Bioinformatics, 31(15), 2015, 2537–2544
doi: 10.1093/bioinformatics/btv186
Advance Access Publication Date: 2 April 2015
Original Paper

OXFORD

Systems biology

Regulatory network inferred using expression data of small sample size: application and validation in erythroid system

Fan Zhu[1,†], Lihong Shi[2,†], James Douglas Engel[3] and Yuanfang Guan[1,4,5,*]

[1]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA, [2]State Key Laboratory of Experimental Hematology, Institute of Hematology and Blood Diseases Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Tianjin 300020, China, [3]Department of Cell and Developmental Biology, [4]Department of Internal Medicine, and [5]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109, USA

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Inanc Birol

Received on November 7, 2014; revised on March 23, 2015; accepted on March 27, 2015

Abstract

Motivation: Modeling regulatory networks using expression data observed in a differentiation process may help identify context-specific interactions. The outcome of the current algorithms highly depends on the quality and quantity of a single time-course dataset, and the performance may be compromised for datasets with a limited number of samples.

Results: In this work, we report a multi-layer graphical model that is capable of leveraging many publicly available time-course datasets, as well as a cell lineage-specific data with small sample size, to model regulatory networks specific to a differentiation process. First, a collection of network inference methods are used to predict the regulatory relationships in individual public datasets. Then, the inferred directional relationships are weighted and integrated together by evaluating against the cell lineage-specific dataset. To test the accuracy of this algorithm, we collected a time-course RNA-Seq dataset during human erythropoiesis to infer regulatory relationships specific to this differentiation process. The resulting erythroid-specific regulatory network reveals novel regulatory relationships activated in erythropoiesis, which were further validated by genome-wide TR4 binding studies using ChIP-seq. These erythropoiesis-specific regulatory relationships were not identifiable by single dataset-based methods or context-independent integrations. Analysis of the predicted targets reveals that they are all closely associated with hematopoietic lineage differentiation.

Availability and implementation: The predicted erythroid regulatory network is available at http://guanlab.ccmb.med.umich.edu/data/inferenceNetwork/.

Contact: gyuanfan@umich.edu

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

1 Introduction

In the past decades, significant research efforts have been devoted to infer gene regulatory networks (GRNs) in the bioinformatics field (Altieri, 2008; Garcia-Echeverria and Sellers, 2008; Gitter *et al.*, 2010; Hennessy *et al.*, 2005; Huang, 1999; Kanehisa and Goto, 2000; Shinozaki *et al.*, 2003; Vogelstein and Kinzler, 2004; Welch *et al.* 2014). The challenge lies in that inferring regulatory network is an ill-posed problem, because the number of interactions to be inferred exceeds the number of independent experiments. Today, the standard solution to these $p > n$ problems (p is the number of parameters and n is the number of training samples) is different regularization techniques, adopting or modifying the method Tikhonov developed almost 80 years ago. When $p>>n$, i.e. an extremely limited set of experimental observations are available, it is often challenging to find a stable solution. In this article, we show a method to utilize both small, context-specific expression data, and large, non-specific datasets to infer regulatory networks. The essence of this method is to intrinsically increase the p by giving a weight to non-specific datasets according to their relevance to the context-specific dataset and accuracy. This technique has been used in non-directional networks, but hasn't been applied to regulatory networks with time-course data.

Current methods (Ernst *et al.*, 2008; Faith *et al.*, 2007; Friedman *et al.*, 1998; Hecker *et al.*, 2009; Huttenhower *et al.*, 2009; Ma *et al.*, 2006; Marbach *et al.*, 2010, 2012; Margolin *et al.*, 2013; Michoel *et al.*, 2009; Mordelet and Vert, 2008; Park *et al.*, 2010; Pique-Regi *et al.*, 2011; Poultney *et al.*, 2012; Prill *et al.*, 2010, 2011; Yu *et al.*, 2004; Zhu and Guan 2014; Zhu *et al.*, 2014; Zou and Conzen, 2005) for modeling regulatory networks based on expression data can be divided into two categories. The first category utilizes a single time-course dataset over a development process or under certain environmental perturbations. Methods in this category focus on predicting regulatory networks corresponding to the same context of input data, which includes context likelihood of relatedness (CLR) (Faith *et al.*, 2007), dynamic Bayesian network (DBN) (Zou and Conzen, 2005), smoothing spline clustering (SSC) (Ma *et al.*, 2006) and Learning Module Networks (LeMoNe) (Michoel *et al.*, 2009) and semi-supervised regulatory network discoverer (SEREND) (Ernst *et al.*, 2008). Intensive research efforts have been devoted to these methods, with community-based assessments (e.g. the Dialogue for Reverse Engineering Assessments and Methods challenges) identifying several well-performed algorithms (Marbach *et al.*, 2010; Marbach *et al.*, 2012; Margolin *et al.*, 2013; Prill *et al.*, 2010, 2011), including ours. Numerous great methods have been derived from these crowd sourcing efforts (Flassig *et al.*, 2013; Irrthum *et al.*, 2010; Marbach *et al.*, 2012; Menéndez *et al.*, 2010; Steiert *et al.*, 2012; Yip *et al.*, 2010). This category of methods may capture the cell lineage-specific and other context-specific information. However, the accuracy of these algorithms is expected to be heavily affected by the input time-course data and subject to fluctuation due to experimental observation errors.

Methods in the second category generate regulatory networks by integrating multiple genomic datasets, such as supervised inference of regulatory networks (SIRENE) (Mordelet and Vert, 2008), multi-level integrated inference and analysis of regulatory networks (Poultney *et al.*, 2012), combinatorial algorithm for expression and sequence-based cluster extraction system (COALESCE) (Huttenhower *et al.*, 2009) and simultaneous genome-wide inference of multiple types of interactions including regulatory ones (Park *et al.*, 2010). By leveraging multiple datasets, these algorithms are expected to be more accurate in predicting a global, non-specific network (Hecker *et al.*, 2009). However, the context-specificity of such integrative methods is limited, due to the lack of context-specific input data.

The complementary nature of the two categories motivated us to develop DIFMINE, a novel multi-layer graphical model that leverages multiple non-specific time-course expression datasets, as well as a single context-specific, yet small-scale dataset to infer regulatory networks that are specific to a differentiation process. In this multi-layer model, the regulatory networks of individual non-specific genomic datasets are inferred using network inference algorithms developed for a single dataset. For this step, we tested state-of-the-art algorithms identified in previous community-based blind assessments (Prill *et al.*, 2010, 2011; Marbach *et al.*, 2012). Next, the relevance of these non-specific networks against the context-specific dataset is evaluated using Bayesian rules, and integrated into a single network. The resulting network intends to reveal the context-specific regulatory relationships on the whole genomic scale. In contrast to other state-of-the-art regulatory network prediction methods, our approach benefits from both the large number of publicly available functional genomic datasets and the small-scale cell lineage-specific dataset that defines the context.

We applied our algorithm to a cell lineage-specific expression RNA-seq dataset of human erythropoiesis. We demonstrated that we could significantly improve the performance of the network inference based on a single, limited dataset, compared with one of the most cited, the state-of-the-art single-dataset methods such as CLR (Faith *et al.*, 2007) and DBN. Additionally, we compared the performance of our algorithm against context-insensitive integration without the information from the cell lineage-specific dataset and found that the former successfully identifies regulatory relationships activated during erythropoiesis. Finally, we used a recently generated ChIP-seq data for TR4, an important transcription factor involved into erythropoiesis (Tanabe *et al.*, 2007; Cui *et al.*, 2011), and validated our prediction of regulatory relationships using this data.

2 Materials and methods

2.1 Inferring cell lineage-specific regulatory network with observed expression data with a small number of samples

The problem we address in this article is how to infer cell lineage-specific regulatory network with limited observed expression data. The differentiation of a cell lineage is a confined process. Typically, a short series of time-course expression data is sufficient to describe the transcriptomic profile of this process. Methods that are used to infer regulatory networks tend to be limited in accuracy when only a small number of sampling points are available. In the original paper that reported one of the most cited method in this field, CLR, 445 sampling points were used to reconstruct the regulatory network (Faith *et al.*, 2007). To address this challenge, we developed a multi-level graphical model to leverage both the cell lineage-specific expression data and the generic, publicly available expression data to infer regulatory networks.

We tested our strategy as human erythroid $CD34^+$ progenitor cells differentiated into mature erythroid cells *in vitro* (Goh *et al.*, 2004), in which the erythroid cells differentiated nearly synchronically. We collected RNA-seq datasets for the transcription profiling at day 4, 8, 11 and 14, respectively, after inducing human $CD34^+$ hematopoietic progenitor cells differentiation (Section 2). This time

frame covers the differentiation stages from immature progenitor cells until mature reticular erythroblasts.

Figure 1 is the workflow of DIFMINE, which contains two levels of Bayesian networks:

1. Generate erythroid-specific gold standard pairs. This cell lineage-specific dataset, together with known regulatory interactions, was used to generate the gold-standard regulatory relationships for evaluating the relevance of the non-specific datasets. First, we obtained activation regulatory interactions from Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway (Kanehisa and Goto, 2000). Then gene pairs that are significantly co-expressed in the erythropoiesis dataset are identified. The intersection of the two sets was used to establish the gold standard regulatory relationships involved in erythropoiesis, with defined directions.
2. Collect suitable public available datasets. We first acquired 1335 human microarray datasets from Gene Expression Omnibus (GEO) database (Edgar *et al.*, 2002) as of September 20th, 2013. These genome-wide datasets will allow us to derive a prediction score between almost every transcription and other genes. These datasets are then filtered according to the following criteria. (i) They represent samples of relatively small time intervals, i.e. no more than 48 h between time points. (ii) Each of them has more than or equal to eight points, so that time-lagged analyses can be carried out. There are 52 datasets satisfying these criteria (Supplementary Supporting Information S1) and thus included in the integration. Datasets are imputed using Sleipnir (Huttenhower *et al.*, 2008) KNNImputer. The pre-processing details of the public datasets are available in Supplementary Supporting Information S7.
3. Generate first layer DBN. For each dataset, we calculated a score for each gene pair that is indicative of the probability of one gene regulating the other. The following algorithms were tested for inferring the regulatory network based on a single dataset: (i) Time-lagged correlation (Schmitt *et al.*, 2004), in which correlations between time-shifted transcription profiles were calculated to represent the regulatory relationships. (ii) DBN, in which regulator-target gene pairs are usually identified based on a statistical analysis of their expression relationships across different time slices (Zou and Conzen, 2005), (iii) Lasso regularization, which is a popular method for regularization. (iv) Truncated Singular Value Decomposition (TSVD), which is another regularization method that solves the ill-posed problem (Holter *et al.*, 2001; Yeung *et al.*, 2002; Zhu *et al.*, 2012, 2013). Each of the above methods is capable of producing a score for each gene pair. Note that for all of these methods, score S is not symmetric due to the directionality of regulatory relationships, i.e. $S(i, j) \neq S(j, i)$. More details can be found in Supplementary Supporting Information S4.
4. Generate second layer DBN. For each method listed earlier, we integrated the scores from the 52 datasets together based on the cell lineage-specific gold standard pairs using Bayesian integration (Chikina *et al.*, 2009; Guan *et al.*, 2008; Guan *et al.*, 2012; Huttenhower *et al.*, 2009; Lee *et al.*, 2011; Pop *et al.*, 2010; Wong *et al.*, 2012). The integration workflow is similar with our previous works (Guan *et al.*, 2008, 2012), while the previous workflow is targeting gene co-functional networks (i.e. no

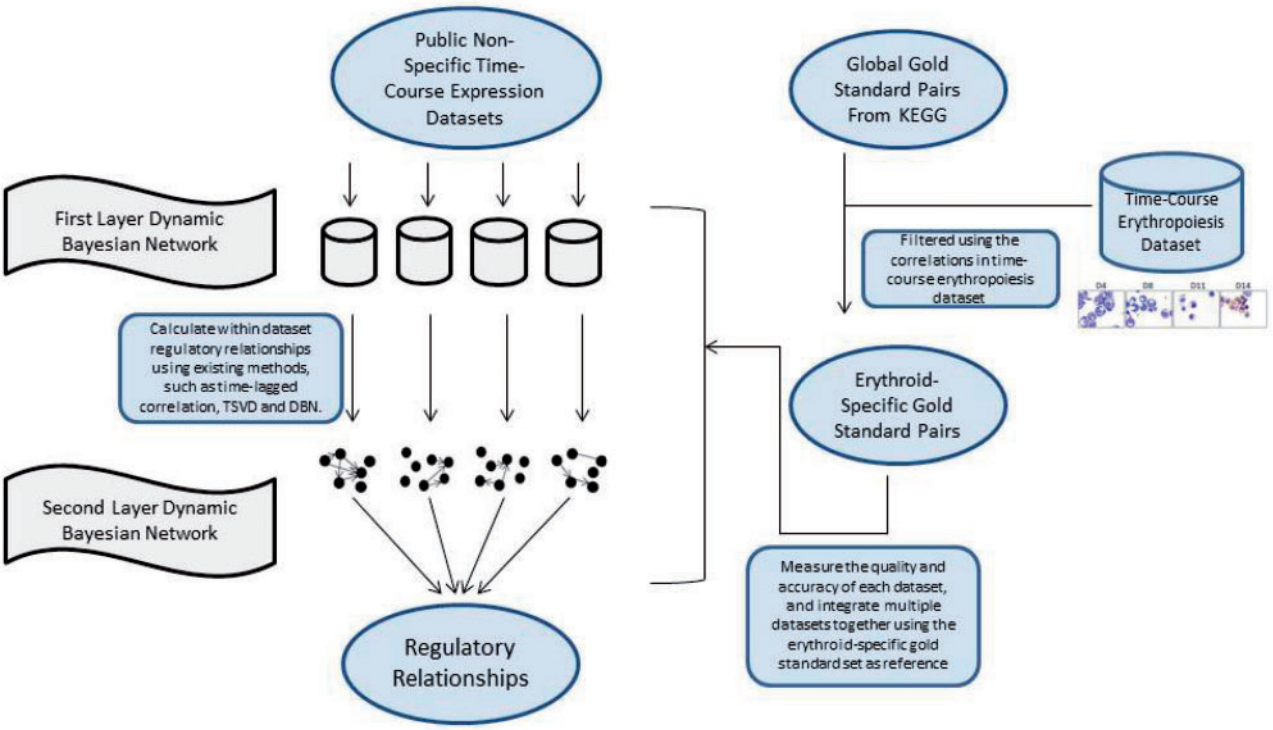

Fig. 1. Strategy for constructing regulatory network through a multi-layer graphical model. Time-course expression datasets were acquired from GEO database. For each selected dataset, we calculated regulatory likelihood score for every gene pair using DBNs. The expression regulation was retreived from public databases and formed the global gold standard, which was then refined into erythroid-specific gold standard according to erythroid expression data. A second layer Bayesian classifier was used to integrate regulatory probablities from multiple datasets together based on the gold standard pairs and to generate the lineage-specific regulatory network

directions) while this workflow is modified and targeting regulatory networks (i.e. with directions). Datasets are weighted automatically based on their qualities and relevance to erythropoiesis. As will be described in the next section, the result indicates that both DBN and Time-Lagged Correlation have satisfying performance in the computational cross validation. In the following analysis, we will use DBN as the base-learner to infer the erythroid-specific network. More details can be found in Supplementary Supporting Information S4.

To evaluate the improvement led by integrating multiple datasets, we generated a model for a global regulatory network in human. This was generated by using gold standard pairs not refined by lineage specificity.

2.2 Collecting gold standard regulatory interactions

2.2.1. Global pairs

Positive gold standard pairs, which represent experimentally validated activation regulatory relationships, were obtained from KEGG (Kanehisa and Goto, 2000). More specifically, a gene pair to be included in the positive gold standards must be marked as positive in 'expression'. A number of 979 positive gold standard regulatory relationships were obtained. Because there is no existing database that defines non-regulatory gene relationships, the negative gold standard was approximated with randomly generated gene pairs.

2.2.2 Erythroid-specific differentiation data and gold standard pairs

To better represent the regulatory relationship in the erythroid differentiation process, we have generated a human erythroid cell differentiation dataset to further refine the gold standards to be cell lineage-specific. Purified human $CD34^+$ hematopoietic progenitor cells were allowed to differentiate *ex vivo* under conditions that were previously reported to achieve 1,000-fold cell proliferation and nearly synchronous erythroid differentiation (Giarratana *et al.*, 2005; Shi *et al.* 2013, 2014). RNA-seq measurements cover four time points, day 4, 8, 11 and 14 after initialization of differentiation, and two biological replicates were taken. This RNA-seq dataset is mapped to human genome using TopHat v2.0.10 (Trapnell *et al.*, 2009) and Cufflinks v2.1.1 (Trapnell *et al.*, 2010) using National Center for Biotechnology Information build 37.2 transcript annotation files. We then calculated Pearson's correlation coefficient for every expressed gene pair (an expressed gene is defined as the gene with Fragments Per Kilobase Of Exon Per Million Fragments Mapped value larger than 1 at least one time point).

An erythroid-specific gold standard pair must satisfy two criteria: (i) it must be included in the generic positive regulatory gold standard pairs; (ii) this gene pair must be co-expressed in the erythroid differentiation dataset ($\rho \geq 0.30$). Two hundred eighty positive gold standard pairs were retained in the erythroid-specific gold standards.

2.2.3 Prior transcription factor information

To further improve our ability to infer the directionality in interactions, a list of 1469 human transcription factors was obtained from (Zhang *et al.*, 2012). The transcriptional factor list was used to refine the prediction result, in which the predicted probability was first up weighted 20 times if i a transcriptional factor and all the probability values are re-scaled to [0,1].

2.3 ChIP-seq experiment

We used ChIP-seq experiment to experimentally evaluate whether the cell lineage-specific modeling can help identify the regulatory interactions specific to erythropoiesis. TR4 binding peaks in ChIP-seq (chromatin immunoprecipitation followed by high-throughput DNA sequencing) data was compared against the predictions. TR4 (orphan nuclear receptor TR4, NR2C2) is a key factor in repressing fetal γ-globin gene transcription and presenting in human and mouse erythropoiesis. TR4 peak annotation was based on the shortest distance from the center of a peak to the transcription start site (TSS) of the nearest RefSeq gene. The TR4 ChIP-seq was performed on Day 8 from CD34+ cell cultures, which generated 32×106 TR4 ChIP-seq enriched, 50-bp short reads. One thousand twenty-five TR4 peaks were identified using a statistical cutoff of $P < 10^{-5}$ (Supplementary Supporting Information S2). This dataset is used as the experimental validation for the erythroid-specific regulatory network. The TR4 downstream targets were validated by knocking down the expression of TR4 followed by transcription profiling with RNA-Seq. From this study, we find out that the genes with TR4 binding at the promoter region, within a 500-bp window, are potential TR4 downstream targets (Shi *et al.*, 2014).

3 Results

3.1 Computational evaluation of the prediction algorithm

We first evaluated the performance of predicting the global regulatory relationships without lineage specificity. To prevent contamination in cross-validation, the gold standard pairs were split into two disjoint graphs, with one serving as training set and the other one as testing set. We have repeated the computational cross-validation 10 times with different partition of the gold standard pairs. We evaluated this model using different base-methods to determine the regulatory likelihood within each dataset, i.e. DBN, time-lagged correlation, Lasso regularization and TSVD. Additionally, we evaluated the improvement gained using prior information of which genes are likely to be transcription factors. Receiver Operating Characteristic curve (ROC) and Precision Recall Curve (PRC) were used to visualize the accuracy of prediction by assessing Type I and Type II error ratios. Area under ROC and PRC were also calculated in order to quantitatively measure the accuracy.

Without prior information of the regulators, the areas under ROC are both 0.568, for DBN and time-lagged correlation (Table 1). The method based on TSVD has lower prediction ability with an area under ROC. Furthermore, Lasso regularization delivers a close-to-random result and was thus removed from the following analysis. This is likely due to the extremely sparse regulatory relationships that were recovered after Lasso regularization, because Lasso's penalty term is not differentiable at 0.

Then, we downloaded all known transcription factors in the human genome (Section 2.2.3). We up-weighted the predicted probability if the pair starts with a transcriptional factor. With prior information of the genes that are likely to be transcription factors, the area under ROC is further improved to extremely high values of

Table 1. Area under receiver operating characteristic curve (AUC) and area under precision recall curve (AUPRC) for the integrated network using different base methods

Without prior	Area under ROC	With prior	Area under ROC
Time-lagged correlation	0.568 (±0.016)	Time-lagged correlation	0.854 (±0.037)
DBN	0.568 (±0.016)	DBN	0.863 (±0.033)
TSVD	0.522 (±0.025)	TSVD	0.783 (±0.056)

0.863 and 0.854 and the area under PRC is increased to 0.347 and 0.346, for DBN and time-lagged correlation, respectively, indicating that the prior information added to this model is correct and is helpful in predicting the directions of the interactions. Figure 2 illustrates the ROC curves and the PRCs of using different base-classifiers to calculate the probabilities. DBN and time-lagged correlation showed high precision (>0.5) in low and median recall area.

For example, in the cross validation, when FOS (FBJ murine osteosarcoma viral oncogene homolog) was a held-out gene, we could observe how well the predictions can recover the interactions of this gene. In the hidden test set, there are nine direct targets (IL1B, CDK4, VEGFB, CCND1, IL6, MMP2, FIGF, IL12B and CCL4) that are activated by FOS. For example, FOS promotes IL1B, IL6 and IL12B via DNA in JAK-STAT signaling pathway (KEGG: hsa04620), and promotes CDK4, VEGFB, MMP2 and CCND1 in PPAR signaling pathway (KEGG: hsa05200). Eight out of the nine targets are correctly identified (Fig. 3) using our regulatory network with a probability value larger than 0.1, which is a cutoff about 20-fold of the randomly predicted probability with a z-score larger than 1. The only case that was missed out in the prediction was CCL4.

3.2 ChIP-seq experiment verifies predicted erythroid-specific regulatory interactions

Because DBN showed decent performance among all base-learners in the cross-validation results of the global network, the erythroid-specific network was calculated using DBN as a base-learner.

In this section, we compare the regulatory probabilities of genes regulated by TR4 against the distance (i.e. the distance between the middle of a peak and the TSS of the nearest gene) measured in the ChIP-seq data. The result is available at Supplementary Supporting Information S3.

We separated the genes into two groups based on the distance from the peak center to the nearest TSS of RefSeq genes. For a random prediction, the probability value should be the baseline—0.05. The distribution of probability values for different groups are displayed (Fig. 4A). For TR4 direct downstream targets, namely the genes within a 500-bp window (low distance group, proximal promoter region), 25% of the regulatory probability is above 0.132. For the genes that are not in the 500-bp window (high distance group), 25% of the predicted regulatory probability is over 0.1512, indicating that TR4 is not regulating the genes in the high distance

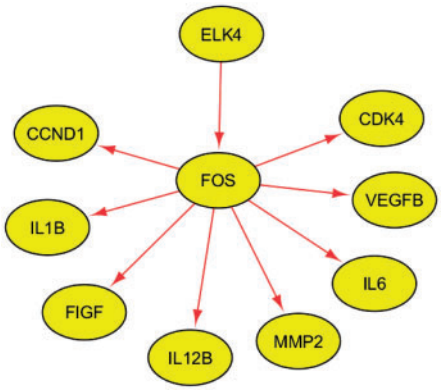

Fig. 3. Example of regulatory relationships predicted in cross-validation. The regulatory network successfully recovered eight out of nine known regulatory relationships in test set for FOS. All these relationships are not included in the training data

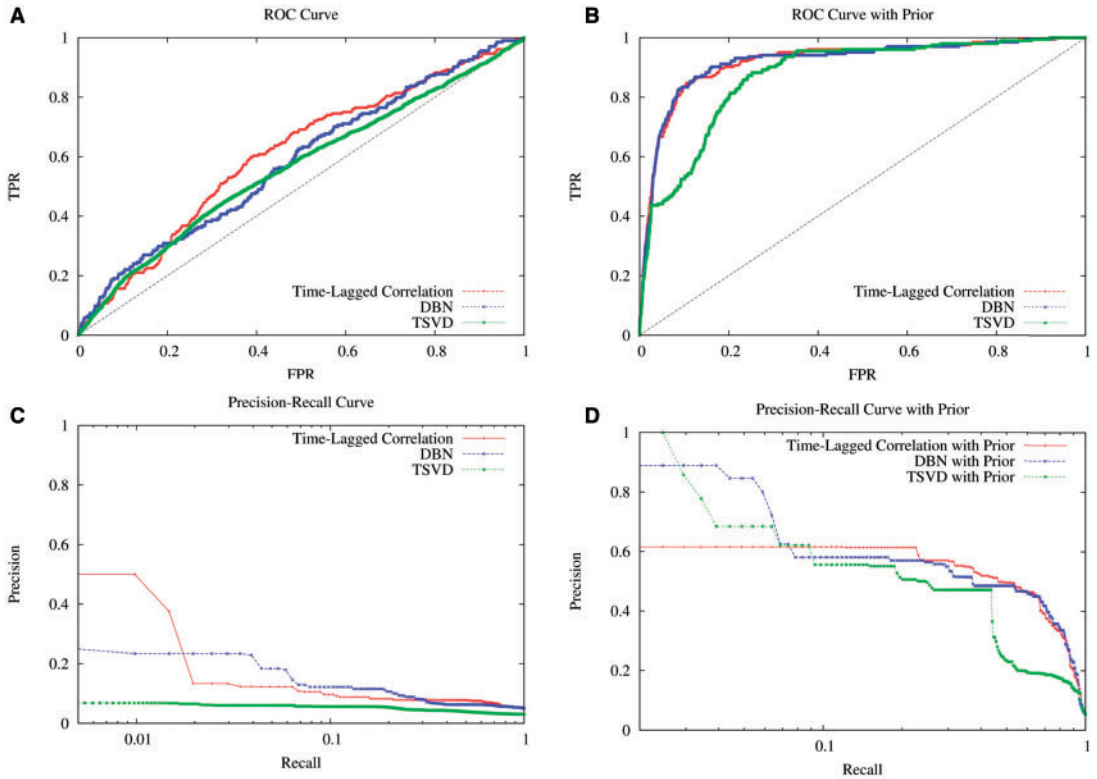

Fig. 2. Cross-validation results for different base classifiers in constructing regulatory networks. This figure shows the cross-validation performance of the models using different base methods to determine the probability within each public dataset. Subfigure A and C contain ROC curves and subfigure B and D contain the PRCs. This figure also illustrates the improvement gained using prior information of the potential regulators (B and D)

group. This shows that DIFMINE predicts the genes in low distance group with a much higher probability than the high distance group ($P = 0.0035$). We have also evaluated the Pearson's correlation coefficient between the predicted probability values and the distance e, the results indicate that they are significantly correlated ($\rho = 0.24$, $P = 2 \times 10^{-5}$).

Supplementary Supporting Information S5 is a representative view of TR4 binding sites of three top DIFMINE predicted genes (in low distance group), MRPL20, NOL7 and CTR9, by integrative genomics viewer (IGV) (Pique-Regi *et al.*, 2011). A clear peak is observed for both these three genes. Based on DIFMINE's prediction, the probability of TR4 regulating MRPL20 is 0.47 ($z = 3.32$); the probability of TR4 regulating NOL7 is 0.43 ($z = 3.02$); the probability of TR4 regulating CTR9 is 0.36 ($z = 2.40$). This functional analysis supports that this method is able to identify direct regulatory relationships that are verified by binding data. Similar trend was observed between genes in the low distance group and all expressed genes.

Furthermore, we found significant improvement achieved by using erythroid-specific integration compared with the global regulatory network (Fig. 4B). The significant difference between low distance group, high distance group and all expressed genes, which is observed in erythroid-specific network, does not exist in the global regulatory network. Additionally, genes in the low distance group and high distance group do not have significantly different probability values ($P = 0.2251$) in the global regulatory network.

3.3 Activities of known erythroid-related genes

Cell lineage-specific integration is further validated by important erythroid genes and their surrounding regulators. Table 2 shows how the erythroid-related genes have significantly higher regulating probability values from their top 10 upstream regulators between erythroid-specific and global inference network. In general, erythroid-related genes are predicted with much higher probabilities in the lineage-specific regulatory network compared with the global network. We repeated this evaluation on 110 expert-curated erythroid-related genes (100 genes from the Hembase database (Goh *et al.*, 2004) and 10 well-known erythroid-related transcriptional factors). Similar trend was observed (Supplementary Supporting Information S1). For example, the top 10 genes regulating hemoglobin beta (HBB) have a significantly different probability values ($P = 5 \times 10^{-5}$) for erythroid-specific network (average probability = 0.652) then in the global

Table 2. Comparison of the predicted regulatory interactions of erythroid-related genes in the lineage-specific network and the global network

Gene name	Average connection strength in erythroid-specific network	Average connection strength in global network	*P* Value
HBB	0.652	0.451	5×10^{-5}
HBD	0.414	0.308	1×10^{-2}
HBA1	0.304	0.199	4×10^{-3}
HBG1	0.145	0.008	6×10^{-4}
HBG2	0.172	0.100	9×10^{-6}
GATA1	0.463	0.266	9×10^{-4}
GYPA	0.354	0.258	1×10^{-2}

This table indicates the difference of regulating probabilities of the top 10 predicted regulators toward erythroid-related genes.

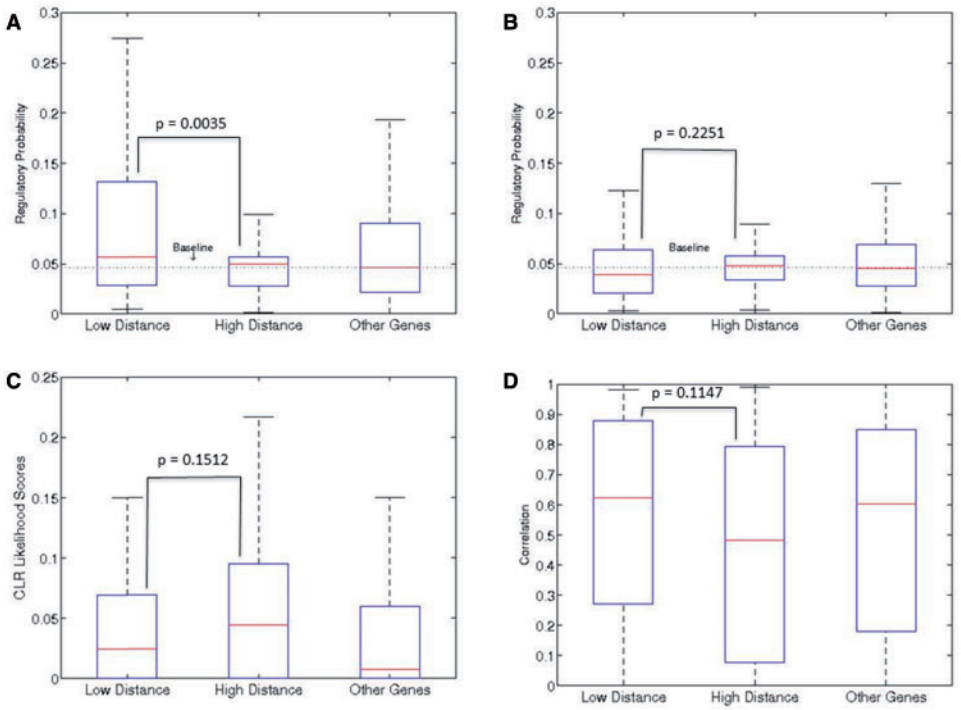

Fig. 4. Validation of predicted probabilities of regulatory relationships using TR4 ChIP-seq data. Genes are split into two groups based on their distance from the peak center to the nearest TSS in RefSeq. The low distance group contains those genes with distances <300 bp. The high distance group contains genes that are >500 bp away and last group contains all the genes that do not have significant peaks. Figure 4A shows the distribution of predicted probabilities using DIFMINE with prior information, Figure 4B is regulatory probability for the global regulatory network inferred with prior information but without the differentiation-specificity. Figure 4C is the result from CLR. Figure 4D is directly using the correlation calculated on the erythropoiesis expression data. The *P* values indicating the difference between low distance group and high distance group are calculated using Wilcoxon rank-sum test

network (average probability = 0.451). This result supports the utility brought by cell lineage-specific integration.

3.4 Ability to predict GRNs with small number of samples compared with previous methods designed for large datasets

One most important utility of DIFMINE is that it is capable of interrogating context-specific regulatory relationships using a dataset with limited samples. To demonstrate this utility, we compared DIFMINE with a classical method, CLR. We used CLR to infer the regulatory relationships based on the same time-course RNA-Seq data we have collected. Figure 4C shows the distribution of CLR likelihood scores for the three groups in the ChIP-seq data, where noticeable difference between these three groups cannot be observed. This is further reflected by the non-significant Wilcoxon rank-sum test result (between the low distance and high distance group, $P = 0.1512$). The correlation coefficient between predicted scores and distance is not significant either ($\rho = 0.072$, $P = 0.117$). We attribute the unsatisfactory performance of CLR to the small number of samples, i.e. CLR usually requires 10 of samples, but in this erythropoiesis dataset, only 8 samples (four time points and two replicates) are available. In fact, most methods based on co-expression, mutual information or any types of correlation could be severely harmed by the lack of samples, if no outside large-scale data is leveraged. This is further supported by directly using the co-expression networks as a prediction of regulatory relationships (Fig. 4D), where the distance between TR4 and the middle point of genes is only marginally correlated with their correlation coefficients ($\rho = 0.172$, $P = 0.049$).

This integration model, DIFMINE, uses both the context-specific and non-specific data to refine the regulatory network that is able to minimize the negative effect of the uncertainty due to insufficient sample size, and to adjust the network to fit the specific context.

3.5 Computational time

Assuming that the number of genes in a dataset is N, the number of dataset is M, and the number of samples is K, the computational complexity to perform each base-learner is $O(MN^2K)$. For the second layer of Bayesian network (the integration step), the computational complexity is $O(MN^2)$.

In the experiment that was performed on a common commercial computer (with PowerEdge R410 processor), the running time for calculating probabilities over all datasets (the first layer of Bayesian network) is similar between different base-learners (DBN, time-lagged correlation or TSVD), which is 5 to 7 days. More specifically, it takes 2–5 h for each dataset depending on the coverage (i.e. number of genes) of this dataset. As the calculation tasks between different datasets are independent, parallelization (Zhu *et al.*, 2011, 2012) can also be used to speed up this step, i.e. giving enough central processing unit/graphics processor unit cores, the computational time can potentially be reduced to $O(N^2K)$. The running time for the second layer of DBN is ~16 h. The computational time is expected to be similar when analyzing different biological systems. Furthermore, results of the first layer of Bayesian network can be re-used for different biological systems.

4 Discussion

Predicting cell lineage-specific GRNs is an important yet challenging problem in system biology. This is mainly due to the fact that lineage-specific expression data are usually limited in sample size. In this study, we developed a novel algorithm that models cell lineage-specific regulatory networks and applied it to erythropoiesis. This algorithm advances previous methods on two aspects: (i) Cell lineage-specificity: experimental validation indicates that it is more accurate than the classical methods and global integration in predicting erythroid-specific regulatory interactions. (ii) Robustness: the interrogation of multiple datasets for their relevance of a specific developmental process makes the integration more robust than using a single dataset.

We have demonstrated the accuracy of the reconstructed regulatory network via computational and experimental validations. In computational cross-validation, results indicate that this model could precisely recover the hidden test set of gold standards. With prior information of a list of transcription factors downloaded from publicly available databases (Zhang *et al.*, 2012), DIFMINE achieved area under ROC and PRC to 0.863 and 0.347, respectively. In experimental validation, results of DIFMINE are validated using ChIP-seq transcriptional factor binding peaks. A strong correlation between the predicted regulatory probabilities and the distances between the peak and the TSS site ($P = 2 \times 10^{-5}$) was observed. Genes in the low distance group and high distance group also have a significantly different probability values ($P = 0.0035$) based on Wilcoxon rank-sum test. Both of the computational and experimental evaluations supports that the accuracy of this model compared with previous models or context non-specific integrations.

Funding

This work is supported by NSF 1452656, NIH 1R21NS082212-01, EU-FP VII Systems Biology of Rare Disease, NIH University of Michigan O'Brien Kidney Translational Core Center, AHA Midwest postdoctoral fellowship N012616, NIH grants R21 HL114368 and U01 HL117658.

Conflict of Interest: none declared.

References

Altieri,D.C. (2008) Survivin, cancer networks and pathway-directed drug discovery. *Nat. Rev. Cancer*, **8**, 61–70.

Chikina,M.D. *et al.* (2009) Global prediction of tissue-specific gene expression and context-dependent gene networks in Caenorhabditis elegans. *PLoS Comput. Biol.*, **5**, e1000417.

Cui,S. *et al.* (2011) Nuclear receptors TR2 and TR4 recruit multiple epigenetic transcriptional corepressors that associate specifically with the embryonic beta-type globin promoters in differentiated adult erythroid cells. *Mol. Cell. Biol.*, **31**, 3298–3311.

Edgar,R. *et al.* (2002) Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.

Ernst,J. *et al.* (2008) A semi-supervised method for predicting transcription factor—gene interactions in Escherichia coli. *PLoS Comput. Biol.*, **4**, e1000044.

Faith,J.J. *et al.* (2007) Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol.*, **5**, e8.

Flassig,R.J. *et al.* (2013) An effective framework for reconstructing gene regulatory networks from genetical genomics data. *Bioinformatics*, **29**, 246–254.

Friedman,N. *et al.* (1998) Learning the structure of dynamic probabilistic networks. In: *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc.

Garcia-Echeverria,C. and Sellers,W. (2008) Drug discovery approaches targeting the PI3K/Akt pathway in cancer. *Oncogene*, **27**, 5511–5526.

Giarratana,M.-C. *et al.* (2005) Ex vivo generation of fully mature human red blood cells from hematopoietic stem cells. *Nat. Biotechnol.*, **23**, 69–74.

Gitter,A. *et al.* (2010) Computational methods for analyzing dynamic regulatory networks. *Methods Mol. Biol.*, **674**, 419–441.
Goh,S.H. *et al.* (2004) Hembase: browser and genome portal for hematology and erythroid biology. *Nucleic Acids Res.*, **32**(Database issue), D572–D574.
Guan,Y. *et al.* (2008) A genomewide functional network for the laboratory mouse. *PLoS Comput. Biol.*, **4**, e1000165.
Guan,Y. *et al.* (2012) Tissue-specific functional networks for prioritizing phenotype and disease genes. *PLoS Comput. Biol.*, **8**, e1002694.
Hecker,M. *et al.* (2009) Gene regulatory network inference: data integration in dynamic models—a review. *Biosystems*, **96**, 86–103.
Hennessy,B.T. *et al.* (2005) Exploiting the PI3K/AKT pathway for cancer drug discovery. *Nat. Rev. Drug Discov.*, **4**, 988–1004.
Holter,N.S. *et al.* (2001) Dynamic modeling of gene expression data. *Proc. Natl. Acad. Sci.*, **98**, 1693–1698.
Huang,S. (1999) Gene expression profiling, genetic networks, and cellular states: an integrating concept for tumorigenesis and drug discovery. *J. Mol. Med.*, **77**, 469–480.
Huttenhower,C. *et al.* (2008) The Sleipnir library for computational functional genomics. *Bioinformatics*, **24**, 1559–1561.
Huttenhower,C. *et al.* (2009) Detailing regulatory networks through large scale data integration. *Bioinformatics*, **25**, 3267–3274.
Huttenhower,C. *et al.* (2009) Exploring the human genome with functional maps. *Genome Res.*, **19**, 1093–1106.
Irrthum,A. *et al.* (2010) Inferring regulatory networks from expression data using tree-based methods. *PLoS One*, **5**, e12776.
Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
Lee,I. *et al.* (2011) Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res.*, **21**, 1109–1121.
Ma,P. *et al.* (2006) A data-driven clustering method for time course gene expression data. *Nucleic Acids Res.*, **34**, 1261–1269.
Marbach,D. *et al.* (2010) Revealing strengths and weaknesses of methods for gene network inference. *Proc. Natl Acad. Sci. USA*, **107**, 6286–6291.
Marbach,D. *et al.* (2012) Wisdom of crowds for robust gene network inference. *Nat. Methods*, **9**, 796–804.
Margolin,A.A. *et al.* (2013) Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. *Sci. Transl. Med.*, **5**, 181re181–181re181.
Menéndez,P. *et al.* (2010) Gene regulatory networks from multifactorial perturbations using graphical Lasso: application to the DREAM4 challenge. *PloS One*, **5**, e14147.
Michoel,T. *et al.* (2009) Comparative analysis of module-based versus direct methods for reverse-engineering transcriptional regulatory networks. *BMC Syst. Biol.*, **3**, 49.
Mordelet,F. and Vert,J.-P. (2008) SIRENE: supervised inference of regulatory networks. *Bioinformatics*, **24**, i76–i82.
Park,C.Y. *et al.* (2010) Simultaneous genome-wide inference of physical, genetic, regulatory, and functional pathway components. *PLoS Comput. Biol.*, **6**, e1001009.
Pique-Regi,R. *et al.* (2011) Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.*, **21**, 447–455.
Pop,A. *et al.* (2010) Integrated functional networks of process, tissue, and developmental stage specific interactions in Arabidopsis thaliana. *BMC Syst. Biol.*, **4**, 180.
Poultney,C.S. *et al.* (2012) Integrated inference and analysis of regulatory networks from multi-level measurements. *Methods Cell Biol.*, **110**, 19–56.
Prill,R.J. *et al.* (2010) Towards a rigorous assessment of systems biology models: the DREAM3 challenges. *PLoS One*, **5**, e9202.
Prill,R.J. *et al.* (2011) Crowdsourcing network inference: the DREAM predictive signaling network challenge. *Sci. Signal.*, **4**, mr7.
Schmitt,W.A. *et al.* (2004) Elucidation of gene interaction networks through time-lagged correlation analysis of transcriptional data. *Genome Res.*, **14**, 1654–1663.
Shi,L. *et al.* (2013) Lysine-specific demethylase 1 is a therapeutic target for fetal hemoglobin induction. *Nat. Med.*, **19**, 291–294.
Shi,L. *et al.* (2014) Biased, non-equivalent gene-proximal and-distal binding motifs of orphan nuclear receptor TR4 in primary human erythroid cells. *PLoS Genet.*, **10**, e1004339.
Shi,L. *et al.* (2014) Developmental transcriptome analysis of human erythropoiesis. *Hum. Mol. Genet.*, ddu167.
Shinozaki,K. *et al.* (2003) Regulatory network of gene expression in the drought and cold stress responses. *Curr. Opin. Plant Biol.*, **6**, 410–417.
Steiert,B. *et al.* (2012) Experimental design for parameter estimation of gene regulatory networks. *PLoS One*, **7**, e40052.
Tanabe,O. *et al.* (2007) The TR2 and TR4 orphan nuclear receptors repress Gata1 transcription. *Genes Dev.*, **21**, 2832–2844.
Trapnell,C. *et al.* (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.
Trapnell,C. *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.
Vogelstein,B. and Kinzler,K.W. (2004) Cancer genes and the pathways they control. *Nat. Med.*, **10**, 789–799.
Welch,L. *et al.* (2014) Bioinformatics curriculum guidelines: toward a definition of core competencies. *PLoS Comput. Biol.*, **10**, e1003496.
Wong,A.K. *et al.* (2012) IMP: a multi-species functional genomics portal for integration, visualization and prediction of protein functions and networks. *Nucleic Acids Res.*, **40**, W484–W490.
Yeung,M.S. *et al.* (2002) Reverse engineering gene networks using singular value decomposition and robust regression. *Proc. Natl Acad. Sci. USA*, **99**, 6163–6168.
Yip,K.Y. *et al.* (2010) Improved reconstruction of in silico gene regulatory networks by integrating knockout and perturbation data. *PLoS One*, **5**, e8121.
Yu,J. *et al.* (2004) Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics*, **20**, 3594–3603.
Zhang,H.-M. *et al.* (2012) AnimalTFDB: a comprehensive animal transcription factor database. *Nucleic Acids Res.*, **40**, D144–D149.
Zhu,F. *et al.* (2011) A parallel deconvolution algorithm in perfusion imaging. In: *Healthcare Informatics, Imaging and Systems Biology (HISB), 2011.* First IEEE International Conference on, IEEE.
Zhu,F. *et al.* (2012) Computed tomography perfusion imaging denoising using Gaussian process regression. *Phys. Med. Biol.*, **57**, N183.
Zhu,F. *et al.* (2012) Parallel perfusion imaging processing using GPGPU. *Comput. Methods Programs Biomed.*, **108**, 1012–1021.
Zhu,F. *et al.* (2013) Lesion area detection using source image correlation coefficient for CT perfusion imaging. *IEEE J. Biomed. Health Inform.*, **17**, 950–958.
Zhu,F. *et al.* (2014) Modeling dynamic functional relationship networks and application to ex vivo human erythroid differentiation. *Bioinformatics*, **30**, 3325–3333.
Zhu,F. and Guan,Y. (2014) Predicting dynamic signaling network response under unseen perturbations. *Bioinformatics*, **30**, 2772–2778.
Zou,M. and Conzen,S.D. (2005) A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics*, **21**, 71–79.