pp. 711-715.

[3] J. Cohen, *Educ Psychol Meas*, **33**(1), 1973, pp. 107-112.

*[4] DREAM5 setting and data: http://wiki.c2b2.columbia.edu/dream/index.php/D5c*