Unravelling the murine osteoblast differentiation pathway by network structure analysis using time-series microarray data

E.P. van Someren
Dept. of Mediametics
Delft Univ. of Technology
Delft, The Netherlands
E.P.vanSomeren@ewi.tudelft.nl

B.L.T. Vaes
Dept. of Applied Biology
Univ. of Nijmegen
Nijmegen, The Netherlands

W.T. Steegenga
Division of Human Nutrition
Wageningen University
Wageningen, The Netherlands

A.M. Sijbers
Target Discovery Unit
N.V. Organon
Oss, The Netherlands,

K.J. Dechering
Target Discovery Unit
N.V. Organon
Oss, The Netherlands,

M.J.T. Reinders
Dept. of Mediametics
Delft Univ. of Technology
Delft, The Netherlands

Abstract

We propose a reverse engineering scheme to discover genetic regulation from genome-wide transcription data that monitors the dynamic transcriptional response after a change in cellular environment. The interaction network is estimated by solving a linear model using simultaneous shrinking of the least absolute weights and the prediction error.

The proposed scheme has been applied to the murine C2C12 cell-line stimulated to undergo osteoblast differentiation. Results show that our method discovers genetic interactions that display significant enrichment of co-citation in literature. More detailed study showed that the inferred network exhibits properties and hypotheses that are consistent with current biological knowledge.

1. Introduction

In order to understand any developmental process, it is imperative to unravel the underlying genetic interaction network. We devised a method for Least Absolute Regression Network Analysis (LARNA) that can unravel genetic network structure from microarray data sampled in time. The interaction network is estimated by solving a linear model using simultaneous shrinking of the least absolute weights and the prediction error. This approach effectively solves the problem of having a limited amount of arrays by focussing on finding the structure of the network.

Current successful methodologies to infer genetic interactions from microarray data, however, have primarily been restricted to the use of perturbation (e.g. knockout) microarray data [2, 4].

2. Network Inference

LARNA is based on a linear model [2] that assumes that the gene expression level of each gene is the result of a weighted sum of all other gene expression levels at the previous time-point $\hat{\mathbf{y}}^t = \mathbf{W} \cdot \mathbf{x}^t + \epsilon$[1]. The interaction parameter, $w_{ij} \in \mathbf{W}$, represents the existence ($w_{ij} \neq 0$) or absence ($w_{ij} = 0$) of a controlling action of regulating gene j on target gene i, whether it is activating ($w_{ij} > 0$) or inhibiting ($w_{ij} < 0$), as well as the strength ($|w_{ij}|$) of the relation.

LARNA distinguishes itself from other regression network models as it provides a unique trade-off between data-fit versus robustness and limited connectivity. This was obtained by augmenting the standard total squared error with a penalty term that sums the absolute values of the weights [5]:

$$\hat{\mathbf{W}} = \arg\min_{\mathbf{W}} \sum_{t=1}^{T} ||\hat{\mathbf{y}}^t - \mathbf{y}^t||^2 + \lambda \sum_{i=1}^{N} \sum_{j=1}^{N} |w_{ij}| \quad (1)$$

3. Results

The proposed scheme has been applied to the murine C2C12 cell-line stimulated to undergo osteoblast differentiation (8 arrays). To test how well our method compares to other approaches, we have also applied three linear models

[1] $t \in 1, 2, \ldots, T$, T is no. array pairs and if $\mathbf{x}^t$ represents a measurement at time s, e.g. $\mathbf{x}^t = \mathbf{x}(s)$, than $\mathbf{y}^t = \mathbf{x}(s + \Delta s)$

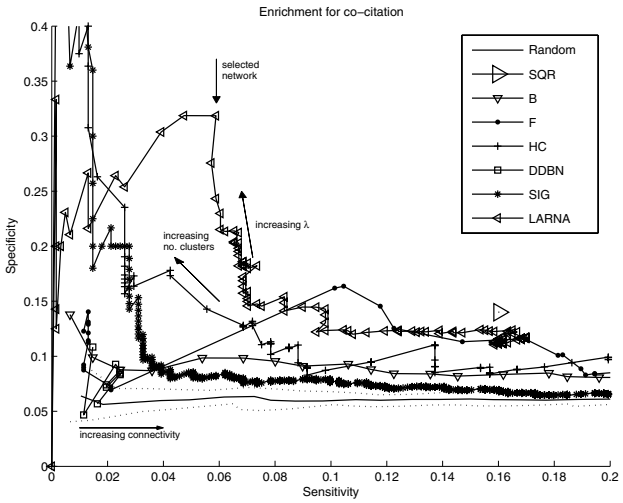

Figure 1. The performance of the network methods in terms of co-citation enrichment.

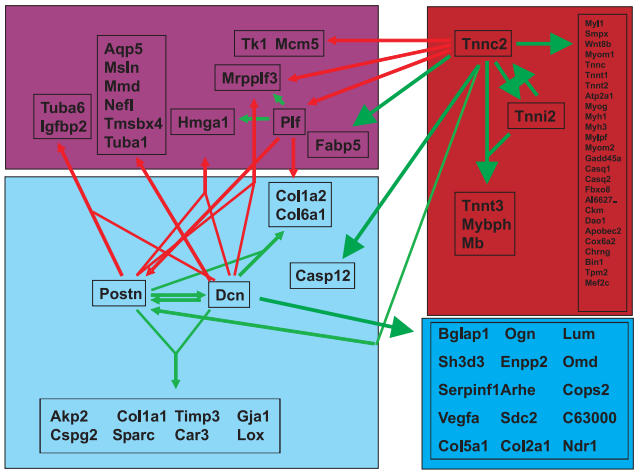

Figure 2. Diagram of the network inferred by *LARNA*

(i.e. *sparse QR (SQR)*, *backward search (B)* and *forward beamsearch (F)*), two non-linear approaches (i.e. *Discrete Dynamic Bayesian Network (DDBN)* and *Sigmoidal Gradient Ascent (SGA)*) to the same data. Furthermore we compare these networks with randomly connected networks and to a co-expression network obtained from hierarchical clustering.

Jenssen [3] has shown that a co-citation network reflects biologically meaningful relationships. One way to validate the network results is to compare them against a co-citation network. For each predicted network, its performance is reflected by its specificity (how often a proposed interaction concerns genes that are related $SP = \frac{TP}{TP+FP}$) and its sensitivity (how much of the 'ground truth' is discovered $SE = \frac{TP}{TP+FN}$).

Figure 1 shows the performance of all methods The results show that standard linear approaches (*B*, *F* and *SQR*) only perform slightly better than. The non-linear *DDBNA* does not outperform random network structures at all, whereas the non-linear *SIG* only achieves strong enrichment at low sensitivity, but does not outperform the best clustering result. The best overall performance of all network inference methods was obtained by *LARNA*.

A single network from LARNA was selected to be analyzed for biological consistency (See Figure 2). The global inter-relationships between genes and modules revealed by the network fit extremely well with current knowledge of differentiation of mesenchymal cells. For example, the differentiation of mesenchymal cells occurs with a concurrent decline in proliferative capacity [1]. Accordingly, the network shows negative feedback from the muscle and osteoblast differentiation modules to the proliferation module.

4 Discussion

In conclusion, we presented LARNA, a regression algorithm to infer sparse genetic networks from microarray data sampled over time. LARNA was shown to outperform other network inference methods and clustering algorithms. Importantly, two global hypothesis raised by LARNA conform to current biological knowledge, i.e. osteoblast maturation is induced by extra-cellular matrix formation and proliferation and differentiation are two mutually exclusive modes of operation. Although the interactions revealed by LARNA on the current limited set of microarrays are of a global character we are confident that future studies with more arrays sampled at smaller intervals are likely to provide a network map with higher resolution.

References

[1] J. Aubin. Advances in the osteoblast lineage. *Biochemistry and Cell Biology*, 76(6):899–910, 1998.

[2] T. gardner, D. di Bernardo, D. Lorenz, and J. Collins. Reverse engineering genetic networks and identifying compound mode of action via expression profiling. *Science*, 2003.

[3] T. Jenssen, A. Laegreid, J. Komorowski, and E. Hovig. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28:21–28, may 2001.

[4] E. Segal, M. Shapira, A. Regev, D. P. D, D. Botstein, D. Koller, and N. Friedman. Module networks: Identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, 34(2):166–176, June 2003.

[5] E. van Someren, L. Wessels, E. Backer, and M. Reinders. Multi-criterion optimization for genetic network modeling. *Signal Processing*, 83:763–775, 2003.