BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 10 2010, pages 1332–1339
doi:10.1093/bioinformatics/btq139

Systems biology

Advance Access publication March 31, 2010

An integer programming formulation to identify the sparse network architecture governing differentiation of embryonic stem cells

Ipsita Banerjee[1,2,*], Spandan Maiti[3], Natesh Parashurama[1] and Martin Yarmush[1]

[1]Center for Engineering in Medicine, Massachusetts General Hospital, Harvard Medical School, Shriners Hospital for Children, 51 Blossom Street, Boston, MA-02114, [2]Department of Chemical Engineering, University of Pittsburgh, 1242 Benedum Hall, 3700 O'Hara Street, Pittsburgh, PA 15261 and [3]Department of Mechanical Engineering – Engineering Mechanics, Michigan Technological University, 1400 Townsend Drive, Houghton, MI 49931, USA

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Primary purpose of modeling gene regulatory networks for developmental process is to reveal pathways governing the cellular differentiation to specific phenotypes. Knowledge of differentiation network will enable generation of desired cell fates by careful alteration of the governing network by adequate manipulation of cellular environment.

Results: We have developed a novel integer programming-based approach to reconstruct the underlying regulatory architecture of differentiating embryonic stem cells from discrete temporal gene expression data. The network reconstruction problem is formulated using inherent features of biological networks: (i) that of cascade architecture which enables treatment of the entire complex network as a set of interconnected modules and (ii) that of sparsity of interconnection between the transcription factors. The developed framework is applied to the system of embryonic stem cells differentiating towards pancreatic lineage. Experimentally determined expression profile dynamics of relevant transcription factors serve as the input to the network identification algorithm. The developed formulation accurately captures many of the known regulatory modes involved in pancreatic differentiation. The predictive capacity of the model is tested by simulating an *in silico* potential pathway of subsequent differentiation. The predicted pathway is experimentally verified by concurrent differentiation experiments. Experimental results agree well with model predictions, thereby illustrating the predictive accuracy of the proposed algorithm.

Contact: ipb1@pitt.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on January 21, 2010; revised on March 12, 2010; accepted on March 29, 2010

1 INTRODUCTION

Phenotype and functionality of a cell is largely governed by the underlying gene regulatory network (GRN). The GRN is of fundamental importance for the developmental process, where a pluripotent progenitor cell gives rise to multiple cell types in a multicellular organism. Acquisition of different cellular phenotypes stems from the differential expression patterns of specific transcription factors that activate a cascade of complex network architecture. While experimental data are fundamental in identifying the level of transcription and the nature of transcriptional control, understanding of the complex network architecture and prediction of the effects of individual interactions in such networks will require their quantitative description in terms of strength of interaction governing the network dynamics. In this article, we report a novel mathematical modeling effort that aims at identifying the transcription factor network governing differentiation of progenitor cells to a specific lineage. We exploit the notion of sparsity, common to many biological networks, to identify the most plausible GRN operative in this scenario. Our model predictions are supported by concurrent experiments in differentiating embryonic stem cells to a specific lineage, for this case the pancreas. As will be discussed subsequently, we believe that our approach will be beneficial for the development of targeted experimental protocols for the production of cells with a pre-specified fate.

Developments in large-scale genomic technologies have made data acquisition more tractable. This feat is increasing the emphasis on the development of meaningful quantitative models utilizing the wealth of experimental data (Bansal *et al.,* 2006). However, it is not always obvious how the data acquired through such techniques can be assembled into unambiguous predictive models. Tremendous effort has been focused to tackle the network identification problem (Davidson, *et al.*, 2002; Foteinou *et al.*, 2009; Yeung *et al.,* 2002) with significant success in analyzing bacterial and yeast (Segal, 2003) networks. However, the generalization of these methods to the inference of networks in higher eukaryotes is not always obvious. Furthermore, developmental GRNs are organized very differently from the GRNs responsible for cellular physiology, house-keeping, cell cycle, etc. (Bolouri and Davidson, 2003). In contrast to most other GRNs, developmental network occurs in a sequence of multiple cascades of transcriptional regulations (Davidson, 2001). Endomesoderm specification in pre-gastrular sea urchin embryo (Oliveri and Davidson, 2004) was among the first attempts in identifying developmental GRN, followed by mesoderm specification in the frog *Xenopus laevis* (Koide *et al.*, 2005), dorsoventral patterning and segmentation of the Drosophila embryo (Stathopoulos *et al.,* 2005), and B-cell differentiation in the

*To whom correspondence should be addressed.

mammalian immune systems (Singh *et al.*, 2005). Parallel efforts in identifying the regulatory networks governing *in vitro* differentiation of embryonic stem cells have been lacking till date, which has been attempted in this report.

The primary purpose of modeling GRNs for developmental process is to reveal pathways of differentiation that can be precisely manipulated to generate different cell types. Currently, it is an area of intense study due to the heightened interest in stem cell biology (Shaywitz and Melton, 2005). The main focus of this article is to capture the regulatory network using its key features: sparsity and cascade-like architecture; and quantify the influence of external environment on the governing network. This endeavor has significant relevance in the field of stem cell differentiation, where cell fate induction is controlled primarily by manipulation of the external environment via extracellular matrix, growth factors, chemical inducers/ repressors, etc. Such mathematical quantification will enable the *in silico* prediction of cell fate by environmental perturbations, resulting in the development of robust differentiation protocols.

The developmental regulatory network is typically organized in a distinctive cascade of control (Blais *et al.*, 2005) that enables the subdivision of the entire complex network into a number of smaller subsets or modules. Each module is under the control of a signature gene or 'hub' that plays a central role in directing the cellular response to a given stimulus. Typically, these hubs connect to very few other nodes, behaving like a small world network with very few steps involved in connecting two nodes. Another characteristic of developmental GRNs is the relative absence of inter-connectivity between hubs that presumably facilitates compartmentalization of various biological processes occurring in a cell.

These observations reinforce the segregation of pancreatic differentiation into specific interconnected modules. Subsequently, the regulatory architecture of a single stage, that of pancreatic endoderm differentiation to pancreatic progenitor, has been treated as a single module. The governing transcription factor of this stage has been identified to be Pdx-1 which was considered to be the 'hub' of the pancreatic differentiation module under consideration.

The mathematical formulation is developed based on the rationale that network sparsity characterizes the regulatory architecture governing development. Consequently, we develop our mathematical model based on the notion of *sparse coding* (Lee *et al.*, 2007). Network sparsity has been experimentally observed in visual system of primates (Vinje and Gallant, 2000), auditory system of rats (DeWeese *et al.*, 2003), and olfactory system of insects. Here, we envisage the notion of network sparsity as the governing criterion determining the regulatory network of differentiating embryonic stem cells and propose a formal mathematical structure to analyze such systems. We describe a novel bilevel optimization algorithm that will identify the underlying regulatory network, and validate it against an *in silico* network (Supplementary Material). The developed algorithm is then applied to a system of embryonic stem cells (ESCs) differentiating towards pancreatic lineage. We show that the identified network largely conforms to a number of observations reported in the literature regarding pancreatic development. Finally, we demonstrate the predictive capability of the mathematical model by simulating a likely mechanism to induce subsequent differentiation, and validating the model prediction with concurrent experiment. The pathway of inducing endocrine differentiation, as predicted by our model, has not yet been reported in literature. However, concurrent experiments in our laboratory successfully validated the salient features of our model prediction. Although developed and validated for the specific case of pancreatic differentiation of mouse ESC, the methodology is sufficiently general in scope to be applicable to any other GRN.

2 METHODS

2.1 Mathematical model

A variety of mathematical models can be used to describe genetic networks, including Boolean logic (Shmulevich *et al.*, 2002), Bayesian networks (Hartemink *et al.*, 2002), graph theory (Wagner, 2001) and ordinary differential equations (Tegner *et al.*, 2003). We have modeled the gene expression profile as a time continuous dynamical system by representing it as a system of coupled ordinary differential equations (Bansal *et al.*, 2006; Yeung, 2002):

$$\dot{\mathbf{X}} = f(\mathbf{X}) \tag{1}$$

where $\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$ represents the array of n mRNA concentrations of interest, and dot denotes the differentiation with respect to time. In the present study, $f(\mathbf{X})$ is modeled by a linear set of equations given by:

$$\dot{\mathbf{X}} = \mathbf{A}\mathbf{X} + \mathbf{B}\mathbf{U} \tag{2}$$

where $\mathbf{A}$ is the $n \times n$ connectivity matrix, $\mathbf{B}$ is the $n \times p$ matrix representing the effect of p perturbations on n genes, and $\mathbf{U}$ is a $p \times 1$ vector representing p perturbations. In the physical context of stem cell differentiation, u_i will represent the concentration of i-th growth factors/inducers/inhibitors used in the differentiation process, and b_{ij} will reflect the effect of these factors on j-th gene in the network.

While the model represented by Equation (2) is a set of ordinary differential equations continuous in time, the experimental measurements are typically performed at discrete time points t_k. Hence, this equation needs to be discretized in time in a manner such that experimental observations can be incorporated. We chose bi-linear transformation towards this end, because of its stability and low computational cost (Ljung, 1999). Using this transformation, Equation (2) is converted to the discrete form:

$$\begin{aligned}\mathbf{X}(t_{k+1}) &= \frac{2+\mathbf{A}\Delta t}{2-\mathbf{A}\Delta t}\mathbf{X}(t_k) + \frac{2\mathbf{B}\Delta t}{2-A\Delta t}\mathbf{U}(t_k) \\ &= \mathbf{A}_d\mathbf{X}(t_k) + \mathbf{B}_d\mathbf{U}(t_k)\end{aligned} \tag{3}$$

where subscript k denotes the value of a quantity at the current sampling point, and $k+1$ is the next sampling point. In the above formulation, the input parameters are the experimentally determined values of gene expression levels at different experimental time points $\mathbf{X}(t_k)$ as well as the external perturbation $\mathrm{U}(t_k)$. The unknown parameters to be determined are the connectivity matrix $\mathbf{A}$ and the effect of the external perturbation, $\mathbf{B}$. Given enough sampled observations of $\mathbf{X}(t_k)$, the estimation problem becomes well posed and a solution providing the best fit in the least square sense can be computed by minimizing the following objective function:

$$\min_{\mathbf{A},\mathbf{B}} \|\mathbf{X}(t_k+1) - \mathbf{A}_d\mathbf{X}(t_k) - \mathbf{B}_d\mathbf{U}(t_k)\|^2 \tag{4}$$

However, this formulation becomes ill-posed in the absence of sufficient experimental data points, which is commonplace in biological systems, specifically in stem cell differentiation. Typically, at this stage, a dimension reduction technique is employed such as principal component analysis (PCA), or singular value decomposition (SVD) (Bansal *et al.*, 2006) to map the original problem space to a lower dimensional subspace. Such techniques, although mathematically tractable, will invariably lead to a loss of information resulting from dimension reduction.

We have developed an alternate strategy to address such underdetermined systems based on inherent properties of biological networks: *network sparsity and modular cascade like architecture*. Yeung *et al.* (2002), were first to

introduce the concept of sparsity in the context of reverse engineering. They used SVD to generate a family of feasible solutions and further constrained the optimal choice to the sparse network. We also believe that constraint of network sparsity will be crucial in determining regulatory network of differentiating stem cells. Consequently, we formulate the reverse engineering problem as a bi-level mixed integer programming problem, where the upper level determines the network topology by promoting network sparsity using an integer programming formulation, while the lower level determines the optimal strength of the connections determined by that topology. Thus, at the end of the procedure we will arrive at a detailed description of the network, consisting of both the architecture and strength of interaction governing the dynamics of gene expression profile. The problem formulation is given by:

$$\begin{aligned} &\min \sum_{i,j=1}^{n} \lambda_{ij} \\ &\text{subject to:} \\ &\arg\min \|\mathbf{X}(t_k+1)-\Lambda\mathbf{A}_d\mathbf{X}(t_k)-\mathbf{B}_d\mathbf{U}(t_k)\|^2 \le tol, k=1,\ldots,m \\ &\mathbf{A},\mathbf{B} \end{aligned} \tag{5}$$

where Λ represents $n \times n$ binary variables corresponding to each component of the connectivity matrix $\mathbf{A}$. $l_{ij}=0$ implies no influence of gene j on gene i, while $l_{ij}=1$ implies that gene j influences gene i. The objective of the upper level formulation is to minimize the total number of network connections, given by $\sum\lambda$, which essentially reduces the density of the connectivity matrix. The upper level thus constitutes an L_0 norm minimization problem, where the number of elements of the matrix $\mathbf{A}$ is minimized to promote sparsity. The constraint evaluated at the lower level is also a minimization problem which ensures that the predicted profile matches the experimental data within user defined accuracy specified by a tolerance. The optimization variables of the lower level are continuous, and it determines the strength and the nature (inducer/inhibitor) of the connectivity matrix.

For a network of n genes, the connectivity matrix $\mathbf{A}$ consists of n^2 elements, and the vector $\mathbf{B}$ consists of n elements. Hence, the upper level contains n^2 binary variables and is solved using a combinatorial optimization technique. This route is chosen since L_0 minimization is an NP hard problem that is better suited to be solved using combinatorial approaches rather than approximation algorithms (Papadimitriou and Steiglitz, 1998). Although there is no efficient algorithm for hard combinatorial problems, evolutionary algorithms have been found to be efficient in finding an approximate solution (Yao, 1999). Note that due to the cascade architecture of pancreatic developmental GRN, the size of the network is conducive to the use of evolutionary algorithms like genetic algorithm (GA). Implementation of GA typically requires coding the continuous variables as bits of binary strings, and decoding the binary bits back to continuous form. However, the present integer programming formulation is particularly conducive to GA, since the binary optimization variables could be directly encrypted in the representative chromosome of the algorithm, hence avoiding additional steps of coding and decoding of continuous variables to binary format. Number of variables to be optimized in the lower level is determined at the upper level as: $\sum_{i,j=1}^{n} \lambda_{ij}$. Upper level integer programming essentially reduces the number of variables to be optimized in the lower level, and as a consequence the estimation problem remains well posed even with a small number of experimental observations. An additional constraint is imposed on the least square minimization program to ensure that the number of estimated parameters does not exceed experimental data points:

$$1 \le \sum_{i,j=1}^{n} \lambda_{ij} < n \times (p+1) \times (m-1) \tag{6}$$

where m is the number of time points, and n and p are as defined before. The above constraint implies that the number of variables to be optimized in the lower level should be less than available data thus ensuring that the problem is solvable.

2.2 Experimental materials and methods

Pancreatic organogenesis occurs in linear cascade of distinct stages starting with endoderm commitment, followed by pancreatic progenitors, endocrine progenitors and finally to mature endocrine cells. In the *in vitro* differentiation of ES cells, a similar sequence is reproduced. The mathematical analysis of regulatory network is similarly treated as a cascade of events, of which we concentrate on a single stage of differentiation, that of pancreatic progenitor commitment, and analyze it for the relevant gene expressions. Details of the differentiation protocol have been included in the Supplementary Materials.

3 RESULTS

3.1 Identification of transcription regulatory network of ES differentiation

The proposed algorithm is first validated against model *in silico* networks (Supplementary Material) before applying it to identify the regulatory architecture of a system of ESCs differentiating to the pancreatic lineage. The experimental details of the differentiation protocol are elaborated in the Supplementary Material section. Figure 1 illustrates the integrated approach we have adopted for this study. The mouse ES cells are differentiated to endoderm-like cells by co-culturing them with primary hepatocytes (Cho *et al.*, 2008). The endodermal cells were harvested from the co-culture and replated on matrigel to induce pancreatic lineage as verified by Pdx-1 expression. Early pancreatic differentiation has been reported to be inhibited by Sonic Hedgehog (Shh) signaling (Kim and Melton, 1998) which in turn can be inhibited by Cyclopamine, a known repressor of Shh. The differentiating population was treated with Cyclopamine which acts as the external perturbation described in the section above. Data from two parallel experiments were considered: (i) the control case cultured in a differentiation media and (ii) Shh inhibition by supplementing differentiation media with Cyclopamine (external perturbation). Both these conditions were analyzed for transcription factors reported to be relevant for early pancreatic differentiation.

While pancreatic organogenesis consists of an extensive network of transcription factors, reported literature establish that the network can be structured as a cascade of smaller modules, of which we will consider that of pancreatic progenitor commitment. A thorough analysis of the literature for early pancreatic markers (Habener *et al.*, 2005 and references therein) reveals 13 major transcription factors (TF) that primarily constitute the TF network at the pancreatic progenitor stage. Hence our population of ESC derived pancreatic

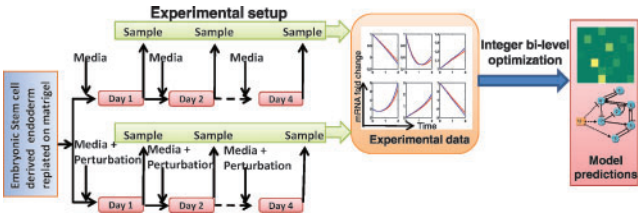

Fig. 1. Schematic representation of the integrated experimental and computational approach applied to understand the regulatory interaction governing stem cell differentiation. The differentiating cell population was sampled every day and analyzed for 13 transcription factors. A bi-level integer programming formulation is solved to identify the regulatory interactions that accurately reproduce the experimentally observed transcription factor dynamics.

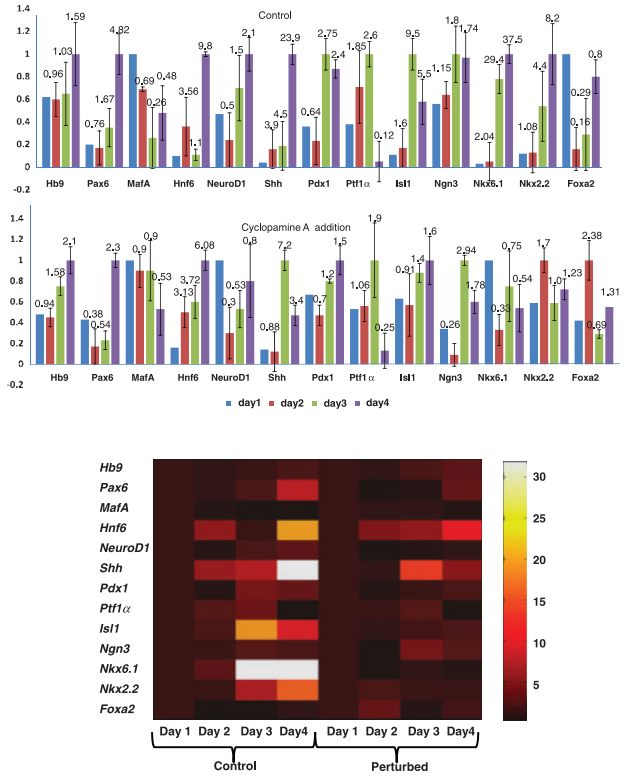

Fig. 2. Experimental data for ES cells differentiating to the pancreatic lineage, cultured in (i) DMEM, 10% FBS and (ii) DMEM, 10% FBS, KAAD Cyclopamine A (external perturbation). Cells were harvested each day and analyzed for 13 relevant TFs. The data represent the relative fold change in TFs as obtained by qRT–PCR and reported as $2^{-\Delta\Delta ct}$ values. (Top) Bar graph represents the scaled (0–1) value of relative fold change and the value above each bar represents the actual fold change. (Bottom) Colormap representation of the TF dynamics.

progenitor cells were also analyzed for these 13 TFs. Figure 2 illustrates the fold change in mRNA levels of these transcription factors over the differentiation time, both for the control and the perturbed conditions. Each transcription factor is represented as fold change of expression levels compared to day 1 of differentiation. The normalized mRNA expression dynamics serves as the input to the network identification algorithm. The bi-level optimization problem is solved with mean value of the experimental data as the input to identify: (i) 13×13 matrix **A**, the regulatory interaction between measured transcription factors including both network topology and connectivity strength, and (ii) 13×1 vector **B**, the effect of external perturbation (Cyclopamine supplementation) on the regulatory network.

Upper level topology optimization problem of the developed algorithm is formulated as an integer programming problem with 169 binary variables representing 13×13 network connectivity. The formulated integer programming problem is solved using GA. The efficiency of the algorithm depends on appropriate choice of starting population, as well as other involved parameters. The initial population size plays an important role in quality and efficiency of the algorithm. A small population size may lead to local convergence or extremely large number of generations. To avoid that a population size of 20 was chosen, and the algorithm

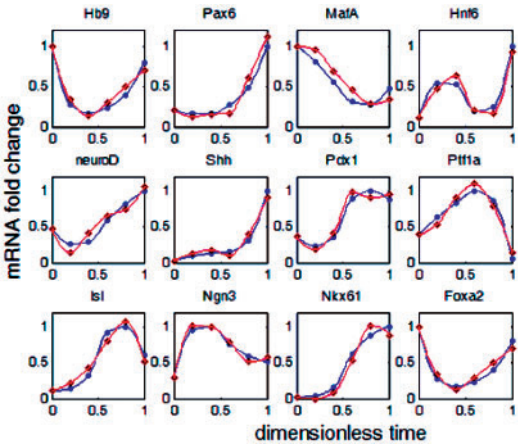

Fig. 3. Comparison of mRNA profiles predicted by the reconstructed network with the experimental data. Network is reconstructed using the proposed bi-level integer programming formulation. An error tolerance of 1.0 captures the experimental data points with excellent accuracy.

was evolved over 200 generations using a tolerance value of 1.0. The tolerance value dictates how closely the predicted profiles are required to match the experimental data. The crossover probability is chosen to be at a standard value of 0.5, and the chosen mutation probability of 0.02 was expected to maintain diversity in population. For each combination of binary variables specifying the network topology in the upper level, lower level regression problem is solved to optimize the connectivity strength against experimental data. Effect of the external perturbation is considered in the lower level as continuous variables, giving rise to a total of $13+\sum\lambda_i$ continuous variables. Dynamics of the gene expression predicted by the reconstructed network **A** is illustrated in Figure 3. Observe from this figure that the computationally predicted gene expressions show an excellent agreement with the experimental data. However, agreement of the mRNA profiles is imposed as a constraint to the lower level optimization problem, which depends on the chosen value to tolerance. Thus, although necessary, it cannot be judged as a sufficient condition indicating the accuracy of reconstruction.

Optimal reconstructed network obtained by solving the upper level integer programming problem results in 54 out of 169 connections, amounting to 68% sparsity as represented in Figure 4. The normalized strength of connectivity for each pair-wise network connection is depicted in Figure 5. A broad range in connectivity strength, varying from 0.1 to 40, is observed in the mathematically constructed network. Accordingly, the pairwise connections are categorized in three groups depending on their connectivity strength, as depicted in Figure 5. Only 9 of the 54 connections exhibited high connectivity strength in the range of 10–40 normalized values; 18 connections had a medium strength in the range of 5–10, while the rest of the 27 connections had a value lower than 5. Among the weaker connections only the ones with values higher than 1 are shown in Figure 5. In order to evaluate the sensitivity of the optimal reconstructed network to the experimental noise, a sensitivity analysis was performed by perturbing each of the experimental data points by 10% and evaluating the corresponding perturbation in the network connectivity. The bars in Figure 5 represent the overall sensitivity of each of the pair-wise connectivities to all the experimental perturbations. It is observed that the optimal network

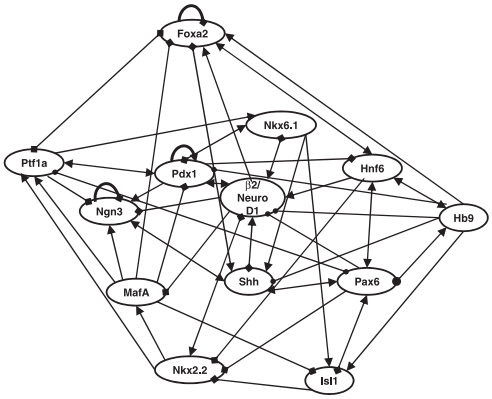

Fig. 4. The reconstructed network **A**, obtained by solving the bi-level mixed integer programming problem, maximizing sparsity at the upper level and minimizing least square error at the second level. The arrow indicates induction and the square represents repression.

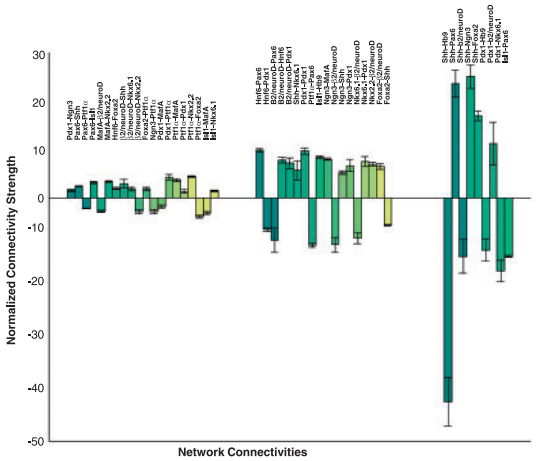

Fig. 5. Normalized connectivity strength of the optimal network reconstructed by solving the bi-level optimization problem. Connectivities with insignificant strength have been omitted from the figure. Pairwise connectivities have been clustered into three groups depending on the strength of the connectivity for ease of viewing. Bars represent the sensitivity of a specific connectivity to experimental noise.

is quite robust against experimental noise, since the sensitivity of most of the network connections remain bounded within 10% of the nominal value of the imposed perturbation.

3.2 Effect of environmental perturbation on gene network

Next, we examine **B**, the effect of external perturbation on the identified regulatory network, as illustrated in Figure 6. The strength of influence is depicted by the thickness of connecting lines, strength being directly proportional to the thickness. It is worth noting that out of the 13 elements of **B**, only three were of appreciable magnitude and the others were negligible. Overall, the strongest effect of Cyclopamine is predicted to be in the inhibition of Shh. Inhibition of Ngn3 and upregulation of Nkx2.2 are also predicted, but with much lower strength.

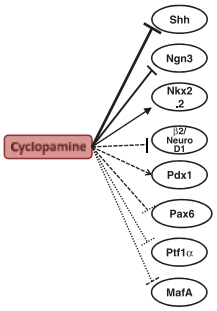

Fig. 6. The predicted effect of external perturbation (Cyclopamine) on the GRN. The thickness of the arrow depicts the strength of connection. The network reconstruction algorithm could accurately predict the most prominent effect of Cyclopamine in inhibiting sonic hedgehog gene.

In vivo studies have reported that Cyclopamine addition leads to ectopic pancreas development (Kim and Melton, 1998), resulting from the inhibition of hedgehog signaling. Both *in vivo* and *in vitro* studies have confirmed that the mechanism of Shh inhibition by Cyclopamine is indirect, resulting from blocking smoothened (Smo) function (Chen *et al.*, 2002; Kawahira *et al.*, 2003). Although such inhibitory effect of Cyclopamine on Shh is an established phenomenon, we did not provide this information *a priori* to the simulation in order to verify the predictive capability of our algorithm. As illustrated in Figure 6, our algorithm could successfully identify the effect of cyclopamine addition as being inhibition of sonic hedgehog. It is important to note that the effect of Cyclopamine on Shh is not direct, but through an indirect signaling cascade. Our model was not provided with enough details to capture the entire signaling pathway of Cyclopamine, but even then our formulation could accurately capture the resultant response of Cyclopamine on *Shh*. This is extremely crucial, since the effect of environmental perturbation on the differentiating cells is likely to be indirect. However, our primary interest is the altered functional behavior of the system in response to these perturbations which our model could predict accurately.

3.3 Identified network captures known interactions

In order to understand the relevance of the mathematically derived network model, the model predictions are compared with some of the well-established experimental studies in the literature. Analysis of the predicted network architecture reveals certain significant regulatory modes which have been reported in the literature by various investigators. Primary controlling transcription factor in the analyzed system is the pancreatic duodenal homeobox gene-1 (*Pdx-1*) which is a master regulator for both pancreatic development and maturation to b-cell phenotype (Habener *et al.*, 2005). Consequently, the reconstructed network is analyzed primarily with respect to *Pdx-1*. Figure 7 illustrates the key comparison of the predicted connections with experimentally observed connections reported in literature. The strength of the predicted connections is depicted by the thickness of the connecting arrows. Overall, an excellent agreement between our reconstructed network and literature reports can be readily observed.

During development, islet progenitors arise from Ptf1a-p48^{+}/Pdx1^{+} cells through the removal of repressive and stimulation of inductive pathways. Ptf1a has also been reported to directly bind

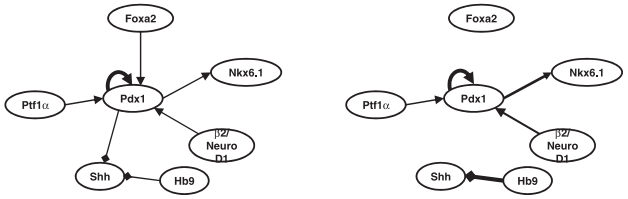

Fig. 7. Comparison of the regulatory interactions of *Pdx-1* as reported in literature (left) with the reconstructed network (right). The strength of interactions in the predicted network is depicted by the thickness of the arrow. The reconstructed network could adequately capture many of the known interactions, without *a priori* information.

to Pdx1 promoter (White *et al.*, 2008). Our differentiation scheme results in strong upregulation of *Ptf1a* and the network analysis reveals the positive induction of *Pdx1* expression by *Ptf1a*.

Pdx1 has also been reported to be directly activated by the transcription factors *NeuroD1* (Cerf, 2006) and *Foxa2* (Ben-Shushan *et al.*, 2001). The positive induction of *Pdx1* by *NeuroD1* is accurately captured by the model, while the effect of *Foxa2* is more indirect. In addition, the strong autoregulatory effect of *Pdx1* in inducing its own transcription is an established phenomenon. The model prediction correctly captures the strong autoregulatory effect of *Pdx1* as well.

Targeted disruption of the *Pdx1* gene in β cells leads to reduced activities of *Pdx1* regulated genes such as *Nkx6.1* and *Glut2* (Holland *et al.*, 2002). A similar phenomenon of *Pdx1* induction of the *Nkx6-1* gene is also observed in the model prediction. As discussed before, one of the earliest events in pancreatic organogenesis is the repression of Shh by the notochord, which in turn promotes *Pdx1* expression in adjacent pancreatic endoderm (Soria, 2001). In parallel, the combined action of *Pdx1* and *Hb9* inhibits *Shh* expression. The model captures a similar effect of *Shh* inhibition by *Hb9*, over and above the suppression of Shh by Cyclopamine.

Endocrine cells originate from lineage-committed progenitors marked by the helix-loop-helix transcription factor neurogenin 3 (*Ngn3*) (Gradwohl *et al.*, 2000). *Ngn3* has also been shown to negatively regulate its own promoter, providing a potential mechanism for self-inactivation and explaining its transient expression during pancreatic development (Smith *et al.*, 2004). The negative auto-regulatory effect of *Ngn3* is correctly predicted in the simulated network.

Nkx2.2 also drives endocrine differentiation and is controlled by alternative promoters at different cellular stages (Watada *et al.*, 2003). During progenitor and endocrine cell stages, *Ngn3* and *NeuroD1* have been shown to activate *Nkx2.2* respectively. Present mathematically derived model identifies *NeuroD1* as the inducer of *Nkx2.2* but not *Ngn3*, suggesting the differentiation stage being endocrine cellular state. The inactivation of *Nkx2.2* gives rise to endocrine-like cells lacking Insulin or *Glut2*, but expressing other endocrine markers such as Amylin and *Isl1*. The current model also indicates that Isl1 may have some effect in negative regulation of *Nkx2-2*. White *et al.* (2008), attempted to identify the regulatory structure of pancreas development and reported the positive effect of *β2/NeudoD1* in the upregulation of *Nkx2-2* and *Foxa2*. *NeuroD1* was also shown to bind to *MafA*. Observe that all these effects of *NeuroD1* are captured in the present reconstructed network.

These results indicate that the proposed algorithm developed on the notion of sparsity of biological networks can successfully extract from the experimental data many of the known interactions which have been independently reported in literature. It is important to note here that the reconstruction algorithm relies on the subset of the transcription factors used in the input data points. We demonstrated with the case of *Shh* inhibition that even in the absence of precise details of intermediate steps, the model could adequately capture the overall behavior of the system. Thus the quality and resolution of the reconstructed network will largely depend on the input data provided to the model.

3.4 Network prediction and experimental validation

While the mathematically reconstructed network agrees well with literature reports on pancreatic developmental networks, the full potential of the model can only be exploited in its predictive capability. The primary purpose of determining regulatory networks governing differentiation is to enable informed protocol design in deriving specific cellular phenotypes. This feat can only be achieved through a combination of successful modeling and concurrent experiment.

The current network is determined for the pancreatic progenitor stage, which precedes the endocrine progenitor stage in pancreatic differentiation. While *Pdx1* controls the differentiation of pancreatic progenitor stage, *Ngn3* is the primary signature gene controlling endocrine progenitor differentiation. In order to verify the predictive capacity of the developed model, we use the reconstructed model to identify a pathway which will significantly up-regulate *Ngn3* expression, thereby inducing endocrine differentiation.

This prediction is achieved by solving Equation (2) with mathematically derived **A**and **B** and a proper choice of **U**. Thus, the effect of silencing the *i*-th gene will be predicted by adjusting u_j, components of **U** as

$$u_j = 0\big|_{j=1,n;j\neq i};\ u_i = -S$$

where $-S$ represents appropriate downregulation of the *i*-th gene.

No other experimental data are used as model input in this stage, which is designed to be completely predictive in nature. This exercise identifies down-regulation of *Foxa2* to be a likely mechanism in up-regulation of *Ngn3* expression levels. In the absence of literature reports relating such an interaction, the validity of this prediction is verified by performing concurrent experiments by silencing *Foxa2* gene in the differentiating stem cell population. Figure 8a represents the dynamic response of the system to *Foxa2* down-regulation and compares the model predictions with experimental observations. Figure 8b illustrates the colorbar representation of the comparison, the predicted versus actual effect of *Foxa2* silencing on the population of differentiating ES cells. Figure 8b clearly shows that the most significant effect of *Foxa2* silencing is the up-regulation of *Ngn3* (~10-fold). This observation compares extremely well with the model prediction of 8-fold up-regulation. *Foxa2* silencing also resulted in significant down-regulation of *MafA* genes, which is also correctly predicted by our reconstructed network, although the magnitude of down-regulation is somewhat underpredicted.

Silencing of Foxa2 had less dramatic effect on many of the other genes, although most of them were affected to some extent. Quite encouragingly, our model could accurately predict most of these

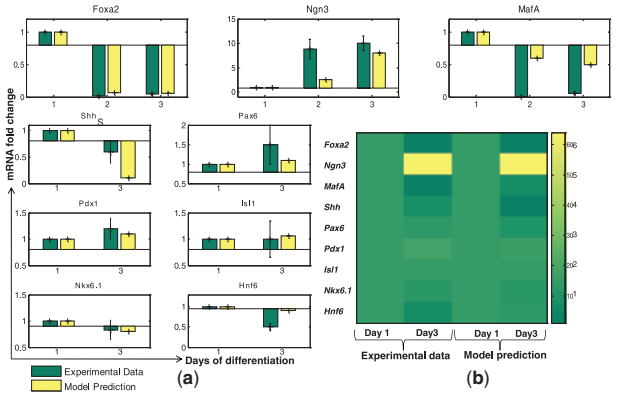

Fig. 8. Effect of silencing of Foxa2 gene on the other components of the regulatory network. (**a**) Comparison of the model predicted effects of Foxa2 down-regulation with experimental data of Foxa2 silencing. The time points analyzed are: 24 h, 36 h and 52 h after initial transfection, depicted as 1, 2 and 3 in the *x*-axis, respectively. (**b**) Colormap comparison of fold changes in gene expression levels between reconstructed model and experimental observation.

effects in both nature and strength. For example, both Pax6 and Pdx1 were up-regulated by Foxa2 silencing, as also predicted by the network. Shh and Hnf6 were both down-regulated by Foxa2 silencing, which is correctly predicted by the model in nature but not in magnitude. The model overpredicts the down-regulation of Shh, while underpredicts the down-regulation of Hnf6. Both isl1 and Nkx6.1 were quite insignificantly affected by the Foxa2 silencing, which again is accurately predicted by the model as well.

Overall, we observe that the mixed integer bi-level programming approach based on the notion of succeeds in accurately capturing the most important and significant connections in the system, although prediction of the weaker connections can be less accurate. The proposed method can determine sufficiently accurate networks with very restricted experimental data, a feature which makes it extremely attractive to stem cell based applications.

4 DISCUSSION

We have developed a bi-level mixed integer programming formulation to reliably identify the GRN from a data-limited scenario. The developed framework can accurately capture the regulatory interactions solely from the data of gene expression profiles, without any *a priori* information regarding regulatory interactions between transcription factors. Our mathematical formulation is based on the understanding that a sparse network reproducing the transcription factor profile will be the optimal one chosen by nature. The network architecture derived by promoting sparsity could adequately predict the experimentally observed regulations reported in the literature. It could also accurately predict the perturbation required to induce subsequent differentiation, an outcome confirmed by concurrent experiments.

The GRN determined by the proposed bi-level methodology could reliably capture the known effects of Shh inhibition of Cyclopamine, along with key governing features of the pancreatic organogenesis. The reconstructed network had excellent predictive capability even outside the domain of experimental data used to determine the network. It could accurately predict the effects of *Foxa2* silencing on up-regulation of *Ngn3* expression, which would induce the next cascade of pancreatic differentiation. Such an effect has not been reported in literature and concurrent experiments validated the nature and even the predicted magnitude of *Ngn3* up-regulation. As detailed in Figure 8, the model prediction could accurately capture the most significant effect of Foxa2 silencing on the genes considered in the network. The prediction of the magnitude of effect had different degrees of accuracy; but the most significant effect, that of *Ngn3* up-regulation, was accurately predicted. This is significant in establishing confidence in the presented reverse engineering approach based on the notion of sparsity, and its usefulness in assisting *in-vitro* differentiation to derive specific cellular phenotypes of interest.

The proposed method offers significant advantage in analyzing differentiating cell population, where the cells are typically induced to a specific lineage by exposing them to different environmental conditions with respect to extracellular matrix, growth factors and chemical inducers or repressors. The proposed method can efficiently utilize this information in analyzing the governing network and does not rely on more laborious gene knockout data. This is even more pertinent in stem cell studies since knockout of certain key genes can have severe consequences with respect to cell survival and differentiation. However, in the event of availability of such data it can be easily incorporated into the network identification formulation. Moreover, the bi-level formulation ensures efficient utilization of experimental data by essentially reducing the number of variables being optimized in the inner loop.

Since the proposed algorithm extracts regulatory information from the dynamic profile of transcription factors, the choice of the transcription factors and the quality of the input experimental data will play a significant role on the reliability of the predicted network. It is crucial to consider all salient transcription factors important for the differentiation stage under consideration. Since the formulation will have no information of the excluded TFs, it is likely that any strong effect of the excluded TFs will be reported as spurious interactions of the existing network. However, as illustrated in the case of Cyclopamine A, the identified network will still capture functional interactions even in the absence of all the intermediate TFs involved in such interaction. The resolution of the identified network will thereby be largely dependent on the input experimental data. However, this interaction will be negligible if the excluded TF have only a mild effect. The mathematical framework proposed in this paper is deterministic in nature, and does not consider the experimental variability and parameter uncertainty in the network prediction. Even then, the posterior sensitivity analysis illustrated in Figure 5 indicates an acceptable level of sensitivity of network connections to the experimental uncertainty.

The developed methodology has been illustrated in the context of differentiating population of embryonic stem cells to pancreatic lineage. However, the mathematical framework is general enough to be used to analyze the differentiation of stem cells, embryonic or adult, to any lineage. Although there is considerable information regarding specific role of transcription factors at different stages of organogenesis, very little information is available at present on how transcriptional networks are organized within these cells. The proposed methodology will be instrumental in identifying such transcriptional networks and the environmental effect on such networks, which will have potential application in designing *in silico* protocols for stem cell differentiation. The proposed scheme holds

the promise of significant impact in efficient protocol development for directed differentiation of embryonic stem cells. This approach will be particularly useful in identifying regulatory networks in data-limited systems like stem cell differentiation and developmental systems in general.

The presented formulation considers transcription factor profiles as input data, whereas the actual regulatory architecture results from a complex interplay of genes, proteins, signaling molecules, etc. Current efforts are under way to extract experimental data of the relevant protein expressions and incorporate that in the network architecture.

ACKNOWLEDGEMENTS

We thank the reviewers for helpful and constructive comments.

Funding: This work was partly supported by NIH (DP2-116520) and start-up funds from University of Pittsburgh to IB.

Conflict of Interest: none declared.

REFERENCES

Bansal,M. *et al.* (2006) Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics*, **22**, 815–822.

Ben-Shushan,E. *et al.* (2001) A pancreatic β-cell-specific enhancer in the human Pdx-1 gene is regulated by HNF-3β, HNF-1α, and SPs transcription factors. *J. Biol. Chem.*, **276**, 17533–17540.

Blais,A. *et al.* (2005) Constructing transcriptional regulatory networks. *Genes Dev.*, **19**, 1499–1511.

Bolouri,H. and Davidson,E.H. (2003) Transcriptional regulatory cascades in development: initial rates, not steady state, determine network kinetics. *Proc. Natl Acad. Sci. USA*, **100**, 9371–9376.

Cerf,M.E. (2006) Transcription factors regulating beta cell function. *Eur. J. Endocrinol.*, **155**, 671–679.

Cho,C.H. *et al.* (2008) Homogenous differentiation of hepaocyte-like cells from embryonic stem cells: applications for the treatment of liver failure. *FASEB J.*, **22**, 898.

Davidson,E.H. (2001) *Genomic Regulatory Systems: Development and Evolution*. Academic, San Diego.

Davidson,E.H. *et al.* (2002) A genomic regulatory network for development. *Science*, **295**, 1669–1678.

DeWeese,M. *et al.* (2003) Binary spiking in auditory cortex. *J. Neurosci.*, **23**, 7940–7949.

Foteinou,P. *et al*. (2009) A mixed-integer optimization framework for the synthesis and analysis of regulatory networks. *J. Glob. Optim.*, **43**, 263.

Gradwohl,G. *et al.* (2000) Neurogenin3 is required for the development of the four endorine cell lineages of the pancreas. *Proc. Natl Acad. Sci. USA*, **97**, 1607–1611.

Habener,J.F. *et al.* (2005) Minireview: transcription regulation in pancreatic development. *Endocrinology*, **146**, 1025–1034.

Hartemink,A.J. *et al.* (2002) Combining location and expression data for principled discovery of genetic regulatory network. *Pacific Symp. Biocomput.*, **7**, 437–449.

Holland,A.M. *et al.* (2002) Experimental control of pancreatic development and maintenance. *Proc. Natl Acad. Sci. USA*, **99**, 12236–12241.

Kim,S.K. and Melton,D.A. (1998) Pancreas development is promoted by cyclopamine, a hedgehog signaling inhibitor. *Proc. Natl Acad. Sci. USA*, **95**, 13036–13041.

Koide,T. *et al.* (2005) Xenopus as a model system to study transcriptional regulatory networks. *Proc. Natl Acad. Sci. USA*, **102**, 4943–4948.

Lee,H. *et al.* (2007) Efficient sparse coding algorithms. *Proceedings of the 19th Annual Conference on Neural Information Processing Systems*. MIT Press, Cambridge, MA, pp. 801–808.

Ljung, L. (1999). *System Identification: Theory for the User*. Prentice Hall PTR, Upper Saddle River, NJ.

Oliveri,P. and Davidson,E.H. (2004) Gene regulatory network controlling embryonic specificatoin in the sea urchin. *Curr. Opin. Genet. Dev.*, **14**, 351–360.

Papadimitriou,C.H. and Steiglitz, K. (1998) *Combinatorial Optimization: Algorithms and Complexity*. Dover, Mineola, NY.

Segal,E. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.*, **34**, 166–176.

Shmulevich,I. *et al*. (2002) gene perturbation and intervention in probabilistic Boolean networks. *Bioinformatics*, **18**, 1319–1331.

Singh,H. *et al.* (2005) Contingent gene regulatory networks and B cell fate specification. *Proc. Natl Acad. Sci. USA*, **102**, 4949–4953.

Smith,S.B. *et al.* (2004) Neurogenin3 activates the islet differentiation program while repressing its own expression. *Mol. Endocrinol.*, **18**, 142–149.

Stathopoulos,A. *et al.* (2005) Genomic regulatory networks and animal development. *Dev. Cell,* **9**, 449–462.

Tegner,J. *et al.* (2003) Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. *Proc. Natl Acad. Sci. USA*, **100**, 5944–5949.

Vinje,W.E. and Gallant,J.L. (2000) Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, **287**, 1273–1276.

Wagner,A. (2001) How to reconstruct a large genetic network from n gene perturbations in fewer than n^2 easy steps. *Bioinformatics*, **17**, 1183–1197.

Watada,H. *et al.* (2003) Distinct gene expression programs function in progenitor and mature islet cells. *J. Biol. Chem.*, **278**, 17130–17140.

Yao, X. (ed.) (1999) *Evolutionary Computation: Theory and Applications*. World Scientific Publishing, Singapore.

Yeung,M.K.S. *et al.* (2002) Reverse engineering gene networks using singular value decomposition and robust regression. *Proc. Natl Acad. Sci. USA*, **99**, 6163–6168.