Network-based comparison of temporal gene expression patterns

Wei Huang1,2,†, Xiaoyi Cao3,† and Sheng Zhong1,3,∗

1Department of Bioengineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA, 2Key Laboratory for Applied Statistics of Ministry of Education, Northeast Normal University, Changchun, Jilin, China and 3Center for Biophysics and Computational Biology, University of Illinois at Urbana-Champaign, Urbana, IL, USA

ABSTRACT

Motivation: In the pursuits of mechanistic understanding of cell differentiation, it is often necessary to compare multiple differentiation processes triggered by different external stimuli and internal perturbations. Available methods for comparing temporal gene expression patterns are limited to a gene-by-gene approach, which ignores co-expression information and thus is sensitive to measurement noise.

Methods: We present a method for co-expression network based comparison of temporal expression patterns (NACEP). NACEP compares the temporal patterns of a gene between two experimental conditions, taking into consideration all of the possible co-expression modules that this gene may participate in. The NACEP program is available at http://biocomp.bioen.uiuc.edu/nacep.

Results: We applied NACEP to analyze retinoid acid (RA)-induced differentiation of embryonic stem (ES) cells. The analysis suggests that RA may facilitate neural differentiation by inducing the shh and insulin receptor pathways. NACEP was also applied to compare the temporal responses of seven RNAi inhibition (RNAi) experiments. As proof of concept, we demonstrate that the difference in the temporal responses to RNAi treatments can be used to derive interaction relationships of transcription factors (TFs), and therefore infer regulatory modules within a transcription network. In particular, the analysis suggested a novel regulatory relationship between two pluripotency regulators, Esrrb and Tbx3, which was supported by in vivo binding of Esrrb to the promoter of Tbx3.

Availability: The NACEP program and the supplementary documents are available at http://biocomp.bioen.uiuc.edu/nacep.

Contact: szhong@illinois.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on July 7, 2010; revised on September 14, 2010; accepted on September 27, 2010

1 INTRODUCTION

Embryonic stem (ES) cells are capable of differentiating into all cell types in an adult body, and can be triggered by different external and internal signals (Ivanova et al., 2006). One of the major themes in ES cell research is to find efficient ways of guiding ES cells to differentiate into a desired cell type. We chose to approach this theme by comparing the differentiation processes triggered by different external stimuli or internal perturbations.

Differentiating mouse ES cells by withdrawing certain growth factors (usually leukemia inhibitory factor (LIF)) leads to generating a mixture of all cell types (spontaneous differentiation) (Hong et al., 2009). Alternatively, exposing ES cells to certain growth factors can lead to enrichment of certain cell lineages during differentiation. Examples include retinoic acid (RA)-induced differentiation that enriches neuronal cells (Jones-Villeneuve et al., 1983), and activin-induced differentiation that enriches mesendoderm cells (Sulzbacher et al., 2009). Besides growth factors, repressing individual regulatory proteins can also induce differentiation. Although we previously showed that the repression of a chromatin modeling factor may encourage ES cells to differentiate and express neural markers (Hong et al., 2009), in general, it is not clear whether the repression of ES cell regulators may encourage differentiation toward any specific cell lineages (Lu et al., 2009).

We hypothesized that if two ES cell regulators are in the same regulatory pathway, the temporal transcriptional responses to repressing them should be similar, as opposed to the temporal responses of repressing a third ES regulator that is not in this regulatory pathway or acting to inhibit this pathway. We also hypothesized that if a transcription factor (TF) inhibits a signaling pathway, the temporal responses to repressing this TF may be closer to the transcriptional responses to adding a growth factor that induces this signaling pathway as compared with the temporal responses to repressing another TF. With these hypotheses in mind, we applied the new method called NACEP to compare spontaneous differentiation against eight types of induced differentiation, including one external stimulation (RA induced) and seven internal perturbations (repressing ES cell regulators). The goals of these comparisons are as follows. First, we wish to gain mechanistic insights into how RA treatment leads to differentiation toward the neural lineage. Second, we wish to test whether any of these induced differentiation processes resemble one another, and thus to infer the relative proximities of these regulators in an ES cell regulatory network. These questions inspired us to revisit analytical methods for comparing time-course gene expression data.

To date, there are very few methods for comparing time-course gene expression data. One method suppresses the temporal information and compares the ‘neighborhood’ genes between two conditions (Cheng et al., 2006). Other methods explicitly model the temporal information, but treat every gene independently (Storey et al., 2005; Telesca et al., 2009). For example, one of these methods fits a spline to the time-course data of a gene in each of two experimental conditions, and then it compares the fitted splines.
This statistical approach enjoys at least two clear advantages, in that it takes full advantage of the time-course data structure, and it implements human intuition in comparing temporal patterns. The limitation is that every gene is modeled independently, and information such as co-expression is ignored. Since the two splines of a gene have to be fitted with typically a dozen data points or even fewer, the fitted splines are sensitive to biological and technical fluctuations. Compared with traditional two-sample comparison procedures that ignore the time information, the spline method can be even more prone to false positives because a random fluctuation on a data point can have a larger chance of inducing a detectable difference to the splines.

Intuitively, utilizing co-expression information, for example, by using co-clustered genes to stabilize the spline fit of a gene, might largely enhance the fit and thus enable much more accurate identification of temporal differences. We wish to formalize this intuition as a co-expression network-based comparison of temporal expression patterns (NACEP).

However, clustering time-course gene expression data by itself is a challenging problem. One group used predetermined gene expression patterns as cluster centers to cluster genes (Schulz et al., 2009). Other teams approached this problem by using a finite mixture model for the clusters and then implementing a spline fit within each cluster (Luan and Li, 2003; Ma et al., 2006). These methods are prominent advances in co-expression analysis of time-course data, but they appear to be far from what is necessary for a co-expression network-based comparison of temporal expression patterns. The outstanding challenges include but are not limited to the following. First, the cluster number has to be preset for these clustering algorithms. Although model-selection criteria such as akaike information criterion (AIC) or Bayesian information criterion (BIC) can in theory be used to judge cluster numbers, in practice the AIC or BIC curves usually do not show clear peaks or charges that are needed to make a decision (Qin, 2006). Second, clustering results are often unstable, in the sense that slight changes to the clustering algorithm or the dataset may generate vastly different clustering results. It is difficult to trust or interpret results that are sensitive to the analytical methods or noise in the data (Quackenbush, 2003). Third, a gene may not have only one function or it may not only participate in one module in the gene regulatory networks (GRNs). Forcing a gene into one cluster makes a strong assumption about the underlying GRNs, and making inferences using such an assumption may defeat the purpose of using network information to improve temporal comparisons. In our opinion, these outstanding difficulties prohibited the invention of a statistical method that explicitly utilizes network information in the identification of genes with different temporal expression patterns. The NACEP model in this article attempts to address these challenges.

2 METHODS

2.1 The NACEP model

NACEP explicitly uses co-expression network information to compare temporal gene expression data. To overcome the difficulties discussed above, NACEP first implements an infinite-mixture model for clustering time-course data. The number of clusters is automatically decided by the data and a Dirichlet process (Antoniak, 1974; MacEachern and Müller, 1998; Neal, 2000; Qin, 2006). Instead of forcing every gene into a cluster, NACEP passes the probabilities of every gene belonging to every cluster into the next step of analysis. In the second step, NACEP infers the temporal pattern of a gene as a weighted average of the temporal patterns of all the clusters, using the probability of assigning this gene to each cluster as the weight of that cluster. Finally, NACEP compares the temporal patterns of a gene between two experimental conditions with a non-parametric test, correcting for multiple hypothesis testing (Fig. 1).

2.2 An infinite-mixture model for clustering time-course data

NACEP implements an infinite-mixture model for clustering time-course data. The cluster memberships are treated as missing data and are assumed to be generated from a Chinese Restaurant Process (CRP; Qin, 2006). Let \( C = (C(1), \ldots, C(N)) \) be the cluster indicator variable, where \( C(i) \in \{1, \ldots, C\} \) denotes that the \( i \)-th gene is assigned to the \( c \)-th cluster, \( 1 \leq i \leq N \). We use \( |C| \) to denote the number of clusters present. \( |C| \) is unknown. \( C \) is treated as missing data in the model.

Given the missing data, the expression levels of a gene are modeled with a mixed-effects model (Luan and Li, 2003). In this mixed-effects model, the cluster mean is modeled as a B-spline. The measured expression level of a gene in this cluster at a time point is modeled as the sum of the cluster mean, a random gene effect and a noise term representing the overall effect of the biological and the technical fluctuations. Let \( Y_{ijkl} \) be the measured expression level for gene \( i \), under experimental condition \( j \), at time-point \( t_k \), from biological or technical replicate \( l \), where \( i = 1, \ldots, N; j = 1, \ldots, J; k = 1, \ldots, K; l = 1, \ldots, L \). Following Luan and Li (2003), the expression levels of the \( c \)-th cluster are modeled as:

\[
Y_{ijkl} = f_{c}(t_k) + \beta_k + \epsilon_{ijkl}
\]

(1)

where \( f_{c}(t_k) \) is the mean profile of the \( c \)-th cluster in the \( j \)-th experimental condition and \( \epsilon_{ijkl} \sim N(0, \sigma^2) \) is the gene effect, which is independent from the measurement error \( \epsilon_{ijkl} \sim N(0, \sigma^2) \). The smooth function \( f_{c}(t_k) \) is modeled as a B-spline, with its basis denoted as \( X \), and

\[
f_{c}(t_k) = X \beta_c
\]

(2)

where \( \beta_c \) is the parameter set of the B-spline.

Thus, a generative probabilistic model for all the time-course gene expression data has been completely specified, with a CRP for generating the cluster indicators and a mixed-effects model for generating expression levels under given cluster indicators.

2.3 The Bayesian formulation and a Gibbs sampling algorithm for model inference

To fit the model parameters from data, we rewrote the NACEP model into a Bayesian form and then developed a Gibbs sampling algorithm to estimate
the model parameters. To put NACEP into a Bayesian form, we used the theoretical developments of Dirichlet processes (Ferguson, 1973). Generally, if \((\Theta, B)\) is a measurable space, on which \(G_0\) is a probability measure and \(\omega\) is a positive real number, a stochastic process \(G\) is a Dirichlet process with base distribution \(G_0\) and concentration parameter \(\omega\) if and only if for any finite partitions \((A_1, A_2, \ldots, A_n)\) on \(\Theta\), \(G(A_1), G(A_2), \ldots, G(A_i) \sim \text{DP}(\omega G_0(A_i))\), where \(G_0(A_i)\) is a special form of Dirichlet process, and a CRP is often used as prior distribution for a Dirichlet process. Neal formulated a model for generating data from a Dirichlet process (Neal, 2000). In Neal’s formulation, the data distribution is a mixture of distributions of form \(F(\theta)\), with the mixing distribution over \(\theta\) being \(G\). Thus, 

\[
y_i | \theta_i \sim F(\theta_i)
\]

\[
\theta_i \sim G(\cdot | G)
\]

\[
G \sim \text{DP}(G_0, \omega)
\]

where \(y_i, i = 1, \ldots, N\) are the data points and \(G\) is a Dirichlet process prior, with concentration parameter \(\omega\) and base distribution \(G_0\).

Inspired by Neal’s formulation, we rewrite the NACEP model as:

\[
Y_i | (\beta, \phi) \sim N(X_i \beta, \phi^2 I)
\]

\[
\beta \sim N(0, \omega^2 I)
\]

\[
\phi^2 \sim \text{InvGamma}(\alpha_1, \alpha_0)
\]

where \(I\) is a column vector of \(1\)s: \((1, \ldots, 1)^T\). We use conjugate priors for \(\alpha\), \(\beta\), and \(\phi\) where \(\alpha\), \(\beta\), and \(\phi\) are hyperparameters. Thus, we provided a Bayesian formulation for the NACEP model. Based on this formulation, we developed a Gibbs sampling algorithm to make our model inference (Bush and MacEachern, 1996) (Fig. 2; Supplementary Material).

### 2.4 Comparison between experimental conditions

How different are the temporal patterns of a gene in two conditions? NACEP computes the false discovery rate (FDR) for every gene that takes multiple hypothesis testing into consideration. NACEP ranks the genes with their FDR.

\[
b_i = \frac{\text{deg}(C_i) \cdot \hat{p}_i}{\text{deg}(C_i) + \hat{p}_i}
\]

where \(C_i\) is a configuration of gene clustering and \(\hat{p}_i\) is the estimated probability for genes \(i\) and \(i'\) being in the same cluster, which can be obtained by counting the fraction of times \(c(i) = c(i')\) in the iterations; in any given clustering configuration \(C_0\), \(\delta_{C_i}(C_0) = 1\), if genes \(i\) and \(i'\) are in the same cluster, otherwise \(\delta_{C_i}(C_0) = 0\).

It should be noted that in the comparison of temporal patterns, NACEP does not force a gene to belong to one cluster; instead all the posterior probabilities of a gene belonging to all clusters are used. Thus, the temporal comparison results are not sensitive to the clustering performance.

### 2.5 Assessing statistical significance

To assess the statistical significance of the difference of temporal patterns of a gene, NACEP computes the distribution of \(d_{ij}\) under the null hypothesis by permuting the expression data of matched time points under the two conditions. Following Storey et al. (2005), NACEP uses the permutation to compute the false discovery rate (FDR) for every gene that takes multiple hypothesis testing into consideration. NACEP ranks the genes with their FDR.

### 2.6 Clustering time-course gene expression data

Although NACEP was designed for utilizing co-expression information to enhance the comparison of temporal gene expression patterns, as a byproduct, NACEP provides a handy and potentially powerful tool for clustering time-course gene expression data. The major improvements of NACEP from other time-course data clustering approaches (Luan and Li, 2003; Ma et al., 2006) are 2-fold. First, NACEP employs an infinite-mixture model and thus the cluster number is judged automatically by the Gibbs sampler algorithm. Second, NACEP enables clustering data from more than one experimental condition, by simultaneously fitting temporal profiles within every experimental condition.

To use NACEP as a clustering algorithm, after running the Gibbs sampler, the clustering inference can be made by:

\[
\hat{C} = \arg \min_{C} \sum_{i,j} \left(\lambda_{ij}(C) \cdot d_{ij}\right)^2
\]

where \(C\) is a configuration of gene clustering and \(\lambda\) is the set of all such configurations; \(d_{ij}\) is the estimated probability for genes \(i\) and \(i'\) being in the same cluster, which can be obtained by counting the fraction of times \(c(i) = c(i')\) in the iterations; in any given clustering configuration \(C_0\), \(\delta_{C_i}(C_0) = 1\), if genes \(i\) and \(i'\) are in the same cluster, otherwise \(\delta_{C_i}(C_0) = 0\).

The correlation between two TFs is the Pearson correlation between the two NACEP distance vectors of the two RNAi experiments.
### 3 RESULTS

#### 3.1 Analysis of synthetic data

NACEP was evaluated on synthetic datasets for clustering performance and for comparison of temporal patterns. Although the results of the two tests are presented sequentially below, it is worth noting that NACEP’s temporal comparison does not rely on a pre-fixed clustering result (see Discussion).

#### 3.2 Clustering

We compared NACEP with three other clustering methods: K-means, MClust and smoothing spline clustering (SSC) (Ma et al., 2006). Both MClust and SSC use a finite-mixture model for clustering. While MClust assumes that the samples are independent, SSC uses splines to model the time-course data structure. As a clustering method, NACEP can be regarded as an extension of SSC to an infinite-mixture model.

We simulated 100 datasets. Each simulation was composed of four clusters, containing 30, 40, 50 and 30 genes in each cluster. Every gene was measured at 10 time points. Following Ma et al. (2006), the mean expression patterns were simulated with the following functions:

\[
y_{ik} = -\exp(t_k) + b_i + \epsilon_{ik}; \quad i = 1, 2, \ldots, 30;
\]

\[
y_{2ik} = \tan(t_k) + b_i + \epsilon_{2ik}; \quad i = 1, 2, \ldots, 40;
\]

\[
y_{3ik} = s(t_k - 4)^2 / 36 + b_i + \epsilon_{3ik}; \quad i = 1, 2, \ldots, 50;
\]

\[
y_{4ik} = \cos(\theta) + b_i + \epsilon_{4ik}; \quad i = 1, 2, \ldots, 30;
\]

where \(k\) is the index of time points; \(i\) is the index of genes; \(b_i\) is the random gene effect; \(\epsilon_{ik}\) is the random measurement error.

Two comparisons were made. First, we compared how often a method incorrectly identifies the cluster number by ‘cluster number prediction error’ (CNPE; Fig. 3B). The BIC was used for MClust and SSC to choose the cluster number. Since K-means cannot automatically determine the cluster number (unless assuming some parametric form of data distribution), we assigned the correct cluster number to K-means and exempted it from the first comparison. Using BIC with SSC (18% CNPE) improved the chances of correctly identifying cluster numbers compared with using BIC with MClust (43% CNPE).

Moreover, the results of NACEP (0% CNPE) were further improved from the results of using BIC with SSC (18% CNPE).

Second, when the cluster number was correct, we compared how often a gene is incorrectly clustered [average misclassification rate (AMR), Fig. 3B]. In this comparison, we assigned the correct cluster number to K-means, MClust and SSC. By doing so, we gave the other algorithms an advantage over NACEP. In this comparison, SSC and NACEP largely outperformed K-means and MClust, consistent with the expectation that explicitly modeling the time-course data structure might boost clustering performance. NACEP further exhibited a 22-fold (1.54/0.0733) improvement of clustering accuracy compared with SSC.

#### 3.3 Comparison of temporal patterns

We compared NACEP with a single-gene-based time-course comparison method called EDGE (Storey et al., 2005).

We did four simulations. In each simulation, time-course gene expression data from two experimental conditions were generated, with 10 time points in each condition. In the first simulation, a total of 1000 genes were simulated. Among these genes, 600 genes were simulated to have the same temporal pattern in the two conditions (Group f, Fig. 4A). The remaining 400 genes had different expression patterns. They were separated into four groups (Groups a–d, Fig. 4A). Their expression patterns differ between the two conditions as follows. Groups a and b had different trends. Group c was generated from gamma functions with different parameters. Group d was generated from sine functions with different phases.

The gene groups were assumed to have different variances on their gene effects (\(b_i\)) and the same variance on the measurement errors (\(\text{Var}(\epsilon_{ik}) = 1\)). We made \(\text{Var}(b_i)\) increase from Group a to Group d, as reflected by the increasing confidence intervals of the mean trajectories in Figure 4A.

To simulate more realistic data, in the second simulation, we increased the variances of the gene effect and the measurement error. To simulate the situations in which some genes cannot be easily clustered (scatter genes), we added a new gene group to the third simulation. This new group (Group g, Fig. 4A) contained 100 genes, each independently generated to have its own temporal pattern, which is the same in the two conditions. Group g genes differ in the two conditions by the average expression level (gene effect \(b_i\)) and measurement variation (\(\epsilon_{ik}\)). In order to challenge NACEP even further, in the fourth simulation, we added yet another group of scatter genes (Group e) with different temporal patterns in the two conditions.

Each simulation was independently repeated 50 times. Both NACEP and EDGE were applied to these datasets to detect genes with different trajectories. True and false positives of these predictions are summarized as receiver operating characteristic (ROC) curves in Figure 4B. NACEP out-performed EDGE in all four simulations. Notably, when the false positive rates are small (using high thresholds), EDGE performed not much better than a random decision, by producing a small number of true positives (Fig. 4B). NACEP largely increased the sensitivities under the same specificities of EDGE, causing the ROC curves to shoot up almost vertically to 20% at a 0% false positive rate in all simulations. This indicates a particularly useful feature of NACEP, in that its top predictions are likely to be reliable.
Fig. 4. Cross-condition comparison. (A) Four synthetic datasets. Five to seven groups of genes were simulated in each synthetic dataset, with each group exhibiting either a different pattern (Groups a–e), or a similar pattern (Groups f and g) between the two experimental conditions. Genes in Groups e and g do not form any clusters. Each gene was generated from its own temporal pattern. Group d differs in the two conditions by a phase shift. The cluster averages and their confidence intervals are shown in solid curves and shaded regions, respectively. Expression values of 10 representative genes are shown for each gene group in dots. The standard deviation of the expression values of all genes in a group is shown as a green band. (B) ROC curves. Panels 1–4 correspond to synthetic datasets 1–4.

3.4 Analysis of ES cell differentiation

3.4.1 Differential temporal responses between spontaneous and RA-induced differentiation of ES cells

We applied NACEP to identify the genes and pathways that mediate RA-induced differentiation of ES cells. RA is known to facilitate ES cell differentiation and to enrich neurogenic precursor cells among the differentiated cells, although the molecular mechanisms of such an effect remain elusive (Glaser and Brüstle, 2005). We hypothesized that the neurogenic effect of RA is mediated by a set of neurogenic regulatory genes, whose expression patterns are different between RA-induced and spontaneous differentiation. To test this hypothesis, we reanalyzed the data by Ivanova et al. (2006), who subjected mouse ES cells to spontaneous and RA-induced differentiation processes (Ivanova et al., 2006). Gene expression was measured in each differentiation condition every day over a 7-day period.

To preprocess the data, we used Gene Ontology (GO) annotations to obtain the genes involved in transcriptional regulation or signal transduction and filtered out the other genes. We also filtered out the genes with small changes of expression levels in both of the two time series and subjected the remaining 783 genes to NACEP analysis. NACEP’s Gibbs sampling computation stabilized after 10 000 iterations and generated 37 clusters (Fig. 5A). Canonical ES cell regulators including Nanog, Oct4, Klf2, Esrrb and Utf1 all showed up in one cluster, lending credibility to the clustering result. Two Ets domain TFs, Etv4 and Etv5, were clustered together with the canonical ES cell regulators, suggesting that these TFs involved in organ morphogenesis might have a neglected role in ES cell regulation. Interestingly, Etv5’s DNA binding motif was reported to be enriched in Nanog bound regions (Zhou et al., 2007).

Another cluster contained Shh and Gli1, the key ligand and TF of the shh signaling pathway, suggesting the shh pathway might be tightly regulated in both spontaneous and RA-induced differentiation of mouse ES cells.

Comparing the two differentiation processes, NACEP reported 156 genes with different temporal patterns (FDR < 10^{-5}). We then separated these 156 genes into two groups, i.e. the induced (134) and the repressed (22) genes, by RA as compared with spontaneous differentiation. The 22 genes repressed by RA included pluripotency and self-renewal regulators Esrrb, Utf1, Nanog, Klf2 and Oct4. These data are consistent with the notion that RA facilitates differentiation.

The top-ranked RA-induced genes included Gli3, Zic3 and others. GLI3 is one of the three Gli family proteins in mice, which serve as key TFs of the shh pathway. Consistent with this result, Zic3 is a known downstream transcriptional target of Gli family TFs (Mizugishi et al., 2001). These data, together with the result that Shh and Gli were clustered together (Fig. 5A), suggest that the shh pathway genes might be activated by RA and mediate the differentiation of ES cells into neural precursors. Consistent with this hypothesis, the induction of the shh pathway promotes neuronal differentiation from embryoid bodies that are differentiated from ES
cells (Vokes et al., 2007). Moreover, in mouse brains, shh activation has been associated with neuroplastic growth and development of brain tumors (Flora et al., 2009). Finally, Etv4, which inhibits Shh during limb bud development (Zhang et al., 2009), was identified by NACEP as one of the top-repressed genes by RA. These data are in-line with the hypothesis that the shh pathway is positively regulated during RA-induced differentiation.

The top RA-induced genes also included Nr2f2 and Igf1. Igf1 can serve as a ligand to the insulin receptor (INSR) pathway. Nr2f2 was shown to be recruited to the Igf1 promoter region, modulating its expression (Kim et al., 2009). These data tempted us to hypothesize that RA-induced neuronal differentiation is at least partially mediated by the INSR pathway. Coincidentally, INSR induces neuronal differentiation of neuroendocrine tumors (Dikic et al., 1994), and it is vital for keeping neural stem cells alive (Siegrist et al., 2010). Taken together, these analyses suggest that RA may induce neuronal differentiation through activating shh and INSR pathways (Fig. 5B).

3.4.2 Comparison of temporal expression responses to RNAi reveals placement of TFs in a GRN. A number of ES cell TFs that sustain self-renewal and inhibit differentiation have been identified. Less clear are the interaction relationships of these ES cell TFs, or the full picture of the ES cell GRN. In order to elucidate the GRN of ES cells, biochemical assays including sequential chromatin immunoprecipitation (ChIP) were used to directly assess the interaction of two TFs (Geisberg and Struhl, 2004); co-localization of binding regions (from ChIP-seq or ChIP-chip data) was used to infer TF interactions (Chen et al., 2008), and co-expression information together with protein-protein interaction data were used to infer GRNs (Müller et al., 2008).

We hypothesized that the transcriptomic responses to knockdown of TFs should also contain useful information on the relative placement of these TFs in a GRN. To test this hypothesis, we started by considering a hypothetical situation, wherein there are three TFs (A, B and C), with A and B ‘closer’ in the GRN as compared with C (Fig. 6A). The relative proximity of A and B can be substantiated in the following examples: A and B often interact with each other, forming a dimer to co-bind and co-regulate target genes; C can only interact with A or B through the assistance of other proteins; C is an upstream regulator of A and B; or C independently regulates a set of its own downstream genes. In these hypothetical examples, the genome-wide temporal transcriptional responses to the knockdown of A and B should be similar, as opposed to the temporal responses to C knockdown.

Using this idea, we applied NACEP to a set of time-course gene expression data generated by seven RNAi experiments (Ivanova et al., 2006), namely the gene expression data of Days 0–7 after the knockdown of Nanog, Oct4, Sox2, Esrrb, Tbx3, Tcf1 and Ccnb1ip1. We compared the temporal responses with each RNAi experiment to the control data, i.e. the gene expression data of wild-type ES cells on matched time (Days 0–7). In every comparison (by NACEP), the genes with a different expression pattern between the RNAi and the control conditions were identified (Fig. 6B). To quantify the similarity of the genes affected by two RNAi experiments, we used the Pearson correlation of the NACEP distances of the two RNAi experiments (see Methods). The Pearson correlation was regarded as a similarity/proximity metric between the two TFs on which the RNAi were performed (Fig. 6C). To obtain a global view of the proximity of all TFs in the GRN, we clustered the TFs by the Pearson’s correlations. For visualization purposes, an edge was drawn between two TFs if their correlation was beyond an ad hoc threshold of 0.54 and if the width of the edge was proportional to the correlation between the two connecting TFs (Fig. 6D). This result indicates that Nanog, Oct4, Sox2 and Esrrb may form a heavily connected regulatory module, and Tbx3, Tcf1 and Ccnb1ip1 are attached to the Nanog-containing module through a few specific links. To check whether the predicted GRN structure was sensitive to the choice of Pearson’s correlation as the similarity/proximity metric or sensitive to the use of all genes, we applied another similarity metric (Supplementary Fig. S1A) and restricted the calculation to the top 5% of genes most strongly affected by all RNAi experiments (Supplementary Fig. S1B), and we found that the predicted TN structure was robust. This analysis based solely on RNAi transcriptomes identified the interactions among Nanog, Oct4, Sox2, Esrrb and Tbx3 consistent with evidence drawn from protein-protein interaction, protein-DNA binding data, and mutation analysis of TF binding sites (dashed edges, Fig. 6D) (Boyer et al., 2005; Chen et al., 2008; Han et al., 2010; Niwa et al., 2009; Pirity and Dinnyes, 2010; van den Berg et al., 2008; Zhang et al., 2008). This analysis failed to predict the transcriptional regulatory relationship between Oct4 and Tcf1 (Matoba et al., 2006), which might be explained by the hypothesis that Tcf1 only specifically regulates a much smaller subset of genes than Oct4, and thus Tcf1 RNAi only reflects a small subset of transcriptomic changes downstream of Oct4 RNAi. The strong correlation of the temporal responses to knockdowns of Esrrb and Tbx3 (Fig. 6D) predicts the proximity of these two TFs in the GRN. Further experiments are needed to test whether these TFs could directly interact with
April 12, 2010, Gene Expression Omnibus (GEO) has documented new methods are needed to exhibit expression profiles that are strikingly similar to results. Some genes known to be involved in a particular pathway analyses of expression data are limited by strong noise in the function.

NACEP can potentially be generalized to handle non-time-course (Luan and Li, 2003) mean generating the gene expression patterns. the clusters and a mixed-effects model (Wang, 1998) with a B-spline probabilistic model with a Dirichlet process (Neal, 2000) generating attempt toward these goals. The heart of this method is a generative intelligently determined. Second, the time-course nature of the data methodological challenges. First, the cluster number has to be intelligently determined. Second, the time-course nature of the data has to be explicitly utilized. The NACEP method represents an attempt toward these goals. The heart of this method is a generative probabilistic model with a Dirichlet process (Neal, 2000) generating the clusters and a mixed-effects model (Wang, 1998) with a B-spline (Luan and Li, 2003) mean generating the gene expression patterns. NACEP can potentially be generalized to handle non-time-course data by not requiring \( f(q_t) \) in Equation (1) to be a time-dependent function.

4.3 GRN structure and TF knockdown

Although gene expression responses to knockdown of TFs are often measured, such data were often used to identify the transcriptional targets of the inhibited TFs. To our knowledge, there is as yet no principled approach to infer the interaction relationships among the inhibited TFs, except in some special cases in which one TF is the transcriptional target of another. As a proof of principle, this work demonstrated that the temporal transcriptional responses to the knockdown of a set of TFs could be used to identify the interaction relationships of these TFs, as well as their relative proximity in the GRN. Our data suggest that Esrb may contribute to maintaining pluripotency through transcriptionally regulating Tbx3, a T-box transcriptional repressor. We were also interested in testing whether any of the TFs maintain pluripotency through, at least partially, inhibiting the RA pathway and its downstream genes. The differences between RA treatment and any TF knockdown are larger than the differences between any two TF knockdowns (Fig. 6C; Supplementary Fig. S1), suggesting that the seven TFs included in our analyses are likely involved in pathways independent to the RA pathway.

ACKNOWLEDGEMENTS

The authors thank Dan Xie, Des Guixuan Lin, Xuming He, Kimberly A. Hughes, Jianhua Guo and Ningzheng Shi for useful discussions.

Funding: NSF DEB 08-48386, NSF DBI 08-45823, NIH R01 DK082605, NIH supplement to GM058686-0851.

Conflict of Interest: none declared.

REFERENCES


Comparison of temporal gene expression patterns


