TGF-β induced epithelial-mesenchymal transition modeling

P Xenitidis1, I Seimenis1, S Kakolyris2, and A Adamopoulos1
1 Medical Physics Lab, School of Medicine, Democritus University of Thrace, Alexandroupolis, Greece
2 Department of Medical Oncology, School of Medicine, Democritus University of Thrace, Alexandroupolis, Greece
E-mail: pxenitid@med.duth.gr

Abstract. Epithelial cells may undergo a process called epithelial to mesenchymal transition (EMT). During EMT, cells lose their epithelial characteristics and acquire a migratory ability. Transforming growth factor-beta (TGF-β) signaling is considered to play an important role in EMT by regulating a set of genes through a gene regulatory network (GRN). This work aims at TGF-β induced EMT GRN modeling using publicly available experimental data (gene expression microarray data). The time-series network identification (TSNI) algorithm was used for inferring the EMT GRN. Receiver operating characteristic (ROC) and precision-recall (P-R) curves were constructed and the areas under them were used for evaluating the algorithm performance regarding network inference.

1. Introduction
The epithelial-mesenchymal transition (EMT) is a process that leads epithelial cells to lose their cell to cell adhesion and, thus, become migratory. Epithelial cadherin (E-cadherin) is a protein that is encoded by the CDH1 gene. Downregulation of E-cadherin decreases cellular adhesion in a tissue and is considered to play an important role in the EMT. There are several pathways that are involved in EMT such as TGF-β, Wnt, NF-kB and Notch [1][2].

In a cell, there are thousands of genes that are up or down regulated in order to produce or stop producing the appropriate proteins that are needed. Some proteins, called transcription factors, are only used to activate other genes. These transcription factors regulate gene expression and the whole system can be viewed as a network called gene regulatory network (GRN). The process of using experimental data to identify how genes interact within a regulatory network is called network identification, network inference or reverse engineering. High-throughput technologies such as DNA microarrays are widely used to measure simultaneously the expression levels of large numbers of genes. Thus, there have been many efforts for GRN inference from microarray data [3]. This work aims at identifying the GRN associated with one of the main pathways involved in EMT by exploiting publicly available microarray data.

2. Modeling Approaches
A GRN is usually modeled as a network that is composed of nodes that represent genes and edges that represent the interactions between the genes involved. The GRN inference is based on two basic elements: network topology and network parameters. The network topology refers to the process of finding the edges. The network parameters refer to the strength and the type of these interactions. The basic steps for GRN inference are shown in Figure 1 and include an initial stimulation (perturbation)
and a set of measurements of the gene expression levels. Many mathematical models have been used for the GRN inference. Among them Boolean networks, differential equations, Bayesian networks, Petri nets, cellular automata, neural networks and Gaussian models [3]. Ordinary differential equations and specifically the TSNI algorithm proposed by Bansal [4] are employed in this study.

![Figure 1. The basic steps for inferring a GRN.](image)

### 3. Experimental Dataset and Pre-processing

The experimental dataset (GSE26858) that was used for the GRN inference was obtained from Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/). It refers to the levels of expression at different time points due to TGF-β treatment of lung epithelial A549 cancer cells and contains four measurements at 0, 2, 12, and 24 h, respectively. At time t=0, the cells are treated with 2 ng/mL of TGF-β [5]. The data were interpolated using piecewise cubic interpolation. Principal component analysis (PCA) can be used in combination with TSNI algorithm for dimensionality reduction [4]. Therefore, a pilot study was performed in relation to the use of PCA on the TGF-β treatment dataset.

Firstly, PCA was applied on the original dataset of expression profiles. The relevant PCA scatterplot is shown in Figure 2(a). Then, genes demonstrating expression profiles with a variance lower than the 10th percentile were filtered out. Genes having very low absolute values (specifically lower than 10% of the whole data set) were also filtered out. The PCA scatterplot of the filtered expression profiles is shown in Figure 2(b).

![Figure 2. Principal component analysis scatterplots of the original (a) and the filtered (b) dataset. Marked dots in (a) correspond to genes known to be involved in EMT.](image)

It is evident that filtering eliminates noisy gene expressions mainly located at the center of the scatterplot. Figure 2(b) also demonstrates four distinct gene clusters. If the expression profiles of the genes belonging to each cluster are plotted, then a similar pattern in each cluster is revealed. In Figure 2(a), the z-scores that belong to a group of genes that are widely accepted to be involved in the TGF-β induced EMT GRN have been superimposed on the original dataset. It is elicited that although PCA can identify strong expression patterns in a dataset, it lacks the power to reveal a group that is
specifically related to the EMT process. It was decided, therefore, not to use the principle component analysis as a method for the assessment of the genes taking part in the inference process.

4. TSNI algorithm evaluation

The performance evaluation of an inference algorithm is typically based on the receiver operating characteristic (ROC) and the precision-recall (PR) curves [6]. ROC and PR curves, along with the areas under them (AUROC and AUPR, respectively) can be found in the supplementary material (Figures S1- S6 and Table S1) for all cases studied.

Initially, the TSNI algorithm performance was tested on artificial data. An artificial GRN containing 10 genes was generated. One gene (Gene 1) was only triggered and the TSNI algorithm was implemented to infer the artificial network. A good agreement was found between the generated and inferred connectivity matrices as far as Gene 1 coefficients are concerned. A poor match was found for the rest of the connectivity coefficients. Although the effect of the triggered gene on the rest of the genes in the GRN is well identified, the genes that were not triggered in the training dataset are not adequately identified. The same conclusion can be reached by simulating all gene expressions in the time domain.

To further explore the reasons behind the above mentioned poor match between generated and inferred data sets, a multi-gene triggering scenario was simulated. In this approach, all genes in the GRN were triggered at different time points (i.e., in a successive time order). Specifically, each gene was triggered with a pulse function type stimulation which allowed the triggered gene to evolve dynamically and provide it with adequate time to reach equilibrium. Afterwards, the next gene was triggered and so on. The time series data matching turned out excellent. A good matching was also found between the generated and inferred connectivity matrices.

This work, therefore, provides evidence that a process of GRN inference which is based on data produced by single-gene triggering may not be able to capture the dynamics of the whole network. Using datasets produced by multi-gene stimulation, at separate time slots, greatly improves the inference performance of the implemented TSNI algorithm.

5. TGF-β induced EMT modeling

Since PCA did not reveal any group of genes that could be characterized EMT relevant, gene selection was based on previously published works [7][8]. For the model studied, seven genes were selected. The CDH1 gene has a dominant role in the specific model since it encodes the protein E-cadherin. Down-regulation of E-cadherin is considered a hallmark of EMT. The next gene selected was the Hmga2 gene which is induced by TGF-β through the Smad proteins. Hmga2 induces expression of Snail. Snail protein is an established regulator of EMT and is encoded by the Snai1 gene which was also taken into account. Snai2, Twist1, Zeb1 and Zeb2 genes were also included in the model.

The TSNI algorithm was implemented for the inference of TGF-β induced EMT GRN and the relevant connectivity matrix A was estimated. To evaluate the inferred system for TGF-β treatment, a simulation was performed in the time domain. Gene expression profiles are plotted in Figure 3, and it is clear that there is a very good agreement between simulated results and microarray measurements.

![Figure 3. TGF-β induced EMT GRN. Microarray (lines) vs simulated (cross signs) data.](image-url)
To evaluate the significance of the connectivity coefficients that were calculated by the TSNI algorithm, it was checked whether all coefficients have an impact on the time-series response of the inferred GRN. For this purpose, each coefficient $a_{ij}$ of the inferred connectivity matrix $A$ was varied and the mean square error (MSE) of each gene expression profile was calculated. All calculated MSEs were then added up to provide a total MSE index that describes how well the inferred system matches the original system in the time domain. After checking all the coefficient $a_{ij}$ of the inferred connectivity matrix $A$, it was found that all coefficients were sensitive and, therefore, it is not convenient to set any coefficient equal to zero without significant total MSE increase. A representative MSE curve can be found in the supplementary material (Figure S7). The inferred system was also checked for stability by producing the relevant zero pole plot. All poles were found to lie within the left-half of the s-plane indicating that the system is stable.

6. Conclusions
This study focused on the modelling of the gene regulatory network that regulates the TGF-β induced EMT process. PCA proved inefficient in facilitating gene selection with regard to EMT. The performance of the TSNI algorithm chosen for GRN inference was tested in two cases. Single-gene perturbation and multi-gene perturbation at distinct time-points. It was shown that single-gene perturbation datasets lead to models that achieve to capture the dynamics of the system regarding the triggered gene but not its dynamics associated with the non-triggered genes. Therefore, we proposed a model with gene-specific triggering time to provide the algorithm with all the information needed for successfully identifying the connectivity matrix coefficients. It was demonstrated that the proposed approach significantly improves the entire modeling performance. The TGF-β perturbation is an intermediate scenario between single and multi-gene triggering. The model is expected to predict the GRN connectivity matrix with adequate accuracy for TGF-β treatment. It is postulated, however, that the model performance will be significantly deteriorated if it is tested with other types of stimulations.

7. Limitations
The fact that available microarray datasets on TGF-β treatment do not allow the performance of a multi-gene perturbation type training for the inferred GRN constitutes the main limitation of the current study. In addition, current work assumes that only 7 genes are closely related to the EMT process, an assumption which is not based on solid scientific data.

References
[1] Jing Y, Han Z, Zhang S, Liu Y and Wei L 2011 Epithelial-Mesenchymal Transition in tumor microenvironment Cell & Bioscience 1 29
[2] De Herreros AG, Peiró S, Nassour M and Savagner P 2010 Snail family regulation and epithelial mesenchymal transitions in breast cancer progression J. Mammary Gland Biol. Neoplasia 15 135–147.
[3] Hecker M, Lambeck S, Toepfer S, van Someren E and Guthke R 2009 Gene regulatory network inference: data integration in dynamic models-a review Biosystems 96 86–103
[4] Bansal M, Della Gatta G and di Bernardo D 2006 Inference of gene regulatory networks and compound mode of action from time course gene expression profiles Bioinformatics 22 815
[5] Zhang Y et al. 2011 High throughput determination of TGFβ1/SMAD3 targets in A549 lung epithelial cells PLoS One 6
[6] Prill RJ et al. 2010 Towards a rigorous assessment of systems biology models: The DREAM3 challenges PLoS One 5
[7] Tan E, Thuault S, Caja L, Carletti T, Heldin C and Moustakas A 2012 Regulation of Transcription Factor Twist Expression by the DNA Architectural Protein High Mobility Group A2 during Epithelial-to-Mesenchymal Transition J. Biol. Chem. 287 7134–45
[8] Tran DD, Corsa CAS, Biswas H, Aft RL and Longmore GD 2011 Temporal and spatial cooperation of Snail1 and Twist1 during epithelial-mesenchymal transition predicts for human breast cancer recurrence Mol. Cancer Res. 9 1644-57