Proceedings of the 31st Chinese Control Conference
July 25-27, 2012, Hefei, China

Inferring Gene Regulatory Network for Cell Reprogramming

DUREN Zhana[1,2], WANG Yong[1,*], SAITO Shigeru[3,4], HORIMOTO Katsuhisa[3]

1 Academy of Mathematics and Systems Science, National Center for Mathematics and Interdisciplinary Sciences, Chinese Academy of Sciences, Beijing 100190, China

2 School of Mathematics and Systems Science, Beihang University, Beijing 100191, China

3 Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, 135-0064, Japan

4 INFOCOM CORPORATION, Tokyo, 150-0001, Japan

E-mails: durenzn@yahoo.cn, ywang@amss.ac.cn, sh.saito@infocom.co.jp, k.horimoto@aist.go.jp

Abstract: The remarkable discovery of induced pluripotent stem cells (iPSCs) demonstrates that cell can be reprogrammed from somatic cell to a pluripotent state by the enforced expression of defined transcriptional factors. However, the underlying mechanism for cell reprogramming remains unknown and the regulatory interactions within this biological process have not been worked out. In particular from the gene regulatory network perspective, it is not clear how the four factors initialize the reprogramming process, propagate the information in a fine tuned way, and finally lead to the dramatic phenotype changes. In this paper, we analyze the time course gene expression data during cell reprogramming in mouse. We propose a three-stage procedure to infer gene regulatory networks. Specifically, we identify the major players during cell reprogramming by selecting differentially expressed genes in the first stage. Then in the second stage we utilize a new method to reveal strong correlations among those selected genes from short time series data. Finally the gene regulatory relationships are modeled by ordinary differential equations (ODE), the correlations are filtered by applying strong regularization, and directed and signed gene regulatory network for cell reprogramming is reconstructed. Preliminary analysis of the inferred network shows that short time series data provide biological insights for the dynamical process during reprogramming.

Key Words: Gene regulatory network, Reconstruction, Induced pluripotent cell, Cell reprogramming, Time series data

1 Introduction

Generation of induced pluripotent stem cells (iPSCs) from somatic cells demonstrates that cell can be reprogrammed to a pluripotent state by the enforced expression of defined factors (Sox2, Oct4, Klf4, and c-Myc) [1]. Similar to the embryonic stem cells (ESCs), iPSCs have the ability of self-renewal and differentiation and can be potentially used on maintaining the growth of human organs and metabolism, repairing the body's aging and diseasing. Furthermore, iPSCs do not have restrictions on the ethics and source materials. Therefore, this remarkable discovery has attracted great attention for its potential applications to drug screening and analyses of disease mechanisms, and even as next generation materials for regenerative medicine [2]. However, understanding the mechanism underlying cell reprogramming is one of the key steps before safely moving to clinical applications. In addition, the mechanism under cell reprogramming provides far-reaching implications for biological sciences [3].

In this paper, we aim to understand the cell reprogramming mechanism from gene regulatory network's perspective. The reasons are in two folds. First, gene regulation is one of the dominant factors for the mechanistic picture of cell reprogramming. It's well-known that reprogramming is initialized by introducing four transcriptional factors (TFs), which will activate or depress thousand of targets and then trigger the dramatic change of gene expression landscape. If we can reconstruct the regulatory interactions during the reprogramming process, we then know how these TFs regulate each other and interact with epigenetic control factors to form a large gene regulatory network. Further it will help to understand that how the four factors initialize the reprogramming process, propagate the information in a fine tuned way, and finally lead to the dramatic phenotype changes. Secondly, a large amount of data has been accumulated in transcription level and makes the inference of large scale gene regulatory network feasible. As we know, DNA microarrays servers as a widely used and standard techniques to measure the expression levels of large numbers of genes simultaneously. Gene expression profiles have been used to compare the difference of iPSCs and ESCs in transcription level and may be further utilized to standardize human iPSCs. More importantly, those well-designed expression microarray experiments can bring us rich information under cell reprogramming and help to reveal the causal regulatory relationships among genes.

Then the problem is how to reconstruct the gene regulatory networks in an accurate and reliable manner by analyzing the available gene expression data for induced pluripotent stem cells. Currently, there are several existing efforts to study gene regulatory networks for cell reprogramming. One direction is to apply the biological technologies to obtain the location (ChIP-chip/ChIP-Seq) data sets to reconstruct a part of this network [4]. Another way is to collect evidences from literature and manually construct a genetic network involved in regulating pluripotency and differentiation [5]. Recently, a novel framework called "network screening" has been applied to detect the active subnetworks for cell reprogramming by integrating ChIP-chip data or existing molecular interactions with conditional gene expression data [6]. These reconstructed

* Corresponding author. This work is supported by National Natural Science Foundation of China (NSFC) under Grant 61171007 and 11131009.

networks demonstrate their power to reveal important biological insights, such as to find new important genes in reprogramming, to reconstruct cell reprogramming landscape, and to systematically search recipes to improve reprogramming efficiency. However, the limitations of those reconstructed networks are also clear. Firstly, those networks are small in scale and are only part of the whole-genome network. They can only offer very limited information since cell reprogramming leads to about 10,000 differentially expressed genes [6]. A large, even whole genome, regulatory network is necessary for a mechanistic picture. Secondly, currently all the above networks are reconstructed by using static data, i. e., the gene expression data and location data are measured after the cells get the pluripotency. As a result, those networks fail to capture the dynamic process during the induction. Here, we will employ mathematical modeling or systems biology methods in reconstructing whole-genome regulatory networks. Particularly, we will computationally analyze newly published time course gene expression data during cell reprogramming.

Gene regulatory network inference is a widely-studied topic in systems biology. There are recently a lot of network inference methods that infer gene regulatory network by information-theoretic methods, Bayesian network predictions, and ordinal differential equation (ODE) models [7-9]. However, reliable gene network inference from gene expression data with short time course measurement remains a challenging and unsolved problem. In this paper, we propose an improved differential equation models based method, which is specifically designed for inferring gene regulatory networks from short time course data of somatic cell reprogramming.

2 Method

2.1 Overview of the method

In this paper, we analyze the time course gene expression data during cell reprogramming in mouse by proposing a three-stage procedure to infer gene regulatory networks. One of our main motivations is to take full advantage of all the available information in time course gene expression data, given the dilemma that data is scarce regarding to the large parameter space for network representation. The schematic plot of the proposed gene regulatory network inference method is illustrated in Figure 1. As the red arrows shows, we identify the major players during cell reprogramming by selecting the differentially expressed genes in the first stage. Then in the second stage we utilize a new method to reveal the strong correlations among the selected genes from short time series data. The directed and signed gene regulatory relationships are modeled by ordinary differential equations (ODE) and gene regulatory network for cell reprogramming is obtained in the third stage.

2.1 Stage I: selecting differential expressed genes

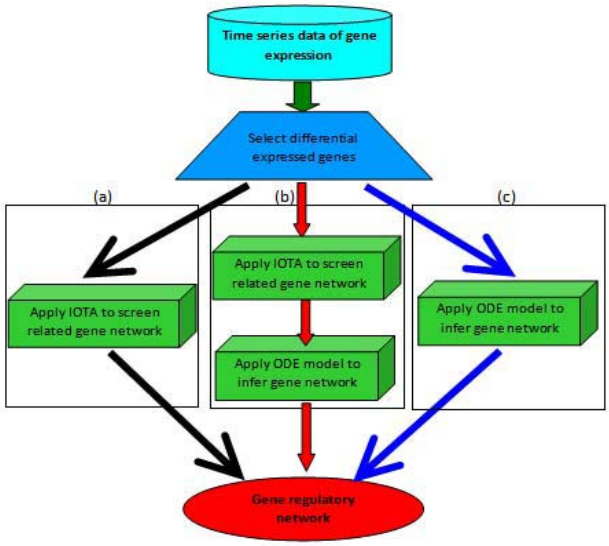


Figure 1. Schematic plot of the proposed gene regulatory network inference method. Our method follows a three stage procedure (indicated by red arrows). At the first stage, we select differentially expressed genes during cell reprogramming. Then the correlations among those genes are calculated. Finally we apply ODE model to infer network. Our method outperforms the direct correlation based method (indicated by black arrows) and direct ODE model based method (indicated by blue arrows).

In the process of cell reprogramming, some genes could be differential expressed and some are not. In our method, we only consider the differentially expressed genes during reprogramming. The genes which are not differential expressed are naturally assumed as insignificant for the cell differentiation process. Our assumption is based on the following three observations. Firstly, a biological gene network is expected to be sparse, in other words, expression of a gene shouldn't be regulated by many genes. Secondly, the large number of genes involved in the modeling makes the inferred network not accurate. Thirdly, reverse engineering such large network is time consuming. Therefore, the first stage of our method is to filter out the genes which are not differential expressed.

The variance is a standard way to measure if a gene is differential expressed or not. We select the gene if variance of the time course profile is larger than a pre-defined threshold. Otherwise, we filter out this gene. As a result, we filter out many genes which are not significantly changed from the time course data and only treat the relationships among the set of genes expressed differentially.

2.2 Stage II: screening co-expressed gene pairs

Although we filter out many genes by Stage I, there are still a lot of remaining genes due to the fact that reprogramming is such a dramatic change inside cell. Given a gene, there are still many potential regulators. At Stage II, we narrow down the number of possible regulators for each gene by finding out their correlated genes during cell reprogramming.

It's a challenging task to detect correlations from a short time course data. In our case, we deal with time series gene expression data, in which a temporal process is measured by time series expression experiments, which exhibit a strong autocorrelation between successive points and escape the

standard association analysis methods. Here we apply a recent method, named the inner composition alignment (IOTA) [10], to identify the coupling and its directionality among genes. In IOTA, a correlation coefficient τ is defined to characterize the correlation of two time series. For the given time series $y^{(l)}$ and $y^{(k)}$, we sort them with the same time order $\varphi^{(l)}$, where $\varphi^{(l)}$ is the permutation which orders $y^{(l)}$ in a non-decreasing order, i.e., $\varphi^{(l)}$: $\forall i \quad [y^{(l)}(\varphi^{(l)})]_i \leqslant [y^{(l)}(\varphi^{(l)})]_{i+1}$. The series $g^{(k,l)} = y^{(k)}(\varphi^{(l)})$ is the reordering of the time series $y^{(k)}$ with respect to $\varphi^{(l)}$. We define the correlation coefficient τ as follows,

$$\tau = 1 - \frac{\sum_{i=1}^{m-2}\sum_{j=i+1}^{m-1}\omega_{ij}\theta[(g_{j+1}^{(k,l)} - g_i^{(k,l)})(g_i^{(k,l)} - g_j^{(k,l)})]}{\nabla} \quad (1)$$

Where m is the length of the time series, $\nabla = \frac{m(m-1)}{2}$ is a normalization constant which corresponds to the maximum number of crossings, ω_{ij} is the weigh and the θ[x] is the Heaviside step function:

$$\Theta[x] = \begin{cases} 1, x > 0 \\ 0, x \leq 0 \end{cases}$$

For the given two time series profiles, the larger the τ, the more related the two genes. In this way, we can rank the potential regulators for each single gene.

2.3 Stage III: applying ODE model to infer the final gene regulatory network

In Stage II, we find all the potential regulators for each gene by checking their correlations in time series data. However, we cannot distinguish if these correlations are casual relationship, and further we do not know the detailed regulatory role (activating or repressing, or inhibiting or inducing) of these correlations. In this subsection, we consider to use the ordinary differential equation model to capture the dynamic relationships among genes [11-14]. A linear differential equation can be used to represent the rate of synthesis of a transcript as a function of the concentrations of other transcripts in a cell and the external perturbations:

$$\frac{d\mathbf{X}}{dt} = J\mathbf{X} + P\mathbf{C} \quad (2)$$

where $X=(x(t_1),...,x(t_m))$ and $dX/dt=(dx(t_1)/dt,...,dx(t_m)/dt)$ are $n \times m$ matrices with the first derivative of mRNA concentration $dx_i(t_j)/dt=[x_i(t_{j+1})-x_i(t_j)]/[t_{j+1}-t_j]$ for i=1,...,n; j=1,...,m is the forward difference approximation. Suppose that there are s external perturbation compounds, then $C=(c(t_1),...,c(t_m))$ is an $s \times m$ matrix representing the s perturbations. In our task, the information of external perturbations is not available and we only focus on regulations among genes, thus we assume that there is no external perturbation. Therefore the unknowns to be calculated are connectivity matrix J. The regulatory relationships can be directed, signed, and weighted. For example, element J_{ij} represents an effect of gene j on gene i, while J_{ji} represents an effect of gene i on gene j. In this way the influence between gene i and gene j is directed. Furthermore, a sign associated with J_{ij} represents a specific role of regulation. For example, if the sign of J_{ij} is positive, gene j is the activator of gene i. On the other hand, if the sign of J_{ij} is negative, gene j is the repressor of gene i. Furthermore the associated weight (the absolute value) of element J_{ij} indicates how strong the regulatory interaction is. Obviously, a zero weight of J_{ij} indicates no interaction between two genes.

Assume that there are N measurement during cell reprogramming with $m_1, m_2, \ldots, m_N$ time points respectively. A biological gene network is expected to be sparse [11], which should also be reflected in the procedure of the network reconstruction. By solving the optimization problem as follows we can infer a sparse gene network.

$$\min_J \frac{1}{2}\left\|\frac{d\mathbf{X}}{dt} - J\mathbf{X}\right\|_2 + \lambda\|J\|_0 \quad (3)$$

Where $\|J\|_0$ is the L_0 norm of matrix J, i.e., the number of non-zero elements in matrix J. $\lambda>0$ is the regularization parameter, and the relative weight of the two terms is controlled by λ. The larger the λ, the sparser the solution is. Unfortunately, the L_0 norm minimization problem is NP-hard and it is difficult to be solved in an efficient way. Therefore we consider some algorithms which can compute an approximate solution. Under certain conditions the L_0 norm can be approximated by the L_1 norm, leading to a convex optimization problem [11].

$$\min_J \frac{1}{2}\left\|\frac{d\mathbf{X}}{dt} - J\mathbf{X}\right\|_2 + \lambda\|J\|_1 \quad (4)$$

We use $\mathbf{X}_i$ represents the time series data of the i-th gene, and use J_i represents the gene regulatory relationships for gene i. So for each single gene, Eq. (4) can be decomposed as subproblems for each gene as follows:

$$\min_{J_i} \frac{1}{2}\left\|\frac{d\mathbf{X}_i}{dt} - J_i\mathbf{X}\right\|_2 + \lambda\|J_i\|_1 \quad (5)$$

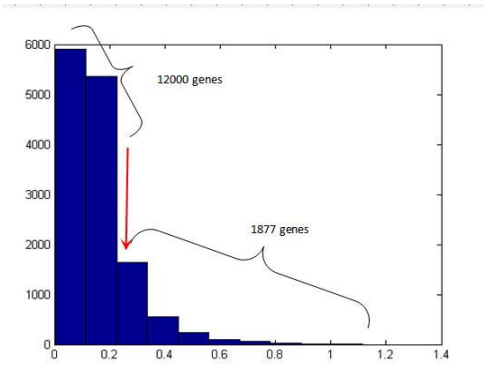
The optimization problem (5) is an L_2-L_1 norm optimization problem, which is a well-studied LASSO model in statistics. Generally the optimal solution of (5) sets as many elements of J_i to zero as possible, thus ensuring a consistent and sparse structure for the inferred gene regulatory network. Here the final output $J=(J_{ij})_{n\times n}$ is an n×n connectivity matrix with elements J_{ij} representing the effect of gene j on gene i with a positive, zero, or negative sign, indicating activation, no interaction, and repression, respectively.

3 Results

In this section we report the preliminary results on inferring gene regulatory network by analyzing short time course gene expression data for cell reprogramming. The expression data is from [15] and measure throughout reprogramming of MEF to iPS using a Dox-inducible promoter. In this data, MEFs were treated with Dox in mES media to turn on the Oct4, Klf4, cMyc, Sox2. Total RNA was extracted at day 0 (no Dox), day 2, 5, 8, 11, 16, and 21 (with Dox) and day 30

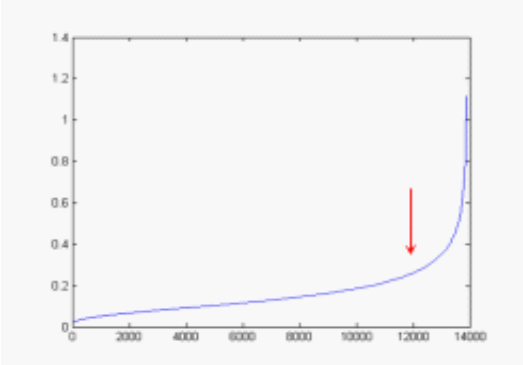
(Dox-independent secondary iPS). Therefore the data we used for network inference is time series data of 13,877 genes at 8 time points. Temporal analysis of this time course data already revealed that reprogramming is a multi-step process that is characterized by initiation, maturation, and stabilization phases. In this paper, we perform further analysis on this dynamic data to understand the process of reprogramming and in particular the gene regulatory network that control progression to a stable pluripotent state.

We apply our three-stage method to the time series data. In the first step, we select differentially expressed genes from the total 13,877 genes (those genes with gene expression data and has corresponding orthologs in human) in mouse. We compute the variance of the gene expression values across 8 time points for each gene. The results are shown in Figure 2. Figure 2 (a) plots the histogram of the variance values for all genes. We find that about 6,000 genes don't change too much during cell reprogramming. While 1,877 genes have relatively large variances (>0.2). Here the threshold 0.2 is selected based on the sharp turn in the curve of ranked variance values in Figure 2(b), where all the variances are ranked in ascending order. Finally, we consider the regulatory relationship among the remaining 1,877genes. The top 20 genes with the largest variances are listed in Table 1. The pluripotency markers, NANOG and POU5F1, are differentially changed during reprogramming.



In addition, TACSTD1(EPCAM) and TDGF1(CRIPT) are known as important ES factors. These facts partially demonstrate that our gene selection procedure is reasonable.

(a)



(b)

Figure 2.In total 1,877 differentially expressed genes are selected from time course gene expression data. (a) Histogram of the variance for all genes (x-axis denotes variance value and y-axis denotes the number of genes). 12,000 genes are filtered since their variance of the expression data are less than 0.2. (b). threshold 0.2 is chosen for the sharp turn in the curve of ranked variance values (y-axis denotes variance value and x-axis denotes the rank of genes).

Gene Name	Entrez Gene ID
CLDN4	1364
KRT16	3868
FETUB	26998
UTF1	8433
EHF	26298
AVIL	10677
CYP4F22	126410
PLA2G1B	5319
ALDH3A1	218
NANOG	79923
POU5F1	5460
LUM	4060
LCN2	3934
KRT6A	3853
NR0B1	190
LY6G6C	80740
MAL	4118
TACSTD1	4072
TDGF1	6997
BEX1	55859

Table 1. The list of top 20 genes with the largest variances in gene expression during cell reprogramming.

In the second step, we calculate the correlations among the 1,877 genes. The correlation coefficient τ, defined in Equation (1) is computed for every two genes, and finally we get a correlation coefficient matrix with size 1,877× 1,877. In Figure 3, we show the procedure to calculate τ for a pair of genes with coefficient 0.9539.

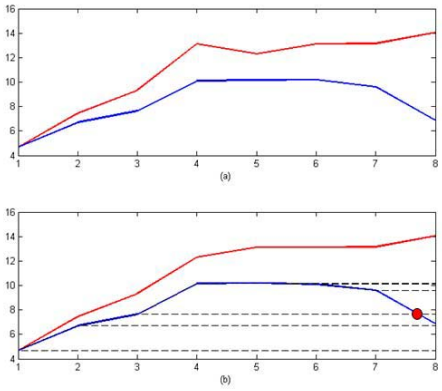


Figure 3.An example to calculate correlation coefficient τ for two genes with coefficient 0.9539. The upper subfigure shows the original time series profiles for two genes. The bottom subfigure shows the reordering of the two time series according to the expression value rank of the gene in red curve. Then τ is calculated under the new order 1,2,3,5,6,4,7,8 by counting the crossings (highlighted by red dot).

Figure 4 shows the results of the correlation coefficient τ for all the gene pairs. The coefficient τ shows an approximate normal distribution with mean value 0.85, which is slightly high because of pre-selection of a small set of differentially expressed genes and the shortage of time points. Then we carefully choose 0.95 as the threshold (see the legend in Figure 4) to select gene pairs with high correlations, which will be feed into the ODE model to further find the true regulatory interactions. In other words, we consider the two genes are related if their correlation coefficient τ is larger than 0.95. For example, we select 305 related genes correlated with BEX1. In total we selected 750,000 gene pairs with high correlations. Averagely, every gene has 400 genes with high correlations as potential regulatory interactions. This number is not evenly distributed. We show in Figure 4 the correlation coefficient distribution for Nanog with other genes. It shows that Nanog has high

correlations with about 700 genes. In Table 2, the top 20 genes with high correlations with Nanog, Pou5f1, and Sox2 are listed.

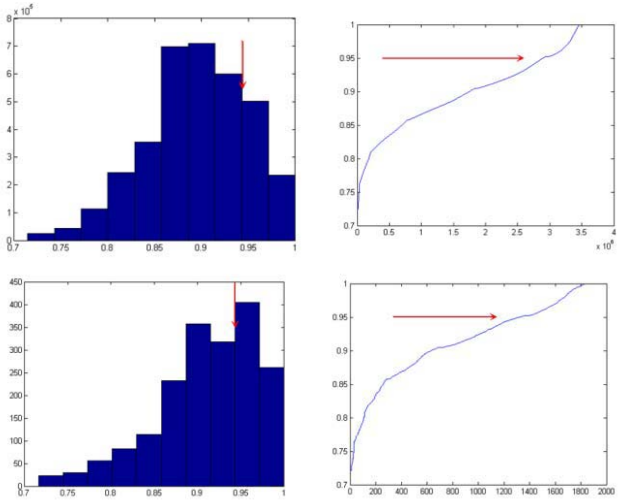


Figure 4.Results of the τ for all the gene pairs. The upper left subfigure shows the histogram of all the correlation coefficients (x-axis denotes correlation coefficient and y-axis denotes the number of pairs). The upper right subfigure illustrates the procedure to select 0.95 (red arrows) as a cutoff. All the pairs are ranked by their coefficients (y-axis denotes correlation coefficient and x-axis denotes the ranks of the pairs). The two subfigures in bottom show the results of the τ for the Nanog with the other genes.

Table 2. The list of top 20 genes with the largest correlations with Nanogo, Pou5f1, and Sox2 during cell reprogramming.

Top 20 genes correlated with Nanog	Correlation	Top 20 genes correlated with Pou5f1	Correlation	Top 20 genes correlated with Sox2	Correlation
TJP2	1	ZNF175	1	NXF2	0.997712626
NIPSNAP1	1	B4GALNT4	1	LRRC2	0.997781886
MFAP5	1	GLUL	1	EFHC2	0.997799354
TMEM180	1	TMEM8	1	ELOVL7	0.99785044
RBMS3	1	TMOD2	1	NFIA	0.997946546
SFRP1	1	GUCY1B3	1	DMRT1	0.998570232
KIAA1199	1	CYP3A7	1	MCF2	0.998997692
RRAGD	1	TLE4	1	TLR4	0.999086184
STK31	1	FNDC4	1	LAMA1	0.999179277
COL6A3	1	STK31	1	STRA8	0.999252468
CTNNAL1	1	MME	1	AVPR1A	0.99946217
MTMR7	1	ATF7IP2	1	PECAM1	0.999508217
GNG3	1	PRKCB1	1	FAM40B	0.999520464
SMTNL2	1	PECAM1	1	SOHLH2	0.999604326
KIT	1	TACSTD2	1	LOC441376	0.999605569
TDRD12	1	TEK	1	KCNJ3	0.999764073
GLDC	1	C1S	1	STK31	0.999800011
CDH1	1	EHF	1	UBA52	1
CCDC3	1	NR6A1	1	SFRP2	1
TDGF1	1	NR0B1	1	EGFL6	1

In the third step, we start with the highly correlated gene pairs and apply the ODE model to infer the gene regulatory network. By solving the optimization problem (5) for each gene, we can obtain a sparse matrix Ji. We use the sophisticated algorithm for LASSO problem and choose the parameter λ=100. From the matrix J_i, we can decide which genes activate or repress gene i. Iterating for each gene, we can infer the whole gene regulatory network.

Figure 5 shows the global picture for the inferred gene regulatory network underlying cell reprogramming. Although we utilize a three stage procedure to make the inferred network sparse. Still we obtain a network with 1,629 genes and 13,295 regulatory interactions. Averagely every gene has 16.25 neighbors. The maximal out-degree is 64 and the maximal in-degree is 320. This huge network provides many candidate interactions for further experimental validations.

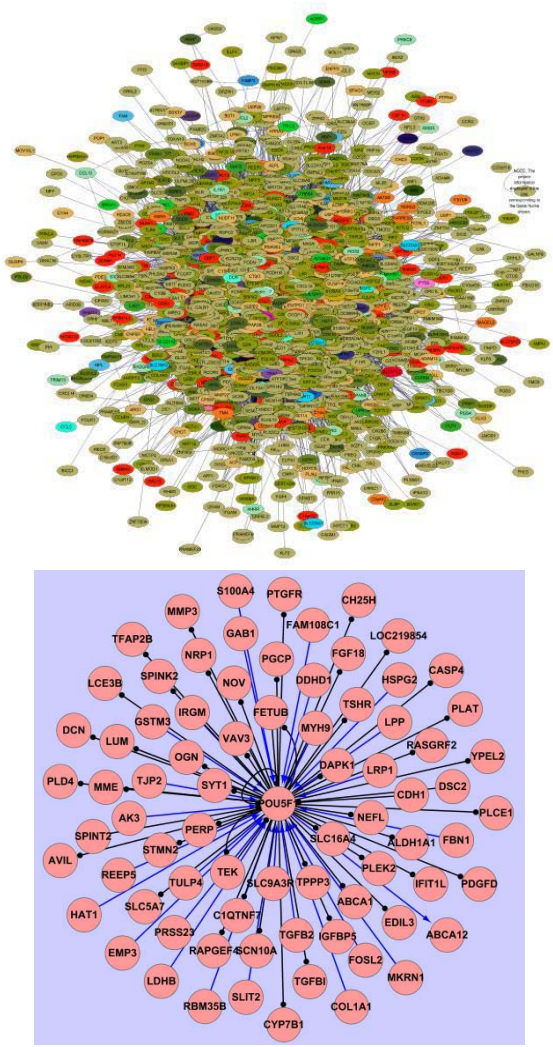


Figure 5.Results of the inferred gene regulatory network (activation: blue arrow, repression: black dot) for cell reprogramming. The upper subfigures shows the whole network. In total there are 1,625 nodes and 13,295 regulatory interactions. The bottom subfigure shows the subnetwork of Pou5f1 with 77 genes and 79 edges.

Then we extract the subnetworks related to gene Pou5f1 (also known as Oct4) from the whole network and illustrate them in Figure 5. There are 77 nodes connected by 79 edges in Pou5f1 subnetwork. We can see that Pou5f1 serves as a network hub to activate and repress many genes to perform biological functions. It's well known that Pou5f1 is a pluripotent marker gene. It's clear in the inferred network that Pou5f1 represses Tfgb1 and Tgfbp5, which are main factors to promote differentiation and development. This fact coincides with the existing studies. We also observed that Cdh1 (previously known as E-cadherin) represses Pou5f1 during the reprogramming. Cdh1 marks adherence junctions and upregulated in the initial phase of reprogramming [15]. Its relationship with Oct4 needs close inspection.

Finally, we perform functional analysis on the inferred regulatory network. The global picture for the main function of the network is shown in Figure 6. It is clear that biological adhesion, multicellular organism processes, and immune systems response are three dominant function terms. Given the facts that cell reprogramming is closely related to cell-cell interaction and adhesion [15], our network shows reasonable results. Also multi-cellular process is related to

differentiation and development, which should be repressed during cell reprogramming.

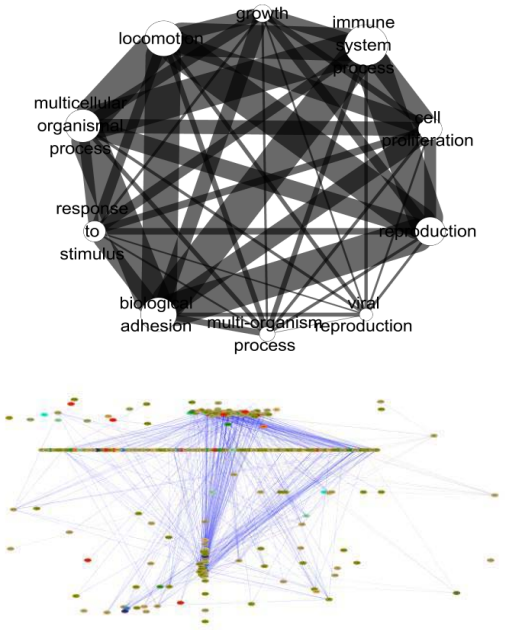


Figure 6. The overview of the GO annotations for the inferred gene regulatory network. The upper subfigure shows the GO slim terms and their interactions. The size of node is proportional to the number genes annotated by this term in the network. The size of the edges is proportional to the number of the overlapped genes of two terms. The bottom subfigure shows the subnetwork of biological adhesion term. The figures are generated by Mosaic plugin in cytoscape http://nrnb.org/tools/mosaic/.

4 Discussions and conclusions

In this paper we propose a new network inference method to reconstruct the gene regulatory network underlying cell reprogramming. As far as we know, this is the first regulatory network inferred from the time series data to capture the cell reprogramming phase. In addition to the well-known dimensionality problem for network inference, the task we treated here is even more challenging, because the cell reprogramming is highly dynamical, a lot of genes are involved, and only very short time course data are measured. We utilize a three stage procedure to respectively select the most relevant genes in the biological process, filter the large amount of lowly correlated gene pairs, and further pick out the regulator by quantitatively modeling regulatory relationships. For conceptual comparison, we also design two control experiments. In the first control experiment, we just follow the Stage I and Stage II. In the second control experiment, we just do the Stage I and Stage III. Compared with the control experiment 1, our current strategy makes network more sparser and further we can infer the regulatory roles as activation or repression. Compared with the control experiment 2, our method pre-selected highly correlated genes and greatly reduces the computational cost and makes the final network sparser.

With all these efforts, still we infer a large scale network. We believe the network brings important insights and a lot of candidates for follow-up biological experiments. As the next step, we are trying to further narrow down the inferred network by utilizing other information in the publically available data and differentiating the direct and indirect regulatory interactions.

Acknowledgements

We thank Dr. Go Nagamatsu from Keio University for the time series data during cell reprogramming in [15].

References

[1] Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 126, 663-676.

[2] Amabile, G., and Meissner, A. (2009). Induced pluripotent stem cells: current progress and potential for regenerative medicine. Trends in molecular medicine 15, 59-68

[3] Li, M., Chen, M., Han, W., and Fu, X. (2010). How far are induced pluripotent stem cells from the clinic? Ageing Research Reviews 9, 257-264.

[4] Zhou, Q., Chipperfield, H., Melton, D.A., and Wong, W.H. (2007). A gene regulatory network in mouse embryonic stem cells.Proceedings of the National Academy of Sciences USA, 104: 16438-16443.

[5] Chang R, Shoemaker R, Wang W (2011) Systematic Search for Recipes to Generate Induced Pluripotent Stem Cells. PLoS Comput Biol 7(12): e1002300.

[6] Saito, S. et al., Potential linkages between the inner and outer cellular states of human induced pluripotent stem cells, BMC Systems Biology, Suppl.4, 2011.

[7] H. de Jong. Modeling and simulation of genetic regulatory systems: a literature review. J. Comput. Biol., 2002, 9,67-103.

[8] T. S. Gardner, J. J. Faith. Reverse-engineering transcription control networks. Physics of Life Reviews, 2005, 2:65-88.

[9] L. Chen, R. S. Wang, X.S. Zhang. Biomolecular Networks: Methods and Applications in Systems Biology. John Wiley & Sons, Hoboken, New Jersey. July, 2009.

[10] S. Hempel, A. Koseska, J. Kurths,1,3 and Z. Nikoloski, Inner Composition Alignment for Inferring Directed Networks from Short Time Series. PhysicalL Review Letters, 2011. 107(054101).

[11] Wang, Y., et al., Inferring gene regulatory networks from multiple microarray datasets. Bioinformatics, 2006. 22(19): p. 2413-2420.

[12] Y. Wang, T. Joshi, D. Xu, X.S. Zhang, L. Chen. Supervised inference of gene regulatory networks by linear programming. Lecture Notes in Bioinformatics, 2006, 4115, 551--561.

[13] Y. Wang, X. S. Zhang and L. Chen. A network biology study on circadian rhythm by integrating various omics data. OMICS: A Journal of Integrative Biology, 2009, 13(4).

[14] Y. Wang, X.-S. Zhang, and Y. Xia. Predicting eukaryotic transcriptional cooperativity by Bayesian network integration of genome-wide data. Nucleic Acids Research, 2009, 37(18):5943--5958.

[15] Payman Samavarchi-Tehrani, Azadeh Golipour, Laurent David, Hoon-ki Sung, Tobias A. Beyer, Alessandro Datti, Knut Woltjen, Andras Nagy, Jeffrey L. Wrana, Functional Genomics Reveals a BMP-Driven Mesenchymal-to-Epithelial Transition in the Initiation of Somatic Cell Reprogramming, Cell Stem Cell, Vol. 7, No. 1, 2010, Pages 64-77